UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

SNP effects from ssGBLUP using the BLUPF90 family (postGSf90)

Daniela Lourenco
BLUPF90 Team – 02/2023

1

Equivalence between GBLUP and SNP-BLUP

GBLUP

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{G}^{-1}\lambda_1 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

$\text{Var}(\mathbf{u}) = ?$

$\text{Var}(\mathbf{u}) = \mathbf{G}\sigma_u^2$

SNP-BLUP (Ridge Regression)

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{I}\lambda_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

↓

SNP effects

$\mathbf{u} = \mathbf{Za}$

2

2

Are GBLUP and SNP-BLUP equivalent?

- Assumption of GBLUP: $\text{Var}(\mathbf{u}) = \mathbf{G}\sigma_u^2$
- In SNP-BLUP: $\mathbf{u} = \mathbf{Za}$

$\mathbf{u} = \mathbf{Za}$

$\text{Var}(\mathbf{u}) = \text{Var}(\mathbf{Za})$

$\text{Var}(\mathbf{u}) = \mathbf{Z}\,\text{Var}(\mathbf{a})\,\mathbf{Z}'$

$\text{Var}(\mathbf{u}) = \mathbf{ZZ}'\sigma_a^2$

$$\sigma_a^2 = \frac{\sigma_u^2}{2\sum_{i=1}^{SNP} p_i(1-p_i)}$$

$$\text{Var}(\mathbf{u}) = \mathbf{ZZ}'\frac{\sigma_u^2}{2\sum_{i=1}^{SNP} p_i(1-p_i)}$$

$$\text{Var}(\mathbf{u}) = \frac{\mathbf{ZZ}'}{2\sum_{i=1}^{SNP} p_i(1-p_i)}\sigma_u^2$$

Genomic relationship matrix VanRaden (2008)

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum_{i=1}^{SNP} p_i(1-p_i)}$$

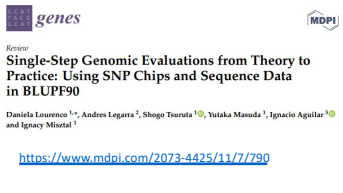
$\text{Var}(\mathbf{u}) = \mathbf{G}\sigma_u^2$ ➡ GBLUP assumption!!!

3

3

GBLUP and SNP-BLUP are equivalent!

If we can get **u** ($\mathbf{u} = \mathbf{Za}$) from SNP-BLUP, we can get **a** from GBLUP!


genes
MDPI
Review
Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90
Daniela Lourenco [1,*], Andres Legarra [2], Shogo Tsuruta [1], Yutaka Masuda [1], Ignacio Aguilar [3] and Ignacy Misztal [1]

https://www.mdpi.com/2073-4425/11/7/790

Pages 11-12

4

4

ssGBLUP and ssSNP-BLUP are also equivalent!

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{H}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

ssGBLUP

Misztal et al. (2009)
Legarra et al. (2009)
Aguilar et al. (2010)
Christensen & Lund (2010)

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'ZM} & \mathbf{X}_n'\mathbf{Z}_n \\ \mathbf{M'Z'X} & \mathbf{M'Z'ZM}+\mathbf{I}\frac{\sigma_e^2}{\sigma_\alpha^2} & \mathbf{M}_n'\mathbf{Z}_n'\mathbf{Z}_n \\ \mathbf{Z}_n'\mathbf{X}_n & \mathbf{Z}_n'\mathbf{Z}_n\mathbf{M}_n & \mathbf{Z}_n'\mathbf{Z}_n+\mathbf{A}^{nn}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \\ \hat{\epsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{M'Z'y} \\ \mathbf{Z}_n'\mathbf{y}_n \end{bmatrix}$$

ssSNPBLUP or **ssBR**

Fernando et al. (2014)
Liu et al. (2014)
Mantysaari & Stranden (2016)

J. Dairy Sci. 101:10082–10088
https://doi.org/10.3168/jds.2018-14913

***Short communication:* Genomic prediction using different single-step methods in the Finnish red dairy cattle population**

H. Gao,*† M. Koivula,‡ J. Jensen,* I. Strandén,‡ P. Madsen,* T. Pitkänen,‡ G. P. Aamand,† and E. A. Mäntysaari‡
*Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark
†Nordic Cattle Genetic Evaluation, DK-8200 Aarhus, Denmark
‡Natural Resources Institute Finland (Luke), FIN-31600 Jokioinen, Finland

We confirmed that regular ssGBLUP and ssBR with an extra polygenic effect led to the same predictions.

SNP effects in ssGBLUP

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{H}^{-1}\lambda_1 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

$$\hat{\mathbf{a}} = \alpha b \frac{1}{2\sum p_i(1-p_i)} \mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}}$$

Matrix of SNP content ($\mathbf{Z}'$)
Genomic relationship matrix ($\mathbf{G}$)

α = blending parameter for $\mathbf{G}$

$$\lambda = \frac{1}{n^2}\left(\sum_i\sum_j \mathbf{A}_{22_{ij}} - \sum_i\sum_j \mathbf{G}_{ij}\right)$$

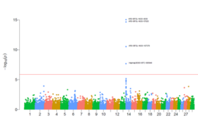
$$b = 1 - \frac{\lambda}{2}$$

What can we do with SNP effects?

1) Predictions for animals not included in the evaluation

Indirect predictions
Indirect Genomic Predictions

2) Genome-Wide Association Studies (GWAS)



1) Indirect Predictions

- Interim evaluations
  - Between official runs
- Not all genotyped animals are in the evaluations
  - Animals with incomplete pedigree increase bias and lower R^2
- Commercial products
  - e.g. GeneMax for non-registered animals

1) Indirect Predictions

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{H}^{-1}\lambda_1 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix} \quad \Longrightarrow \quad \hat{\mathbf{a}} = \alpha b \frac{1}{2\sum p_i(1-p_i)} \mathbf{Z'G}^{-1}\hat{\mathbf{u}}$$

Indirect Prediction: $\mathbf{IP} = \boldsymbol{u}_{\boldsymbol{m}}^{*} = \mathbf{Z}\hat{\boldsymbol{a}}$

9

9

1) Indirect Predictions

Indirect Prediction: $\boldsymbol{u}_{\boldsymbol{m}}^{*} = \mathbf{Z}\hat{\boldsymbol{a}}$

└ Fine if comparing among animals with IP

- Not fine if compare it with GEBV from the main evaluation
  - Put it in the pedigree scale

$$\boldsymbol{u}_{\boldsymbol{m}} = \hat{\boldsymbol{\mu}} + \boldsymbol{u}_{\boldsymbol{m}}^{*}$$

$$\hat{\boldsymbol{\mu}} = \alpha\lambda \mathbf{1'G}^{-1}\hat{\mathbf{u}}$$

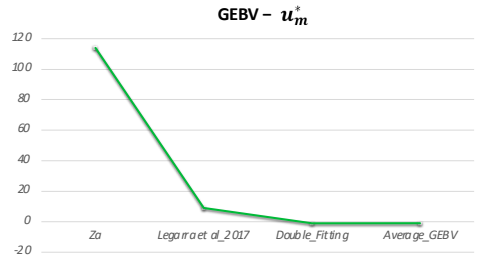
α = blending parameter for $\mathbf{G}$

$$\lambda = \frac{1}{n^2}\left(\sum_i \sum_j \mathbf{A}_{22_{ij}} - \sum_i \sum_j \mathbf{G}_{ij}\right)$$

10

10

1) Indirect Predictions



11

11

How to compute Indirect predictions

1) Pedigree + phenotypes + genotypes

2) `renumf90`

3) preGSf90 to save clean files

4) `blupf90+` (save the clean files)

- Good practice to save time: `OPTION saveGInverse` + `OPTION saveA22Inverse`

5) `postGSf90` (with clean files)

- BLUPF90 family software to compute SNP effects (+more)
- Same parameter file as `blupf90+`
- Good practice to save time: `OPTION readGInverse` + `OPTION readA22Inverse`

12

12

Output from postGSf90

snp_sol — http://nce.ads.uga.edu/wiki/doku.php?id=readme.pregsf90

contains solutions of SNP and weights

- 1: trait
- 2: effect
- 3: SNP
- 4: Chromosome
- 5: Position
- 6: SNP solution
- 7: weight

snp_pred

- 1st line: model, tuning, blending information
- 2nd line: Trait/effect info
- AF in 10 columns
- mu_hat, var_mu_hat
- SNP effects

13

13

How to compute Indirect Predictions

6) predf90

- Have to provide a SNP file for the new genotyped animals to receive IP
  - same SNP as in the clean file

```
predf90 --snpfile newgen.txt --use_mu_hat
```

- The last statement adds the base, so that we have: $\boldsymbol{u_m = \hat{\mu} + u_m^*}$

http://nce.ads.uga.edu/wiki/doku.php?id=readme.predf90

14

14

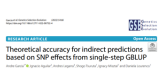
Output from predf90

SNP_predictions

Animal ID	SNP call rate	Indirect Predictions
UGA50014	1.00	0.17414457
UGA50016	1.00	0.72332874E-01
UGA50042	1.00	1.0016705
UGA50058	1.00	0.17190497
UGA50060	1.00	0.98674759E-01
UGA50065	1.00	-0.60623702E-01
UGA50073	1.00	-0.17060851
UGA50077	1.00	-0.21597147
UGA50079	1.00	-0.69586390
UGA50084	1.00	1.0600574
UGA50085	1.00	-0.28602412
UGA50088	1.00	-0.12758011

`predf90` can also compute accuracy of indirect predictions

OPTION snp_p_value	#in `blupf90+`
OPTION snp_var	#in `postGSf90`
--acc	#in `predf90`



15

15

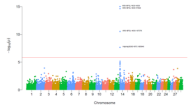
2) Genome-wide Association Studies

16

16

Current standard for GWAS

- Single marker regression with **G** to compensate for relationships
  - $\mathbf{y} = \mathbf{Xb} + \mathbf{z}_i\mathbf{a}_i + \mathbf{u} + \mathbf{e}$
    - **z:** gene content {0,1,2}
    - **a:** SNP effect
- Estimate SNP effects
- Get p-values as $pval_i = 2\left(1 - \Phi\left(\left|\frac{\hat{a}_i}{sd(\hat{a}_i)}\right|\right)\right)$
- Apply Bonferroni to correct for multiple testing



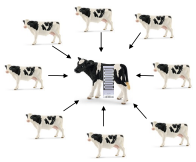
- **Assumption: Genotyped individuals have phenotypes**

17

17

GWAS in livestock populations

- Most animals are non-genotyped
- Animals may not have phenotypes
- Some traits are sex-limited
  - milk, fat, protein
- Single marker regression
  - Only genotyped animals with phenotypes
  - Deregressed EBV
- Need a method that fits the livestock data
  - ssGWAS



18

18

Single-step GWAS (historical)

SNP effects

GEBVs

$$\hat{\mathbf{a}} = \alpha b \frac{1}{2\sum p_i(1-p_i)} \mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}}$$

VanRaden 2008
Stranden and Garrick 2009
Wang et al. 2012

a) Quadratic SNP variance (Falconer & Mackay, 1996)

$$d_i = \hat{a}_i^2 2p_i(1-p_i)$$

b) NonlinearA SNP variance (VanRaden, 2008)

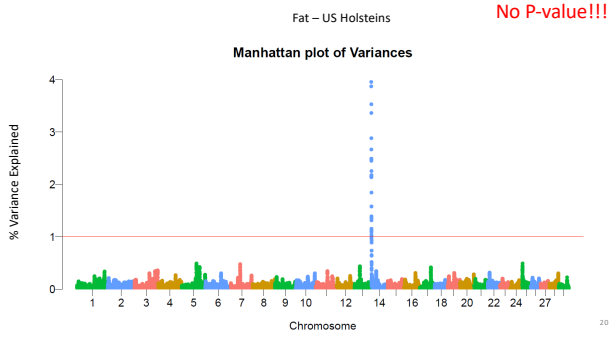
$$d_i = 1.125^{\frac{|\hat{a}_i|}{sd(a)}-2}$$

19

19

Single-step GWAS

No P-value!!!

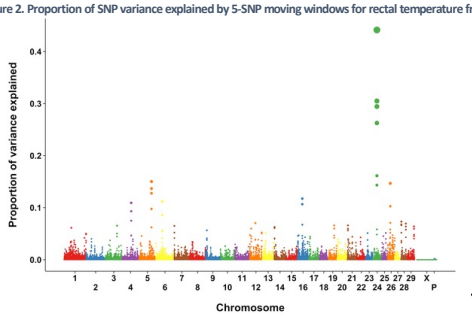


20

20

Single-step GWAS

Figure 2. Proportion of SNP variance explained by 5-SNP moving windows for rectal temperature from a single-step GBLUP analysis



No P-value!!!

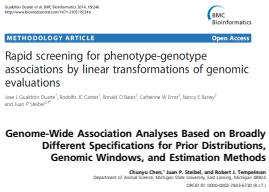
PLOS ONE

Dikmen S, Cole JB, Null DJ, Hansen PJ (2013) Genome-Wide Association Mapping for Identification of Quantitative Trait Loci for Rectal Temperature during Heat Stress in Holstein Cattle. PLOS ONE 8(7): e69202. https://doi.org/10.1371/journal.pone.0069202
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0069202

21

21

Can we have p-values in ssGWAS?


BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations

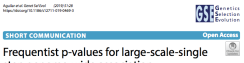
Genome-Wide Association Analyses Based on Broadly Different Specifications for Prior Distributions, Genomic Windows, and Estimation Methods


ANIMAL GENETICS

Meta-analysis of genome-wide association from genomic prediction models


J. Dairy Sci. 101:3148–3154 https://doi.org/10.3168/jds.2017-13364

Genome-wide association analyses based on a multiple-trait approach for modeling feed efficiency


SHORT COMMUNICATION Open Access

Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle

22

22

P-values in ssGWAS

1) Factorize and Invert LHS of ssGBLUP with YAMS (Masuda et al., 2014)

2) Solve the MME for $\begin{bmatrix}\hat{\beta}\\ \hat{u}\end{bmatrix}$ using the sparse Cholesky factor

blupf90+


Ignacio Aguilar

3) Extract coefficients for genotyped animals ($\mathbf{C}^{u_2u_2}$) from LHS^{-1}

4) Obtain individual prediction error variance of SNP effects:

$$Var(\hat{a}_i) = \alpha b \frac{1}{2\sum p_i(1-p_i)} \mathbf{z}_i'\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{u_2u_2})\mathbf{G}^{-1}\mathbf{z}_i \frac{1}{2\sum p_i(1-p_i)} \alpha b$$

(Gualdron-Duarte et al., 2014)

postGSf90


Andres Legarra

5) Backsolve GEBV to SNP effects ($\hat{a}$): $\hat{a} = \alpha b \frac{1}{2\sum p_i q_i} \mathbf{Z}'\mathbf{G}^{-1}\hat{u}$


Yutaka Masuda

6) p-value$_i = 2\left(1 - \Phi\left(\left|\frac{\hat{a}_i}{sd(\hat{a}_i)}\right|\right)\right)$

Φ is the cumulative standard normal function

23

23

How to run ssGWAS with p-values in BLUPF90

- After renumf90 and preGSf90 to save clean files:
  - blupf90+ to estimate GEBV
    - OPTION SNP_file snp.dat_clean
    - OPTION map_file mrkmap.txt_clean
    - OPTION saveGInverse
    - OPTION saveA22Inverse
    - OPTION snp_p_value
  - postGSf90 to backsolve GEBV to SNP effect
    - OPTION SNP_file snp.dat_clean
    - OPTION map_file mrkmap.txt_clean
    - OPTION readGInverse
    - OPTION readA22Inverse
    - OPTION snp_p_value
    - OPTION windows_variance X #if need variance explained by X SNP

24

24

Output from postGSf90

`chrsnp_pval`

contains data to create plot by GNUPLOT

- 1: trait
- 2: effect
- 3: -log10(p-value)
- 4: SNP
- 5: Chromosome
- 6: Position in bp

Pft1e2.gnuplot

Pft1e2.R

`chrsnp`

contains data to create plot by GNUPLOT

- 1: trait
- 2: effect
- 3: values of SNP effects to use in Manhattan plots → [abs(SNP_i)/var(SNP)]
- 4: SNP
- 5: Chromosome
- 6: Position

Sft1e2.gnuplot

Sft1e2.R

25

25

Output from postGSf90

`snp_sol`

contains solutions of SNP and weights

- 1: trait
- 2: effect
- 3: SNP
- 4: Chromosome
- 5: Position
- 6: SNP solution
- 7: weight

if `OPTION windows_variance` is used

- 8: variance explained by *n* adjacents SNP.

if `OPTION snp_p_value` is used

- 9: variance of the SNP solution (used to compute the p-value)

26

26

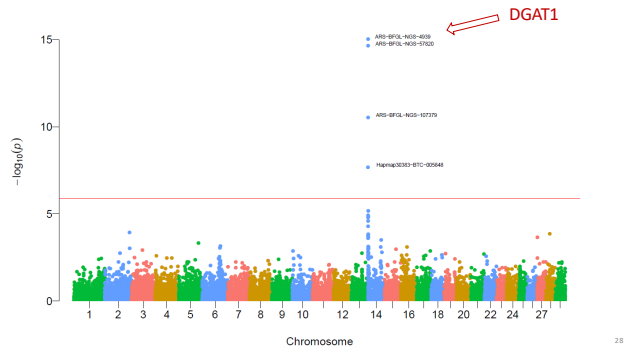
P-values in ssGWAS for US Holsteins

- US HOL 2009 data: milk, fat, protein
- Single-trait models
  - 10k genotyped bulls
  - 752k records for 100k daughters
  - 303k animals in ped

27

27

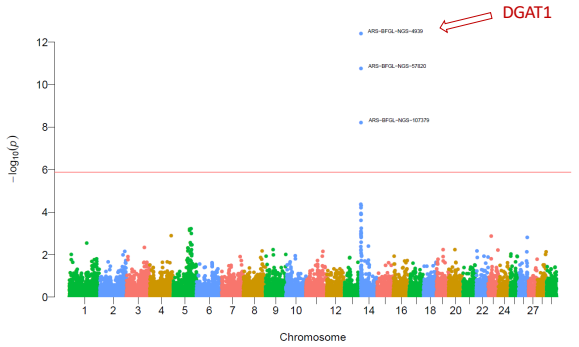
P-values in ssGWAS - Milk



28

28

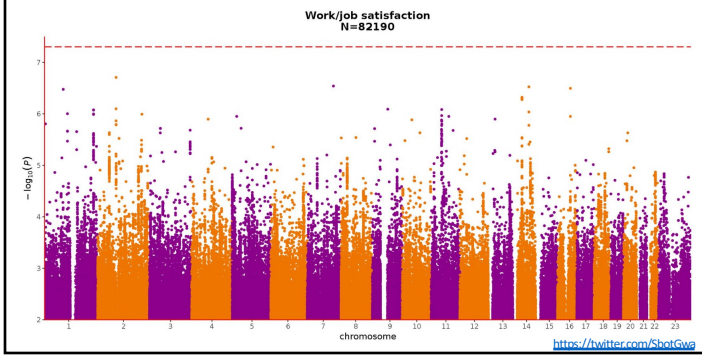
P-values in ssGWAS - Fat



29

29

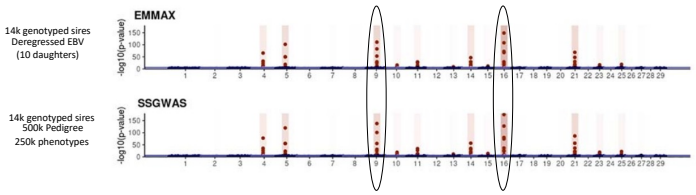
Non-significant hits



30

ssGWAS vs. EMMAX

- Simulated population (1 QTN per CHR)



Association	EMMAX (Khang et al., 2010)	ssGWAS (Aguilar et al., 2019)
True Positive	55.2[a] (3.7)	61.6[a] (8.7)
False Positive	0.0	0.0

Mancin et al. (2021)

31

31

postGSf90 options

http://nce.ads.uga.edu/wiki/doku.php?id=readme.pregsf90

```
OPTION Manhattan_plot
```

Uses GNUPLOT to plot the Manhattan plot (SNP effects) for each trait and correlated effect.

```
OPTION Manhattan_plot_R
```

Uses R to plot the Manhattan plot (SNP effects) for each trait and correlated effect.
pdf images are created. *manplot_Str1e2.pdf*, but other formats can be specified.
Note: *1*e*2* corresponds to trait 1, effect 2.

```
OPTION Manhattan_plot_R_format <format>
```

Control the format type to create images in R
`format` values accepted:

- pdf (default)
- png
- tif

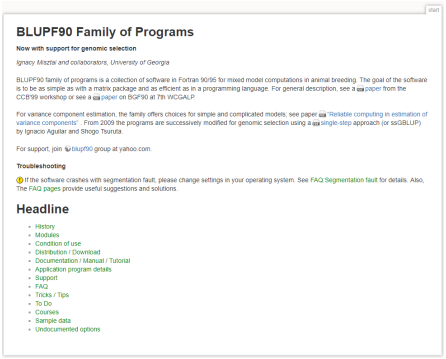
```
OPTION plotsnp <n>
```

Control the values of SNP effects to use in Manhattan plots

- 1: plot regular SNP effects: abs(val)
- 2: plot standardized SNP effects: abs(val/sd) (default)

32

nce.ads.uga.edu/wiki



start

BLUPF90 Family of Programs

Now with support for genomic selection

Ignacy Misztal and collaborators, University of Georgia

BLUPF90 family of programs is a collection of software in Fortran 90/95 for mixed model computations in animal breeding. The goal of the software is to be as simple as with a matrix package and as efficient as in a programming language. For general description, see a paper from the CCB'99 workshop or see a paper on BGF'90 at 7th WCGALP.

For variance component estimation, the family offers choices for simple and complicated models; see paper "Reliable computing in estimation of variance components". From 2009 the programs are successively modified for genomic selection using a single-step approach (or ssGBLUP) by Ignacio Aguilar and Shogo Tsuruta.

For support, join blupf90 group at yahoo.com.

Troubleshooting

If the software crashes with segmentation fault, please change settings in your operating system. See FAQ Segmentation fault for details. Also, The FAQ pages provide useful suggestions and solutions.

Headline

- History
- Modules
- Condition of use
- Distribution / Download
- Documentation / Manual / Tutorial
- Application program details
- Support
- FAQ
- Tricks / Tips
- To Do
- Courses
- Sample data
- Undocumented options

33