



UNIVERSITY OF  
**GEORGIA**

# Creation of genomic relationship matrices with preGSf90

Ignacio Aguilar  
INIA Uruguay

Daniela Lourenco  
UGA USA

# Genomic Relationship Matrix - G

- $G = ZZ'/k$

- Z = matrix for SNP marker

- Dimension Z =  $n \times p$

- n animals,

- p markers

## Genotype Codes


0 – Homozygous

1 – Heterozygous

2 – Homozygous

5 – No Call (Missing)

Data file with SNP marker



80	21101011002012011011010110111111211111210100
8014	21110101511101120221110111511112101112210100
516	21100101202252021120210121102111202212111101
181	21110111112201120550200020101022212211111100

# HOWTO: 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
- Efficient methods
  - creation of the genomic relationship matrix, relationship based on pedigree
  - Inverse of relationship matrices
- Performs Quality Control of SNP information

# 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)
- Pedigree file
- Cross Reference file for renumber ID
  - Links genotypes files with codes in pedigree, etc.
- Map file (optional)

# OPTIONS – BLUPF90 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`

# RENUMF90

- Add keyword to the “animal effect”  
SNP\_FILE  
marker.geno
- Renumber tool to prepare:
  - data
  - pedigree
  - genotypes
  - parameter files for all other BLUPF90 programs
  - Check wiki:
- <http://nce.ads.uga.edu/wiki/doku.php>



# Parameters file

RENUMF90  
renum.par

```
DATAFILE
phenotypes.txt
TRAITS
3
FIELDS_PASSED TO OUTPUT

WEIGHT(S)

RESIDUAL_VARIANCE
0.9038
EFFECT
1 cross alpha # mu
EFFECT
2 cross alpha # animal
RANDOM
animal
FILE
pedigree
SNP_FILE
marker.geno.clean
(CO)VARIANCES
0.9951E-01
```

BLUPF90  
renf90.par

```
DATAFILE
renf90.dat
NUMBER_OF_TRAITS
1
NUMBER_OF_EFFECTS
2
OBSERVATION(S)
1
WEIGHT(S)

EFFECTS: POSITIONS_IN_DATAFILE NUMBE
2 1 cross
3 15800 cross
RANDOM_RESIDUAL VALUES
0.90380
RANDOM_GROUP
2
RANDOM_TYPE
add_animal
FILE
renadd02.ped
(CO)VARIANCES
0.99510E-01
OPTION SNP_file marker.geno.clean
```

# Pedigree file from RENUMF90

- 1 - **animal number**
- 2 - parent 1 number or UPG
- 3 - parent 2 number or UPG
- 4 - 3 minus number of known parents
- 5 - known or estimated year of birth
- 6 - number of known parents;  
    **if animal is genotyped 10 + number of known parents**
- 7 - number of records
- 8 - number of progenies as parent 1
- 9 - number of progenies as parent 2
- **10 - original animal ID**

# SNP file & Cross Reference Id

## SNP File

First col: Identification, could be alphanumeric

Second col: SNP markers {codes: 0,1,2 and 5 for missing}

```
80 211010110020120110110101101111
8014 211101015111011202211101115111
516 211001012022520211202101211021
181 211101111122011205502000201010
```

## Cross Reference ID

```
1732 80
8474 8014
406 516
9441 181
```

Renumber ID

## Pedigree File (from RENUMF90)

```
1732 11010 10584 1 3 12 1 0 0 80
8474 8691 9908 1 3 12 1 0 0 8014
406 8691 9825 1 3 12 1 0 2 516
9441 8691 8829 1 3 12 1 0 0 181
```

Original ID

# 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**<sub>22</sub>

- OPTION AlphaBeta alpha beta
  - $G = \alpha * G + \beta * A$
- OPTION tunedG
  - 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)

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

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

# 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

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

# 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

# Looking stored matrices

- Avoid open with text editors, huge files !!!
- For example:
- 1500 genotyped individuals => 1,125,750 rows
- Inspection could be done by Unix commands:
  - `head G` => first 10 lines
  - `tail G` => last 10 lines
  - `less G` => scroll document by line/page
  - `wc -l G` => count number of lines

good for checks with the number of  
genotypes  $(n) = (n*(n+1))/2$

# head G

---

1	1	.999382118619
1	2	.355052761478
2	2	1.014521277458
1	3	-.048184197960
2	3	-.057513012886
3	3	.976558921904
1	4	-.101734083083
2	4	-.007644724611
3	4	.196757165096
4	4	1.018165021903

# 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

# 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 saving **G**  
run blupf90 with option to read **G** from the file
  - Option 2:  
run blupf90 with clean files saving **G**