UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences
Animal Breeding and Genetics Group

Creating genomic relationship matrices with preGSf90

BLUPF90 TEAM – 05/2024

preGSf90

- Performs Quality Control of SNP information
- Creates the genomic relationship matrix (**G**)
  - and relationships based on pedigree ($\mathbf{A}_{22}$)
  - Inverse of relationship matrices

BLUP-based models

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

BLUP

Henderson, 1963

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

GBLUP

Nejati-Javaremi et al., 1997
Fernando, 1998
VanRaden, 2008

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

ssGBLUP

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

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

Realized relationship matrix (**H**)

Animal	Sire	Dam
1	0	0
2	0	0
3	1	2
4	1	2

$$\mathbf{H} = \begin{pmatrix} var(u_1) & cov(u_1,u_2) \\ cov(u_2,u_1) & var(u_2) \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G}-\mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{pmatrix}$$

Pedigree Relationship Matrix (**A**)

$$\begin{bmatrix} 1.0 & 0.0 & 0.5 & 0.5 \\ . & 1.0 & 0.5 & 0.5 \\ . & . & 1.0 & 0.5 \\ . & . & . & 1.0 \end{bmatrix}$$

Genomic Relationship Matrix (**G**) for animals 3 and 4

$$\begin{bmatrix} 1.0 & 0.52 \\ . & 1.0 \end{bmatrix}$$

Realized Relationship Matrix (**H**)

$$\begin{bmatrix} 1.004 & 0.0 & 0.507 & 0.507 \\ . & 1.004 & 0.507 & 0.507 \\ . & . & 1.0 & 0.52 \\ . & . & . & 1.0 \end{bmatrix}$$

PreGSf90

- Created to construct the matrices used in ssGBLUP

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

$\mathbf{G}$ $\mathbf{G}^{-1}$

$\mathbf{A}_{22}$ $\mathbf{A}_{22}^{-1}$

$\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$

OPTION required to run preGS90

- PreGSF90
  - controled by adding OPTION to the parameter file

```
OPTION SNP_file marker.geno.clean
```

  - Reads:
    - `marker.geno.clean`
    - `marker.geno.clean.XrefID` (created by renumf90)

    - Pedigree file
    - Map file (optional)

PreGSf90

- Efficient methods

$$\mathbf{G} \qquad \mathbf{G}^{-1}$$

$$\mathbf{A}_{22} \qquad \mathbf{A}_{22}^{-1}$$

- Computes statistics for the matrices
  - Means, Var, Min, Max
  - Correlations between diagonals
  - Correlations for off-diagonals
  - Correlations for the full matrices
  - Regression coefficients

Genomic Matrix default options

- $\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}$ (VanRaden, 2008)

- With:
  - **Z** centered using current allele frequencies
    - Current genotyped animals

Genomic Matrix Options

- `OPTION whichG` *`x`*
  - 1: **G**=**ZZ**'/k ; as in VanRaden, 2008 (default)
  - 2: **G**=**ZDZ**'/n ; where D=1/2p(1-p) as in Amin et al. (2007); Leuttenger et al. (2003)
  - 3: As 2 with modification from Yang et al. (2010)
    - Diagonal of **G** is independent of AF

Genomic Matrix Options

- `OPTION whichfreq` *`x`*
  - 0: read from file *freqdata* or other specified name (needs OPTION FreqFile)
  - 1: 0.5
  - 2: current calculated from genotypes (default)

- `OPTION FreqFile` *`file`*
  - Reads allele frequencies from a file if `OPTION whichfreq 0`

Genomic Matrix Options

- `OPTION whichfreqScale` *`x`*
  - 0: read from file *freqdata* or other specified name (needs `OPTION FreqFile`)
  - 1: 0.5
  - 2: current calculated from genotypes (default)

- `OPTION FreqFile` *`file`*
  - Reads allele frequencies from a file if `OPTION whichfreqScale 0`

Genomic Matrix default options

- **Tuning**
  - Adjust **G** to have mean of diagonals and off-diagonals equal to $\mathbf{A}_{22}$

    - Base of GBLUP is *genotyped* animals
    - Base of pedigree is *founders of the pedigree*
    - For SSGBLUP modelled as a mean for genotyped animals
      - $p(\boldsymbol{u}_2) = N(\mathbf{1}\mu, \mathbf{G})$
      - Integrate μ : $\mathbf{G}^* = 11'\lambda + (1 - {}^{\lambda}/_{2})\mathbf{G}$
      - μ = (Genomic base) – (Pedigree base)
      - Vitezica et al. 2011

Genomic Matrix default options

- `OPTION tunedG` *x*
  - 0: no adjustment
  - 1: mean(diag(G))=1, mean(offdiag(G))=0
  - 2: mean(diag(G))=mean(diag(A_{22})), mean(offdiag(G))=mean(offdiag(A_{22})) (default)
  - 3: mean(G)=mean(A_{22})
  - 4: Use Fst adjustment. Powell et al. (2010) & Vitezica et al. (2011)

$$\lambda = \frac{1}{n^2}\left(\sum_i\sum_j \mathbf{A}_{22_{ij}} - \sum_i\sum_j \mathbf{G}_{ij}\right) \qquad \mathbf{G}^* = 11'\lambda + (1 - \lambda/2)\mathbf{G}$$

- 9: arbitrary parameters: specify two additional numbers a and b in a+bG

```
OPTION tunedG 9 a b
```

Genomic Matrix default options

Default: `OPTION tunedG 2`

Chen et al. (2011)
Christensen et al. (2012)

Effect of different genomic relationship matrices on accuracy and scale
C. Y. Chen, I. Misztal, I. Aguilar, A. Legarra and W. M. Muir

J ANIM SCI 2011, 89:2673-2679.
doi: 10.2527/jas.2010-3555 originally published online March 31, 2011

*"This suggests that the optimal **G** should have AvgD and AvgOff close to that of $\mathbf{A}_{22}$. Although similar AvgD – AvgOff in **G** and $\mathbf{A}_{22}$ ensured unbiased estimates of the additive variances, identical AvgOff seemed to remove biases for the EBV of genotyped birds"*

Single-step methods for genomic evaluation in pigs

O.F. Christensen [1], P. Madsen [1], B. Nielsen [2], T. Ostersen [2], G. Su [1]

Forni *et al.* (2011) suggested that *G* should be scaled such that the average of diagonal elements equal the average of diagonal elements of A_{11}, whereas Chen *et al.* (2011) and Vitezica *et al.* (2011) suggested that a small number should be added to all elements of *G* such that the average of all elements equal the average of elements of A_{11}. Here, we combined these two ideas and adjusted *G* to

$$\mathbf{G_a} = \beta \mathbf{G} + \alpha, \qquad (4)$$

where β and α solved the system of equations

$$\text{Avg}(\text{diag}(G))\beta + \alpha = \text{Avg}(\text{diag}(A_{11})),$$
$$\text{Avg}(G)\beta + \alpha = \text{Avg}(A_{11}).$$

Genomic Matrix default options

- **Blending** - to avoid singularity problems

$$\mathbf{G} = 0.95*\mathbf{G}_0 + 0.05*\mathbf{A}_{22}$$

- `OPTION AlphaBeta 0.95 0.05` #(default)

- Beta may vary from 0.01 to 0.3
  - Some places may use 0.5

Genomic Matrix options

- `OPTION GammaDelta` *`x1 x2`*

$$\mathbf{G} = \alpha\mathbf{G}_0 + \beta\mathbf{A}_{22} + \gamma\mathbf{I} + \delta$$

- Objective: blend 95% of **G** with 5% identity instead of $\mathbf{A}_{22}$

$$\mathbf{G} = 0.95\mathbf{G}_0 + 0.0\mathbf{A}_{22} + 0.05\mathbf{I} + 0.0$$

- `OPTION AlphaBeta 0.95 0.0` #default = 0.95 0.05
- `OPTION GammaDelta 0.05 0.0` #default = 0.0 0.0

Order of procedures

Tuning → Blending

McWhorter et al. (2022)

Storing and Reading Matrices

- preGSf90 saves $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$ by default (file: GimA22i)

To save the 'raw' genomic matrix:

- `OPTION saveG  [all]`
  - If the optional *all* is present all intermediate **G** matrices will be saved!!!

To save $\mathbf{G}^{-1}$

- `OPTION saveGInverse`
  - Only the final **G**, after blending, scaling, etc. is inverted!!!

To save $\mathbf{A}_{22}$ and $\mathbf{A}_{22}^{-1}$

- `OPTION saveA22` and `OPTION saveA22Inverse`

Storing and Reading Matrices

- `OPTION saveG  [all] ,OPTION saveGInverse`, …

  - Saves in binary format
  - “Dumped” format to save space and time
  - To save as row, column, value:
    - `OPTION no_full_binary`
    - Still binary, but can be easily read and converted to text

Storing with Original IDs

- Some matrices could be stored in text files with the original IDs extracted from *renaddxx.ped* created by the RENUMF90 program (col #10)
- For example:
  - `OPTION saveGOrig`
  - `OPTION saveDiagGOrig`
  - `OPTION saveHinvOrig`
- Values
  - origID_i, origID_j, val

Genomic Matrix - Population structure

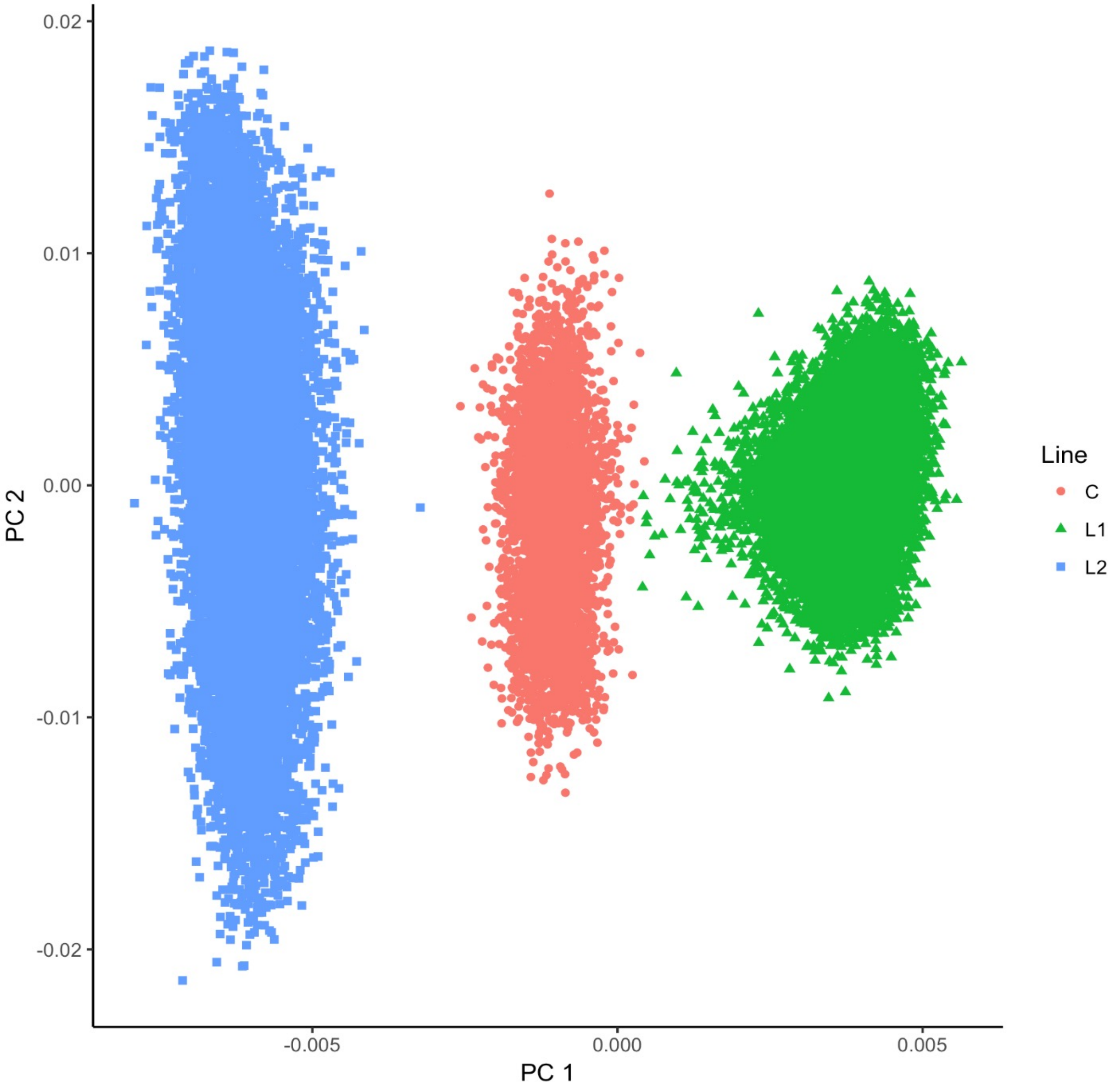
```
OPTION plotpca
```

Plot first two principal components to look for stratification in the population.

```
OPTION extra_info_pca file col
```

Reads from *file* the column *col* to plot with different colors for different classes.

Genomic Matrix - Population structure



Tricks to setup **G** for GBLUP #1

- Tricks are needed because preGSf90 is set up for ssGBLUP

1) Use a dummy pedigree

```
1 0 0
2 0 0
…
```

2) Use PED_DEPTH 1 in renumf90

3) Change blending parameters

- `OPTION AlphaBeta 1.00 0.00` → G = 1.00***G** + 0.00***I**
- `OPTION AlphaBeta 0.95 0.05` → G = 0.95***G** + 0.05***I**

4) No adjustment for compatibility with $\mathbf{A}_{22}$

- `OPTION tunedG 0`

Tricks to setup **G** for GBLUP #2

- Yet another ways to run GBLUP in BLUPF90
- Replace steps 1 and 2 by:

A) In renum.par, remove any information about the pedigree file

```
FILE
pedigree.txt
FILE_POS
1 2 3 0 0
PED_DEPTH
3
```

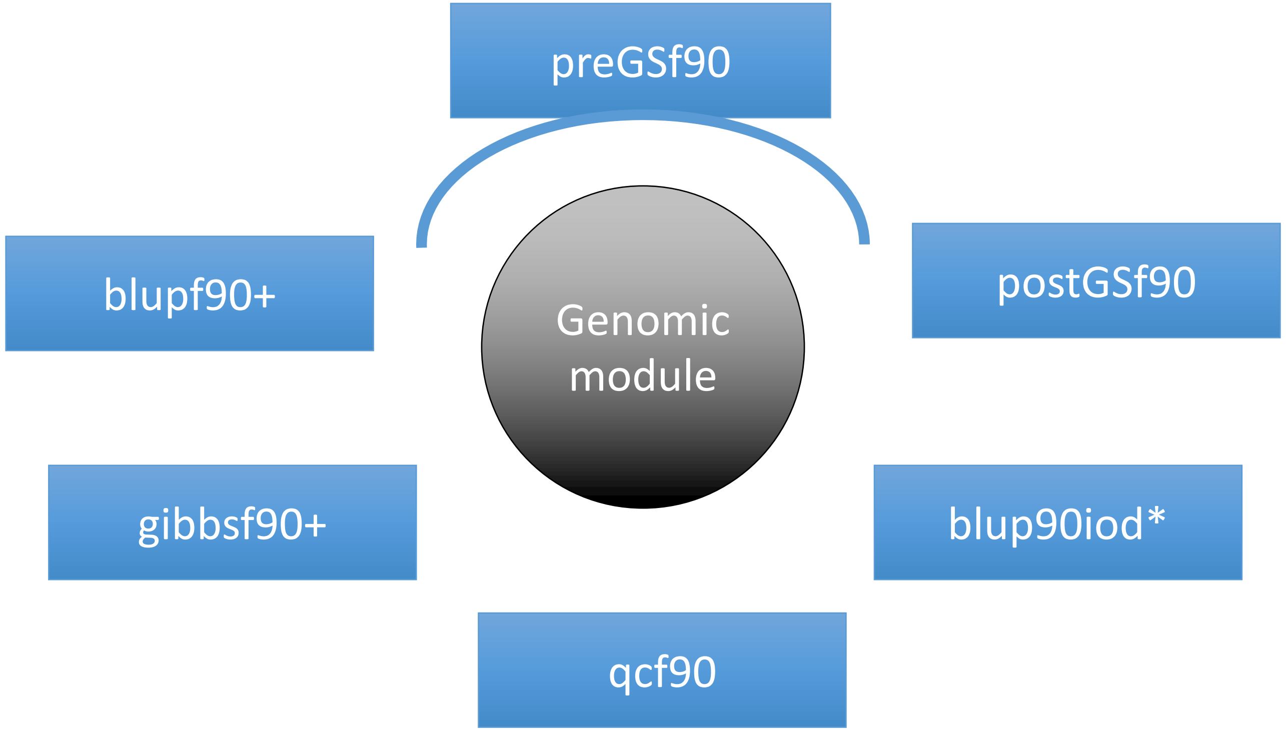
OR

B) Add this option to the renf90.par parameter file:

```
OPTION omit_ainv
```

PreGSf90 inside BLUPF90 ??

- Almost all programs from BLUPF90 support creating genomic relationship matrices
    - `OPTION SNP_file xxxx`



- When to use preGSF90 ?
    - Same genomic relationship matrix for several models, traits, etc.
    - Just do it once and store GimA22i or Gi and A22i separate

Use in application programs

- Use renumf90 for renumbering and creating XrefID and other files

```
SNP_FILE
marker.geno
```

- Option 1:

  run blupf90+

- Option 2:

  run preGSf90 with quality control, saving clean files
  run blupf90+ with clean files

- Option 3:

  run preGSf90 (program saves **GimA22i**)
  run blupf90+ with option to read **GimA22i**

preGSf90 is highly parallelized!

```
OPTION num_threads_pregs n
```

Specify number of threads to be used with MKL-OpenMP for creation and inversion of matrices

Be careful: It has advantages and disadvantages!