UNIVERSITY OF GEORGIA

Creating genomic relationship matrices with preGSf90

BLUPF90 TEAM, 06/2022

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

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

BLUP

Henderson, 1963

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

GBLUP

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

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

ssGBLUP

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

$$\mathbf{H}^{-1} = \left[\begin{array}{c|c} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \hline \mathbf{A}^{21} & \mathbf{A}^{22} \end{array}\right] + \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{array}\right]$$

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

PreGSf90

- Created to construct the matrices used in ssGBLUP

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

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

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

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

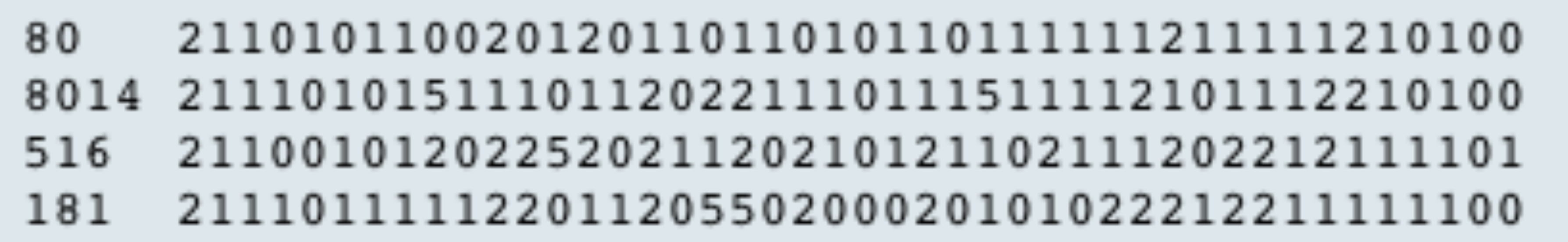
Genomic Relationship Matrix - **G**

- $\mathbf{G} = \dfrac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)}$ (VanRaden, 2008)
  - **Z** = matrix for SNP marker
  - Dimension of **Z** = $n*i$
  - *n* animals
  - *i* markers

Genotype Codes

0 – Homozygous
1 – Heterozygous
2 – Homozygous
5 – No Call (Missing)

SNP file



```
80    21101011002012011011010110111111211111210100
8014  21110101511101120221110111511112101112210100
516   21100101202252021120210121102111202212111101
181   21110111112201120550200020101022212211111100
```

HOW TO: Create **G**

- Read SNP marker information => **M**

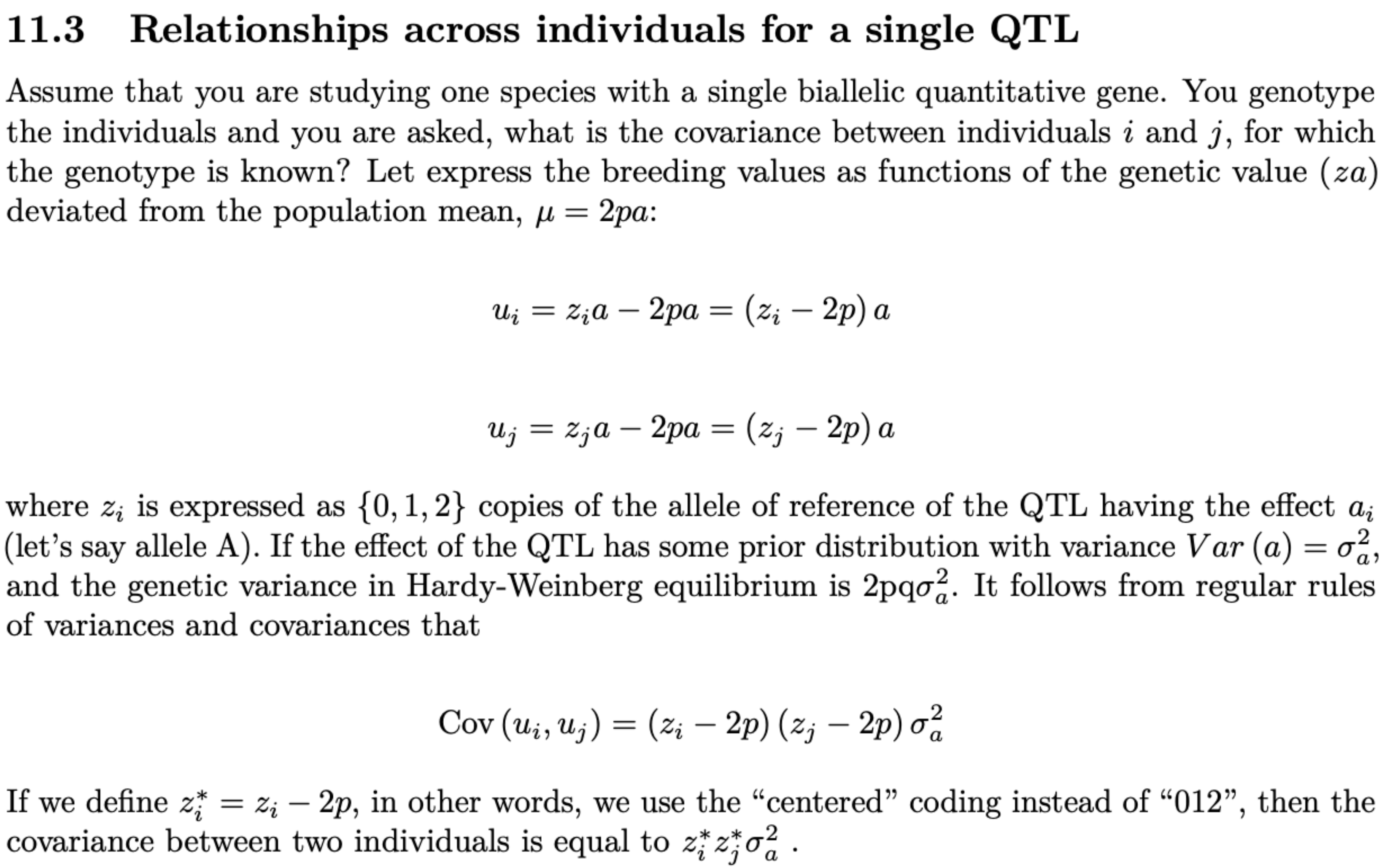
$$\begin{bmatrix} 2 & 1 & 2 & .. \\ 0 & 1 & 0 & .. \\ .. & .. & .. & .. \end{bmatrix}.$$

- Get ‘means’ to center
  - Calculate allele frequency from observed genotypes (p_i)
  - p_i= sum(SNPcode$_i$)/2n

- Centered matrix $\mathbf{Z} = \mathbf{M} - 2\mathbf{P}$

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

Why to center **G**?



11.3 Relationships across individuals for a single QTL

Assume that you are studying one species with a single biallelic quantitative gene. You genotype the individuals and you are asked, what is the covariance between individuals i and j, for which the genotype is known? Let express the breeding values as functions of the genetic value (za) deviated from the population mean, $\mu = 2pa$:

$$u_i = z_i a - 2pa = (z_i - 2p)\, a$$

$$u_j = z_j a - 2pa = (z_j - 2p)\, a$$

where z_i is expressed as $\{0, 1, 2\}$ copies of the allele of reference of the QTL having the effect a_i (let's say allele A). If the effect of the QTL has some prior distribution with variance $Var\,(a) = \sigma_a^2$, and the genetic variance in Hardy-Weinberg equilibrium is 2pqσ_a^2. It follows from regular rules of variances and covariances that

$$\mathrm{Cov}\,(u_i, u_j) = (z_i - 2p)\,(z_j - 2p)\,\sigma_a^2$$

If we define $z_i^* = z_i - 2p$, in other words, we use the "centered" coding instead of "012", then the covariance between two individuals is equal to $z_i^* z_j^* \sigma_a^2$.

Bases for Genomic Predictions – Page 60

Why to scale **G**?

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

Creating **G**

- Issues
  - Large number of genotyped individuals
  - Large number of SNP markers
  - Matrix multiplication ~ cost $n^2 * i$

- Large amounts of data put in (cache) memory to do matrix multiplication for each pair of animals and indirect memory access (center)

PreGSf90

- Efficient methods
  - create the genomic relationship matrix and the relationship matrix based on pedigree
  - Invert the relationship matrices

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

OPTIONS – preGS90 parameter file

- PreGSF90
  - controled by adding OPTION commands 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)

Genomic Matrix default options

- $\mathbf{G}_0 = \dfrac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}$ (VanRaden, 2008)

- With:
  - **Z** centered and scaled 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 UAR 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

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

Adjusting **G** to $\mathbf{A}_{22}$

- **Tuning**

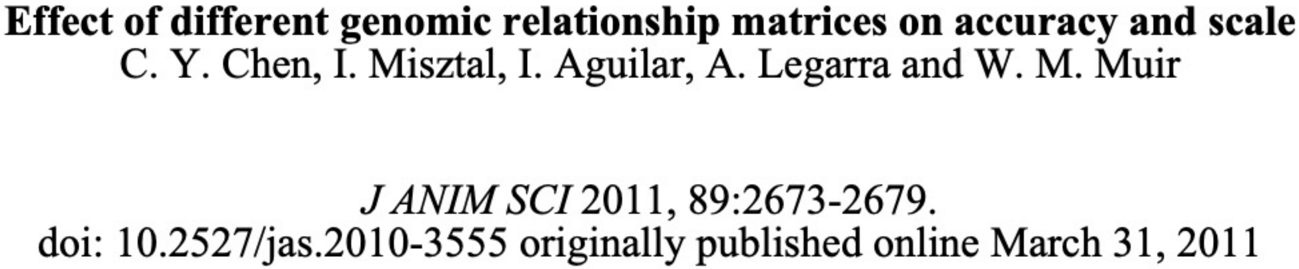
– Adjust **G** to have same mean diagonal and off-diagonal as $\mathbf{A}_{22}$

- 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 2`   #(default)

– suggested by Chen et al. (2011)



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 **A**$_{22}$. Although similar AvgD − AvgOff in **G** and **A**$_{22}$ ensured unbiased estimates of the additive variances, identical AvgOff seemed to remove biases for the EBV of genotyped birds"*

Options for matching **G** to $\mathbf{A}_{22}$

- `OPTION tunedG` *`x`*
  - 0: no adjustment
  - 1: mean(diag(**G**))=1, mean(offdiag(**G**))=0
  - 2: mean(diag(**G**))=mean(diag($\mathbf{A}_{22}$)), mean(offdiag(**G**))=mean(offdiag($\mathbf{A}_{22}$)) (default)
  - 3: mean(**G**)=mean($\mathbf{A}_{22}$)
  - 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*+*b*G

```
OPTION tunedG 9 a b
```

Genomic Matrix default options

- **Blending** - to avoid singularity problems

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

  – `OPTION AlphaBeta 0.95 0.05` #(default)

  – Beta may vary from 0.2 to 0.01

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 $\mathbf{A}_{22}$

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

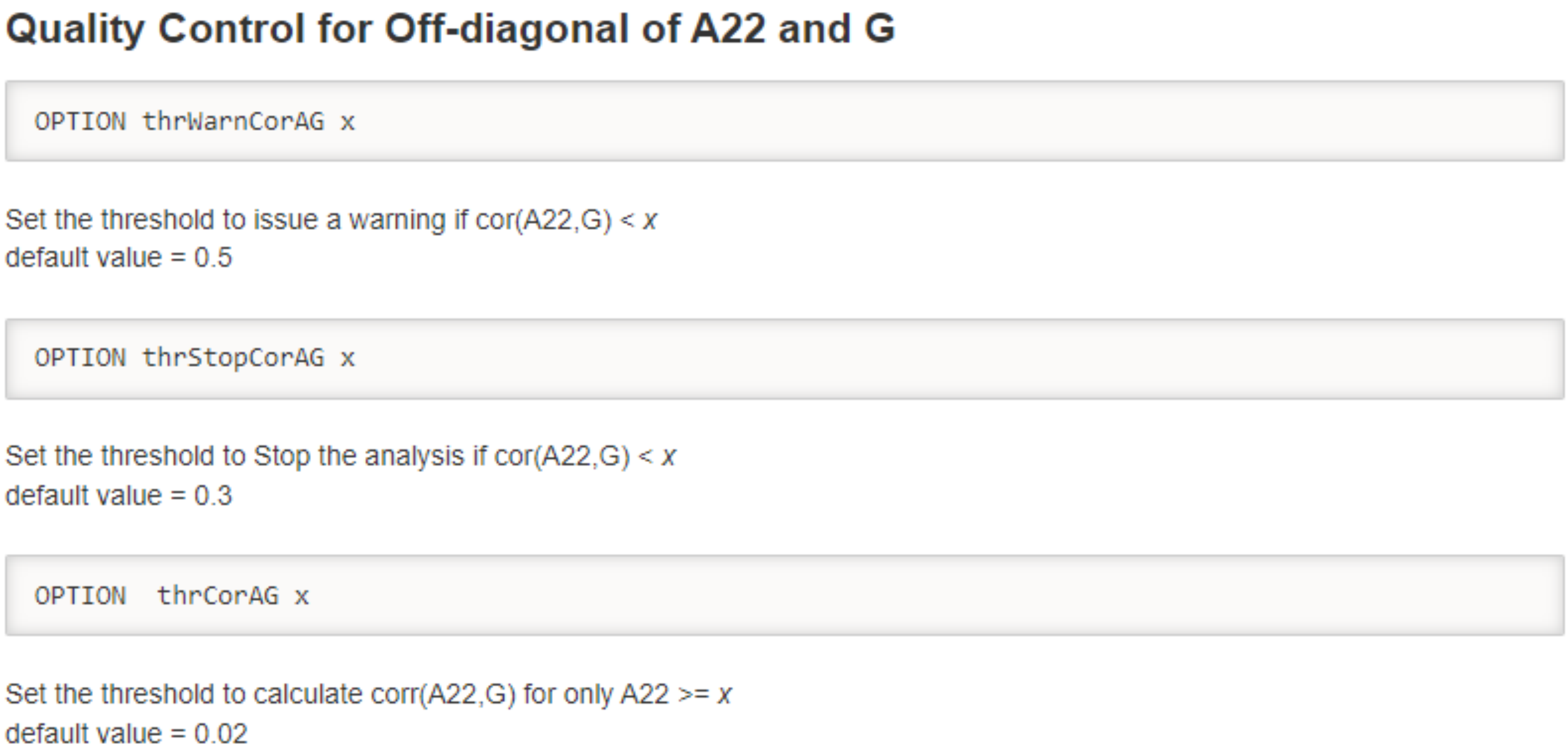
- `OPTION AlphaBeta 0.95 0.0` #default = 0.95 0.05
- `OPTION GammaDelta 0.05 0.0` #default = 0.0 0.0

Order of procedures

Tuning → Blending

McWhorter et al. (2022)

Quality control for off-diag of **G** to **A**$_{22}$



Quality Control for Off-diagonal of A22 and G

```
OPTION thrWarnCorAG x
```

Set the threshold to issue a warning if cor(A22,G) < *x*
default value = 0.5

```
OPTION thrStopCorAG x
```

Set the threshold to Stop the analysis if cor(A22,G) < *x*
default value = 0.3

```
OPTION  thrCorAG x
```

Set the threshold to calculate corr(A22,G) for only A22 >= *x*
default value = 0.02

Storing and Reading Matrices

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

To save 'raw' genomic matrix:

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

To save **G**$^{-1}$

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

To save **A**$_{22}$ and inverse

- `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

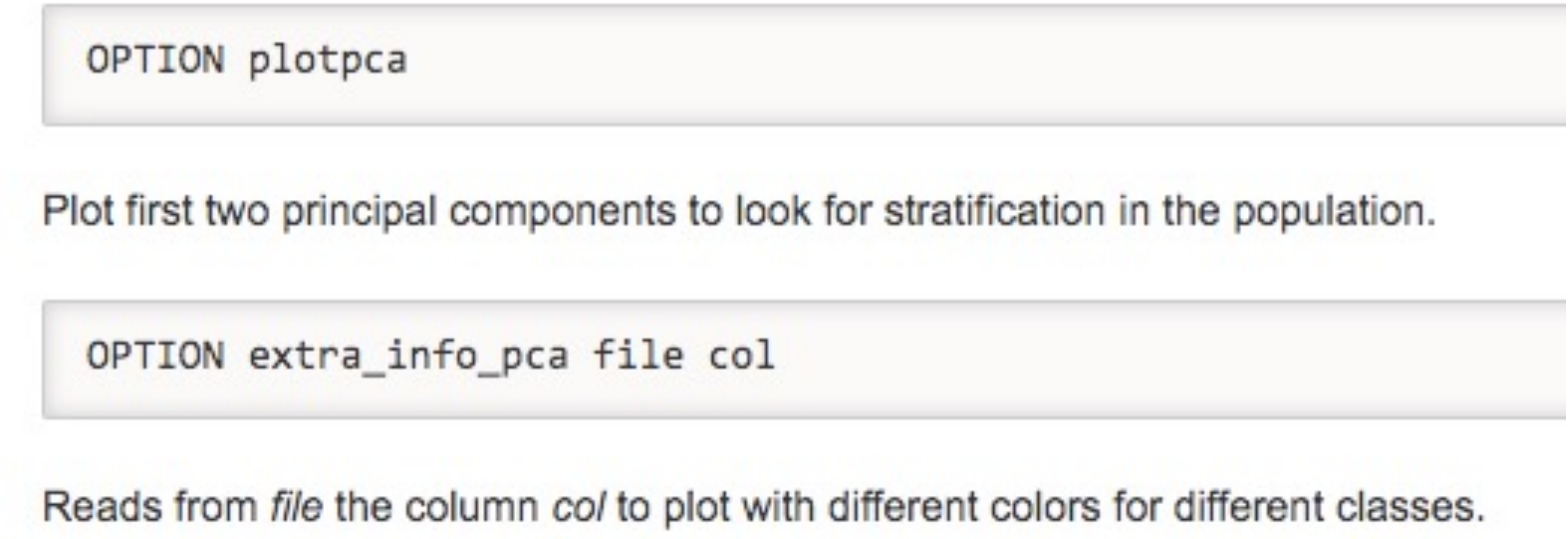
- Some matrices can be stored in text files with the original IDs extracted from *renaddxx.ped* created by the RENUMF90 program (col #10)

- For example:
  - `OPTION saveGOrig`
  - `OPTION saveDiagGOrig`
  - `OPTION saveHinvOrig`

- Values
  - origID_i, origID_j, val

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

Genomic Matrix - Population structure



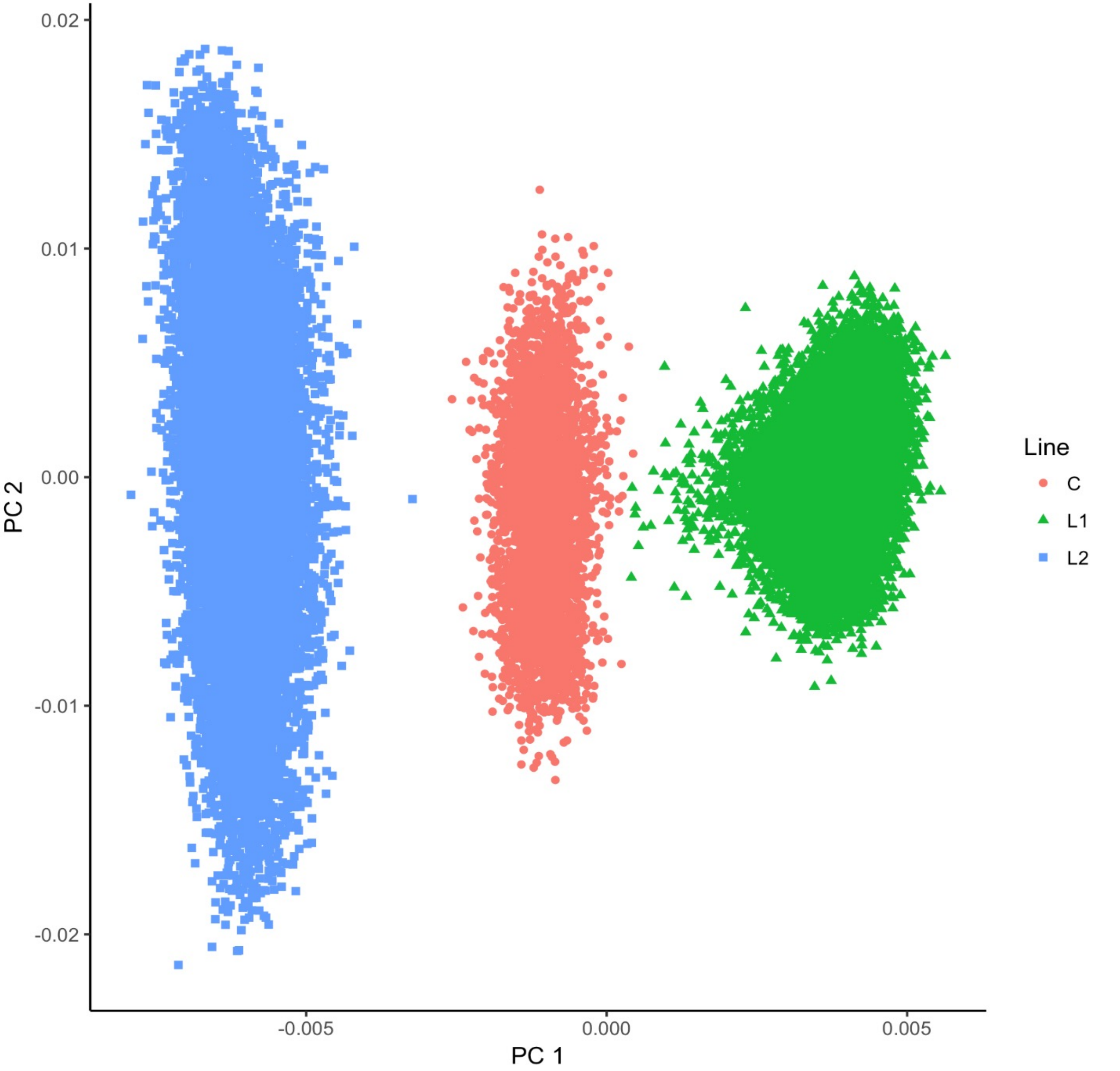
```
OPTION plotpca
```

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



PreGSf90 inside BLUPF90 ??

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

- Why preGSF90 ?
  - Same **G** for several models, traits, etc.
  - Just do it once and store GimA22i or Gi and A22i

Use in application programs

- Use renumf90 for renumbering and creating XrefID and other files

```
SNP_FILE
marker.geno
```

- Option 1:

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

- Option 2:

  run blupf90+

- Option 3:

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

Tricks to setup **G** for GBLUP

- 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` → $G = 1.00*\mathbf{G} + 0.00*\mathbf{I}$
- `OPTION AlphaBeta 0.99 0.00`
- `OPTION GammaDelta 0.01 0.00`

→ $G = 0.99*\mathbf{G} + 0.01*\mathbf{I}$

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

- `OPTION tunedG 0`

Tricks to setup **G** for GBLUP

- 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

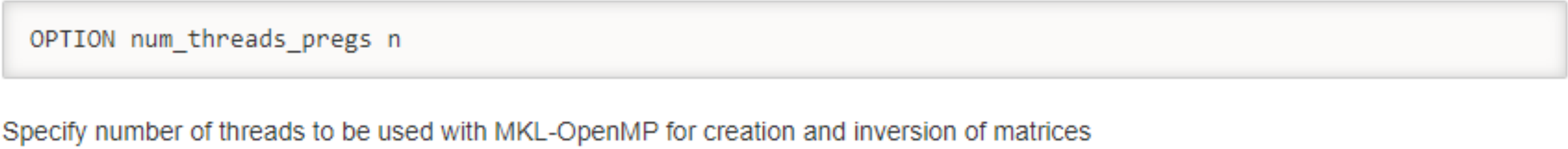
```
FILE
pedigree.txt
FILE_POS
1 2 3 0 0
PED_DEPTH
3
```

OR

B) Add this option to the parameter file:

```
- OPTION omit_ainv
```

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!