



UNIVERSITY OF
GEORGIA

Creation of genomic relationship matrices with preGSf90

Ignacio Aguilar
INIA Uruguay

Daniela Lourenco
UGA USA

Genomic Relationship Matrix - G

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

- Z = matrix for SNP marker

- Dimension of Z = n * p

- n animals

- p 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: Creation of Genomic Matrix

- Read SNP marker information => **M**
$$\begin{bmatrix} 2 & 1 & 2 & \dots \\ 0 & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$
- Get 'means' to center
 - Calculate allele frequency from observed genotypes (p_i)
 - $p_i = \text{sum}(\text{SNPcode}_i) / 2n$
- Centered matrix **Z** = **M** – 2**P**
- **G** =
$$\frac{\mathbf{ZZ}'}{2 \sum p_i(1-p_i)}$$
 (VanRaden, 2008)

Creation of Genomic matrix

- Issues
 - Large number of genotyped individuals
 - Large number of SNP markers
 - Matrix multiplication $\sim \text{cost } n^2 * p$
- Large amount of data put in (cache) memory to do matrix multiplication for each pair of animals and indirect memory access (center)

PreGSf90

- Interface program to the genomic module to process the genomic information for the BLUPF90 family of programs
- Besides Quality Control of SNP information:
- Efficient methods
 - creation of the genomic relationship matrix, relationship based on pedigree
 - Inverse of relationship matrices

PreGSf90

- Created to construct the matrices using 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}

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

PreGSf90

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

Input files

- Same parameter file as for all BLUPf90 programs
 - But with “OPTION SNP_file xxxx”
 - Turns on the genomic module
- SNP file (marker information)
- Cross Reference file for renumber ID
 - Links genotypes files with codes in pedigree, etc.
- Pedigree file
- Map file (optional)

OPTIONS – preGS90 parameter file

- PreGSF90
 - controled by adding OPTION commands to the parameter file
 - `OPTION SNP_file marker.geno.clean`
 - Read 2 files:
 - `marker.geno.clean`
 - `marker.geno.clean.XrefID` (created by renumf90)

Genomic Matrix default options

- $\mathbf{G}^* = \frac{\mathbf{Z}\mathbf{Z}'}{2 \sum p_i(1-p_i)}$ (VanRaden, 2008)
- With:
 - \mathbf{Z} centered using allele frequencies estimated from SNP
- $\mathbf{G} = \mathbf{G}^*0.95 + \mathbf{A}_{22}*0.05$ (to invert)
- Tuning of \mathbf{G} (see Vitezica et al., 2011)
 - Adjust \mathbf{G} to have mean of diagonals and off-diagonals equal to \mathbf{A}_{22}

Genomic Matrix Options

- OPTION whichfreq *x*
 - 0: read from file *freqdata* or other specified
 - 1: 0.5
 - 2: current calculated from genotypes (default)
- OPTION FreqFile *file*
 - Reads allele frequencies from a file
- OPTION maxsnp *x*
 - Set the maximum length of string for reading marker data from file => BovineHD chip

Options for Blending **G** and **A**₂₂

- OPTION AlphaBeta *alpha beta*
 - $G = \alpha * G + \beta * A$
- OPTION tunedG *x*
 - 0: no adjustment
 - 1: $\text{mean}(\text{diag}(G))=1$, $\text{mean}(\text{offdiag}(G))=0$
 - 2: $\text{mean}(\text{diag}(G))=\text{mean}(\text{diag}(A))$,
 $\text{mean}(\text{offdiag}(G))=\text{mean}(\text{offdiag}(A))$ (default)
 - 3: $\text{mean}(G)=\text{mean}(A)$
 - 4: Use Fst adjustment. Powell et al. (2010) & Vitezica et al. (2011)

$$\rho = \frac{1}{n^2} (\sum_i \sum_j A_{22\ ij} - \sum_i \sum_j G_{ij})$$

$$G^* = (1 - \rho / 2) G + \mathbf{1}\mathbf{1}' \rho$$

Storing and Reading Matrices

To save our 'raw' genomic matrix:

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

or its inverse

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

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

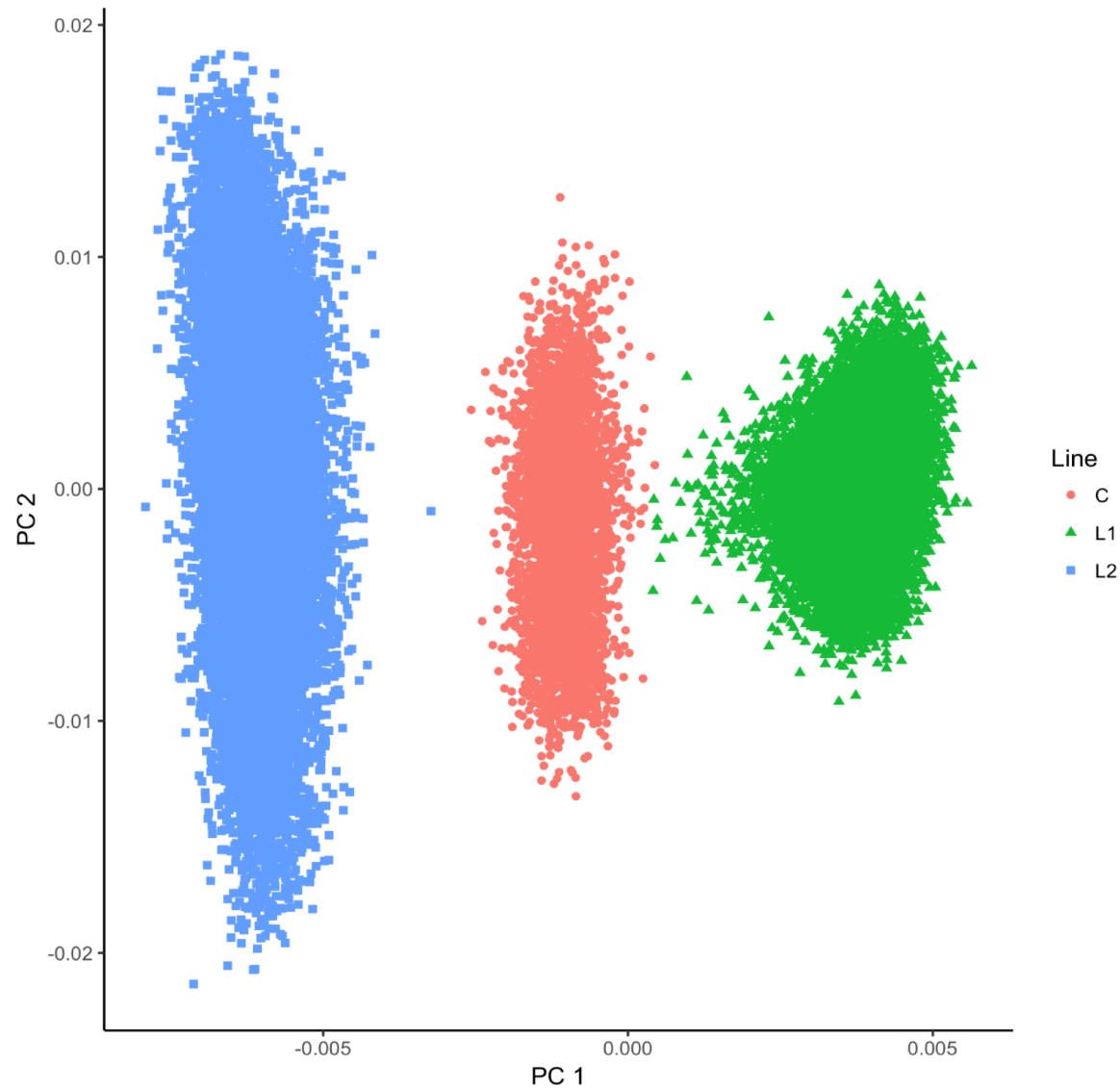
```
OPTION plotpca
```

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



Creation of 'raw' genomic matrix 'GBLUP'

- Tricks:
- Use dummy pedigree

```
1 0 0
2 0 0
...
```
- Change blending parameters
 - OPTION AlphaBeta 0.99 0.01
- No adjustment for compatibility with \mathbf{A}_{22}
 - OPTION tunedG 0

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

PreGSf90 inside BLUPF90 ??

- Almost all programs from BLUPF90 support creation of genomic relationship matrices
- `OPTION SNP_file xxxx`
- Why preGSF90 ?
 - Same genomic relationship matrix for several models, traits, etc. Just do it once and store GimA22i

Use in application programs

- Use renumf90 for renumbering and creation of XrefID and files
SNP_FILE
marker.geno
- Run preGSf90 with quality control, saving clean files
 - Option 1:
run preGSf90 with clean files (program saves **GimA22i**)
run blupf90 with option to read **GimA22i** from the file
 - Option 2:
run blupf90 with clean files