

# Quality Control of SNP data and creation of genomic matrices with BLUPF90 software

**Daniela Lourenco**  
BLUPF90 TEAM – 03/2023



**UNIVERSITY OF  
GEORGIA**

**College of Agricultural &  
Environmental Sciences**

*Animal Breeding and  
Genetics Group*

# SNP data

SNP

ANIMAL

025	110101110511110111110010001221151205122125022511110250122010201021000221121025000122010
036	21101101022012122222012101222010120222111112021222111112102020101101020111112011012110:
050	121010021112021111200021212222100021122122122110000020220000211022122212122020001112020:
054	120001200220121211100121002222110211221102011212221200220021212121111202112022002022100:
066	200002020221021221120022001222211101220202110202222020220001222121011201021022010011010:
097	101102120220121122111021001111100102211212022111111020221001201222012111021021021012000:
101	121002120220011221100011112220100101120112121211121201221002102002021211222022010022110:
151	111001020221220210201011012220200121221111221221121111222002201112011212111022000022012:
172	211012020211112101211021102220101001221212221102220201221020212112010211122022112011010:
224	220001110221012210101021102520201112120222122212220110121011102220050210121022010022125:
277	210102200121221211212021012222002012210212110201121021221002211011020211021112021012010:
314	122011120122220210210010002121001120120202001210020021210011201022021212111022010101100:
419	221112210121120222221022102110201021121211122000000111220002211122020222112120012121110:
439	200202100122121210101021012221101112220202022110010111210011201022012220211021010011020:
456	12000102022111220010102100221100020222121222220010110221110212012011212211111102112010:
501	111000021221121201212121002221101202222101022112222110220011202110020201102022100021020:
571	110000120202200221212022001210200011122110110222221200220020212001010212121022102010110:
579	11210021021001010111102200222120002221111202022222110222101202012111222111112011011020:
581	21100202152100122120201100220020112512121502252222250221011201121051202222112111012110:
657	110011120220111211101020012221000112221212021211121200220012202220022212212112001112011:
660	210002120221120221121021012221011012221222121211120201221012201121111211112022000012101:
730	210002020220020222220012002220001220222220021102252200122001202111151001012022001012025:
732	212102121521002201200012101121201215110215122521121150220011102111050202221122011022010:
764	111102121520012212211020001220201225222115021522221150220110202120050202022022111112110:
780	121101021122220210101022002221201201121221012111110111221020202001010112212121002021021:
800	22100012022122221020202110222110101211202212022222200221002211121021202011022010111010:
816	110001220220121220110022011121100011021122121220020112222002222111021111212022011022010:
832	121010011120011211110021112220111112122221210201111020221002112221001212111121012111110:
900	210100110220122121211021102121012120221212121101111110221001202121110211011021100022020:
901	121001020221121212210010002120201111221112122001111110221002201022012212121021000012020:

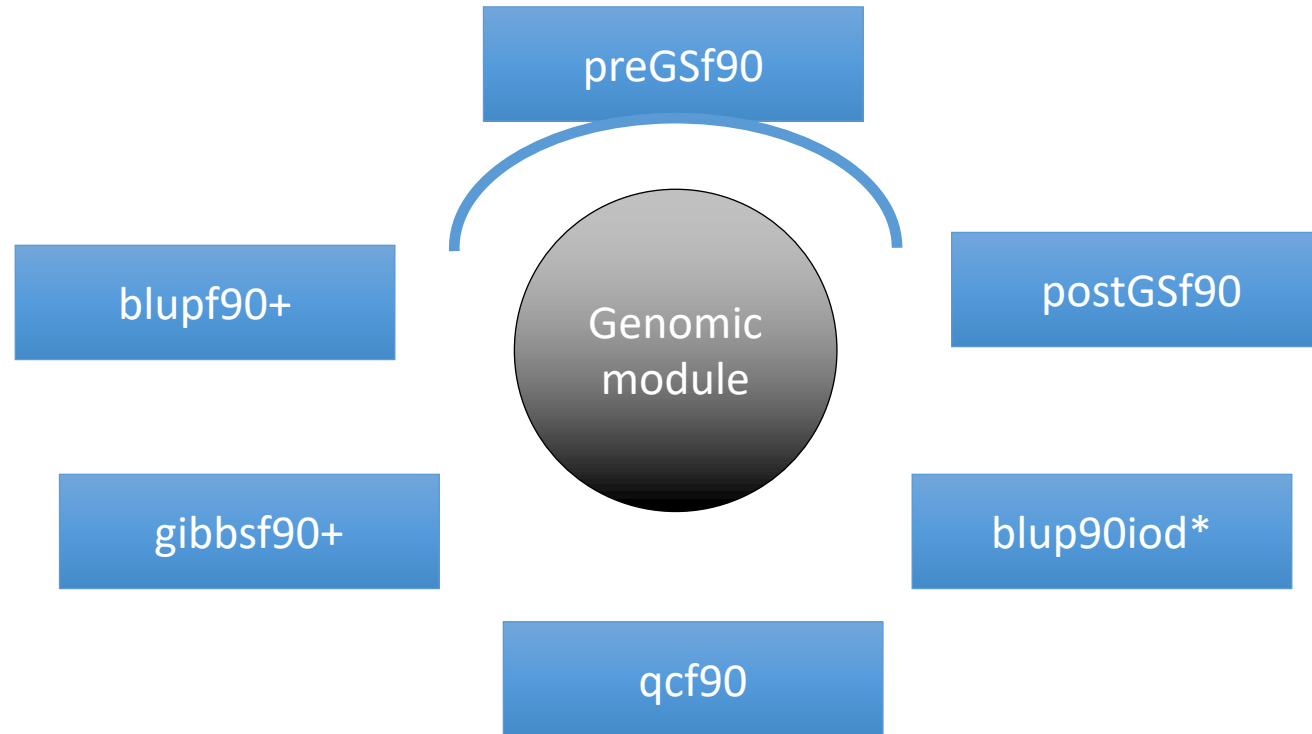
# Quality control

Which software in the  
BLUPF90 family?

- Call rate
  - Animals
  - SNP
- Minor Allele Frequency (MAF)
- Hardy-Weinberg Equilibrium (HWE)
- Non-mapped SNP
- Mendelian Conflicts
- Duplicate genotypes
- Linkage disequilibrium (LD)

# preGSf90

- Interface program to the genomic module to process the genomic information in the BLUPF90 family of programs



# preGSf90

- Performs Quality Control of SNP information
- Creates the genomic relationship matrix (**G**)
  - and relationships based on pedigree (**A**<sub>22</sub>)
  - Inverse of relationship matrices



# preGSf90

- Same parameter file as for all BLUPF90 programs
- Needs an extra OPTION in renf90.par
  - `OPTION SNP_file marker.geno`
- Reads 2 extra files (besides data and pedigree):
  - `marker.geno`
  - `marker.geno_XrefID` (created by renumf90)

`_XrefID` has 2 columns: Renumbered ID Original ID

# Run renumf90 before preGSf90

- Use renumf90 for renumbering data and creating XrefID and files

```
EFFECT
1  cross alpha
RANDOM
animal
FILE
ped3.txt
FILE_POS
1 2 3 0 0
SNP_FILE
marker.geno
PED_DEPTH
0
(CO) VARIANCES
0.30
```

# Parameter files

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
(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.
```



# New pedigree file from RENUMF90

- 1 - **renumbered animal ID**
- 2 - parent 1 number or UPG
- 3 - parent 2 number or UPG
- 4 - 3 minus number of known parents or inbreeding code
- 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, XrefID, and ped from renumf90

## SNP File

First col: original ID

Second col: SNP genotypes {codes: 0,1,2, and 5 (missing)}

All SNP should start in the same column!!!

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

No changes!!!

Renumbered ID

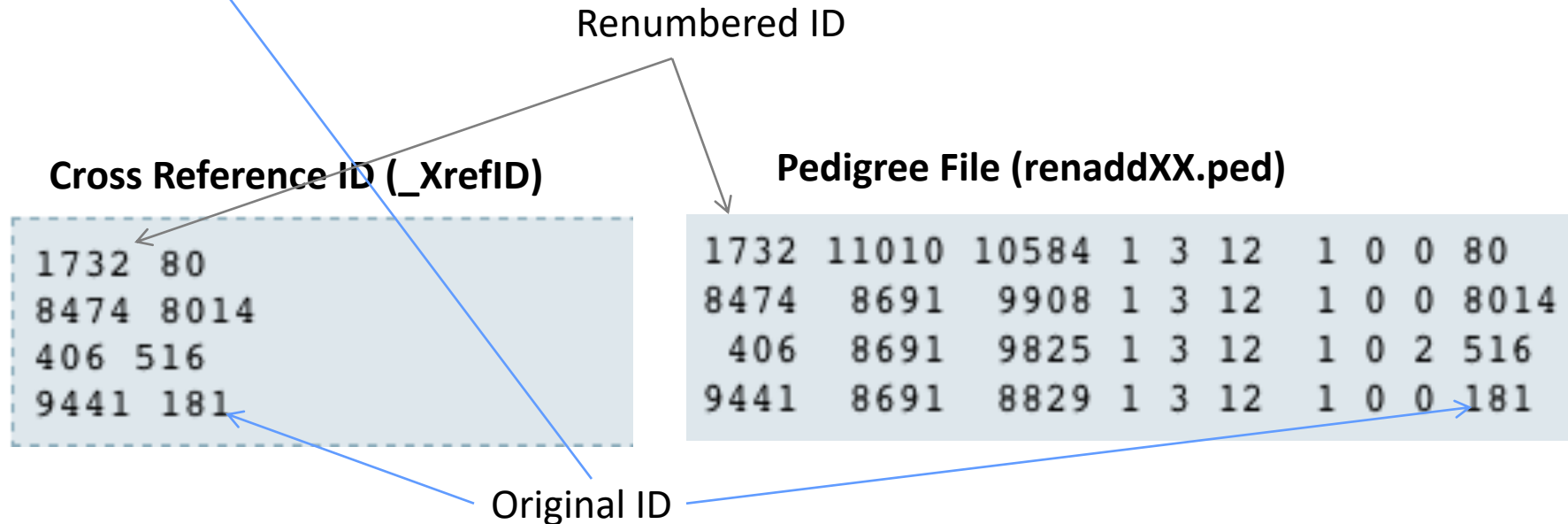
Cross Reference ID (\_XrefID)

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

Pedigree File (renaddXX.ped)

```
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



# preGSf90

- Same parameter file as for all BLUPF90 programs
- Needs an extra OPTION in renf90.par
  - `OPTION SNP_file marker.geno`
- Reads 2 extra files (besides data and pedigree):
  - `marker.geno`
  - `marker.geno_XrefID` (created by renumf90)

`_XrefID` has 2 columns: Renumbered ID Original ID

# SNP map file – new default

- `OPTION chrinfo <file>`
- `OPTION map_file <file>`
  - For GWAS and QC
- Format:
  - A header must be provided
    - Names for SNP, chromosome, and physical position are mandatory
  - SNPID for SNP
  - CHR for chromosome
  - POS for position

```
NUM CHR   POS      SNPID      NUM2
31428 14 7928189 ARS-BFGL-BAC-1020 2
32005 14 31819743 ARS-BFGL-BAC-10245 3
31371 14 6133529 ARS-BFGL-BAC-10345 4
31679 14 17544926 ARS-BFGL-BAC-10591 7
32053 14 34639444 ARS-BFGL-BAC-10867 8
31993 14 31267746 ARS-BFGL-BAC-10919 9
23506 10 18882288 ARS-BFGL-BAC-10952 10
23550 10 20609250 ARS-BFGL-BAC-10960 11
23566 10 21225382 ARS-BFGL-BAC-10975 12
23612 10 26527257 ARS-BFGL-BAC-10986 13
24705 10 78512500 ARS-BFGL-BAC-10993 14
24712 10 79252023 ARS-BFGL-BAC-11000 15
24732 10 80410977 ARS-BFGL-BAC-11003 16
24741 10 80783719 ARS-BFGL-BAC-11007 17
24827 10 84516867 ARS-BFGL-BAC-11025 18
25865 11 21276136 ARS-BFGL-BAC-11039 21
```

# Saving 'clean' files

- SNP excluded from QC are set to missing (i.e., Code=5)
  - 5 is replaced by 0 in calculations
- OPTION saveCleanSNPs
- Save clean genotype data without excluded SNP and individuals
  - For example, for a SNP\_file named *marker.geno*
- Clean files will be:
  - *marker.geno\_clean*
  - *marker.geno\_clean\_XrefID*
- Removed SNP/animals will be output in files:
  - *marker.geno\_SNPs\_removed*
  - *marker.geno\_Animals\_removed*

# Only QC in preGSf90

- Quality control
- Genomic relationship matrices and inverses
  - Inverse is costly
- How to do only QC avoiding the inverses:
  - `OPTION SNP_file marker.geno`
  - `OPTION saveCleanSNPs`
  - `OPTION createGInverse 0`
  - `OPTION createA22Inverse 0`
  - `OPTION createGimA22i 0`

# No QC in the application programs

- ONLY use:
  - If QC was performed in a previous run and “clean” genotype file is used
- `OPTION SNP_file marker.geno_clean`
- `OPTION no_quality_control`

# Use in application programs

- Use `renumf90` for renumbering and creation of XrefID and files


```
SNP_FILE  
marker.geno
```

```
EFFECT  
1 cross alpha  
RANDOM  
animal  
FILE  
ped3.txt  
FILE_POS  
1 2 3 0 0  
SNP_FILE  
marker.geno  
PED_DEPTH  
0  
(CO) VARIANCES  
0.30
```

- Run `preGSf90` with quality control, saving clean files
- Run further programs with clean files as needed
  - `blupf90+`, `gibbs2f90+`, ...



# PreGSf90 wiki

 BLUPF90

Log In

Search

Media Manager Sitemap

Trace: [start](#) · [application\\_programs](#) · [readme.pregsf90](#)

readme.pregsf90

## PreGSF90 / PostGSF90

PreGSF90 is an interface program to the `genomic` module to process the genomic information for the BLUPF90 family of programs

This page also describes some options for PostGSF90 which is designed for genome-wide association study (GWAS).

Ignacio Aguilar and Ignacy Misztal, University of Georgia  
email: [iaguilar@inia.org.uy](mailto:iaguilar@inia.org.uy); [ignacy@uga.edu](mailto:ignacy@uga.edu)  
01/29/09 - 07/30/14

### Summary

Program PreGSF90 helps to implement the genomic selection following the single-step methodology as presented by [Aguilar et al. 2010 JDS](#). In this methodology the relationship matrix **A** based on the pedigree information is replaced by matrix **H**, which combines the pedigree and genomic information.

The main difference between  $\mathbf{A}^{-1}$  and  $\mathbf{H}^{-1}$  is matrix of structure  
 $\text{GimA22} = \text{inv}(\mathbf{G}) - \text{inv}(\mathbf{A}_{22})$ ,  
where **G** is a genomic relationship matrix and **A<sub>22</sub>** is a relationship matrix for genotyped animals.

Efficient methods for the creation of the genomic relationship matrix, relationship based on pedigree and their inverses are described in [Aguilar et al., 2011 JABG](#). Program PreGSF90 could be run after RENUMF90. It is also run automatically by application programs like BLUPF90, REMLF90, GIBBSxF90 or BLUP90IOD when their parameter file contains OPTION SNP\_file filename.

#### Table of Contents

- ♦ PreGSF90 / PostGSF90
  - ♦ Summary
  - ♦ Input files
  - ♦ Output files
  - ♦ Options for creation of genomic relationship Matrix (G)
  - ♦ Quality Control (QC) for G
  - ♦ Quality Control for Off-diagonal of A22 and G
  - ♦ Options for H
  - ♦ GWAS options (PostGSF90)
  - ♦ Output files for GWAS (postGSf90)
  - ♦ Misc options
  - ♦ Save and Read options
  - ♦ Save and Read intermediate files
- ♦ DEPRECATED OPTIONS

# preGSf90

- Performs Quality Control of SNP information
- Creates the genomic relationship matrix (**G**)
  - and relationships based on pedigree (**A**<sub>22</sub>)
  - Inverse of relationship matrices



# BLUP-based models

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

BLUP

Henderson, 1963

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

GBLUP

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

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

ssGBLUP

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

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

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

# PreGSf90

- Created to construct the matrices using in ssGBLUP

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

$\mathbf{G}$

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

$\mathbf{A}_{22}$

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

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

# Genomic Relationship Matrix - G


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

- $\mathbf{Z}$  = matrix for SNP marker
- Dimension of  $\mathbf{Z} = n \times i$
- $n$  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
```

# 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

# Genomic Matrix default options

- $\mathbf{G}_0 = \frac{\mathbf{Z}\mathbf{Z}'}{2 \sum p_i(1-p_i)}$  (VanRaden, 2008)
- With:
  - $\mathbf{Z}$  centered using current allele frequencies
    - Current genotyped animals

# 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 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)`
- Numerically: Beta can vary from 0 to 1
- In practice: Beta may vary from 0.01 to 0.2

# Genomic Matrix default options

- **Tuning**

- Adjust  $\mathbf{G}$  to have mean of diagonals and off-diagonals equal to  $\mathbf{A}_{22}$
- OPTION tunedG 2    #(default)    Chen et al. (2011)
- Base of GBLUP is *genotyped* animals
- Base of pedigree is *founders of the pedigree*
- For SSGBLUP modelled as a mean for genotyped animals
  - $p(\mathbf{u}_2) = N(\mathbf{1}\mu, \mathbf{G})$
  - Integrate  $\mu : \mathbf{G}^* = \mathbf{1}\mathbf{1}'\lambda + (1 - \lambda/2)\mathbf{G}$
  - $\mu = (\text{Genomic base}) - (\text{Pedigree base})$
  - Vitezica et al. 2011

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

- OPTION tunedG  $x$ 
  - 0: no adjustment
  - 1:  $\text{mean}(\text{diag}(\mathbf{G}))=1$ ,  $\text{mean}(\text{offdiag}(\mathbf{G}))=0$
  - 2:  $\text{mean}(\text{diag}(\mathbf{G}))=\text{mean}(\text{diag}(\mathbf{A}_{22}))$ ,  $\text{mean}(\text{offdiag}(\mathbf{G}))=\text{mean}(\text{offdiag}(\mathbf{A}_{22}))$  (default)
  - 3:  $\text{mean}(\mathbf{G})=\text{mean}(\mathbf{A}_{22})$
  - 4: Use Fst adjustment. Powell et al. (2010) & Vitezica et al. (2011)

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

# 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 the optional *all* is present all intermediate  $\mathbf{G}$  matrices will be saved!!!

To save  $\mathbf{G}$  and inverse:

- OPTION saveG and OPTION saveGInverse
  - Only the final  $\mathbf{G}$ , after blending, scaling, etc. is inverted !!!

To save  $\mathbf{A}_{22}$  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 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

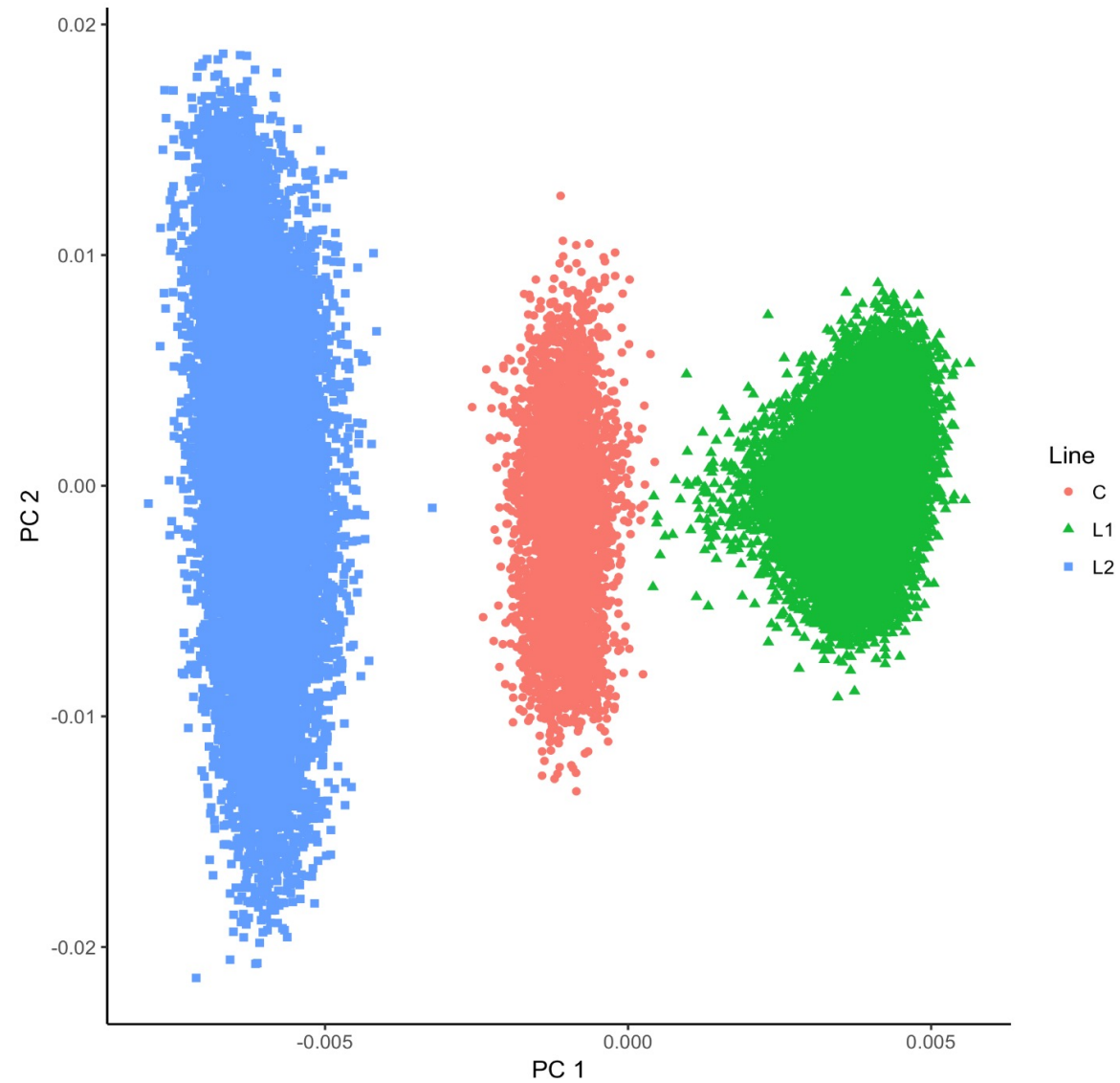
```
OPTION plotpca
```

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  $\rightarrow G = 1.00*\mathbf{G} + 0.00*\mathbf{I}$
- OPTION AlphaBeta 0.95 0.05  $\rightarrow G = 0.95*\mathbf{G} + 0.05*\mathbf{I}$

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

- OPTION tunedG 0

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

1) In renum.par, remove any information about the pedigree. Example:

```
FILE
pedigree.txt

FILE_POS
1 2 3 0 0

PED_DEPTH
3
```

3) Change blending parameters

- `OPTION AlphaBeta 1.00 0.00` →  $\mathbf{G} = 1.00*\mathbf{G} + 0.00*\mathbf{I}$
- `OPTION AlphaBeta 0.95 0.05` →  $\mathbf{G} = 0.95*\mathbf{G} + 0.05*\mathbf{I}$

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

- `OPTION tunedG 0`

# PreGSf90 inside BLUPF90 ??

- Almost all programs from BLUPF90 support creating 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 or Gi and A22i separate

# 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 blupf90+ with clean files
- Option 2:  
run preGSf90 with clean files (program saves **GimA22i**)  
run blupf90+ with option to read **GimA22i** from the file

# Reading external matrices

- BLUPF90 programs accept external matrices created outside
- [http://nce.ads.uga.edu/wiki/doku.php?id=user\\_defined\\_files\\_for\\_covariances\\_of\\_random\\_effects](http://nce.ads.uga.edu/wiki/doku.php?id=user_defined_files_for_covariances_of_random_effects)
- File should be row, column, value in plain text format (lower OR upper triangular)

renf90.par

```
RANDOM_GROUP
# genomic
2
RANDOM_TYPE
user_file
FILE
# matrix file
Gi
```

Valid format

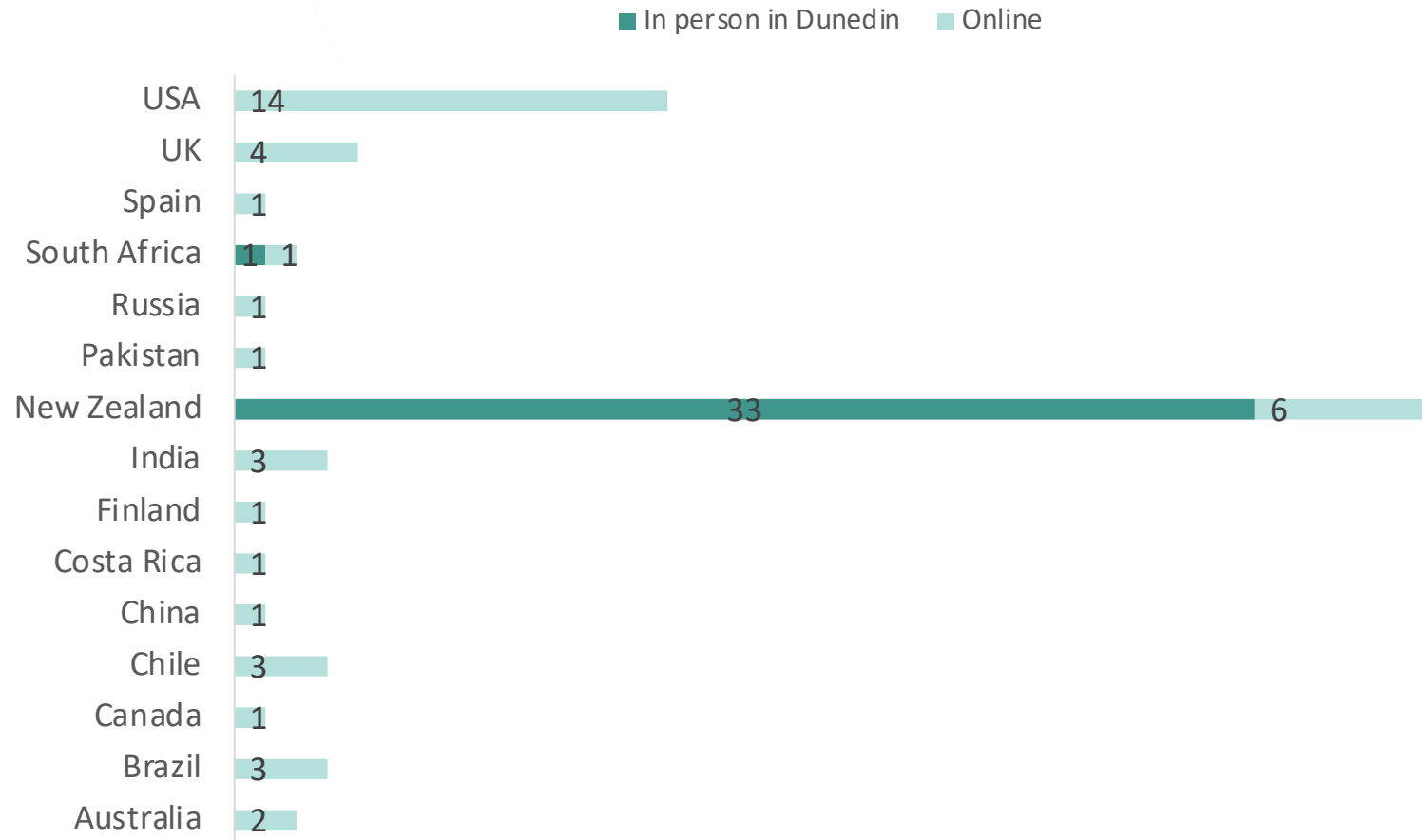
```
1 1 1
1 2 0.5
2 2 1
```

Non-valid format

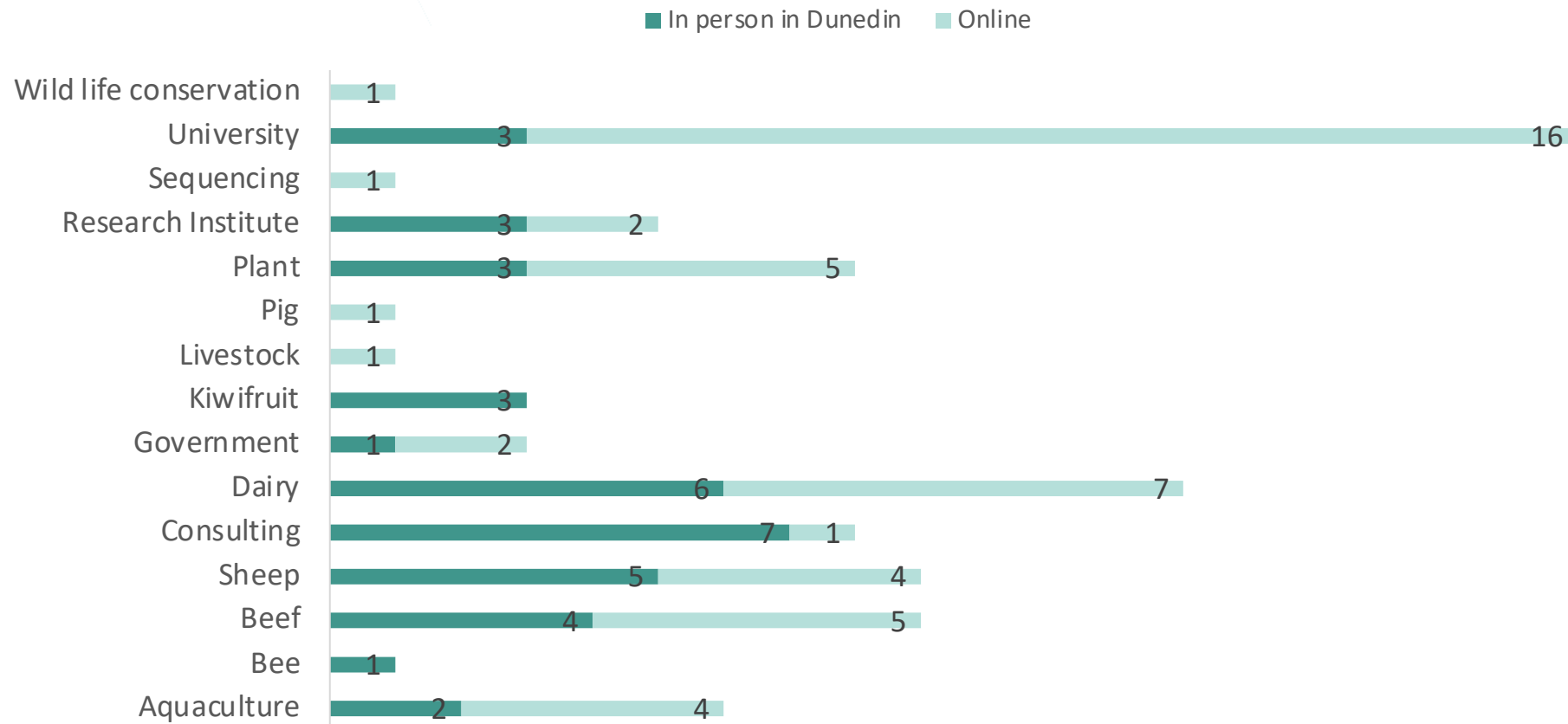
```
1 1 1
1 2 0.5
2 1 0.5
2 2 1
```

- user\_file: if providing the inverse of the covariance structure
- user\_file\_inv: if the program has to invert the covariance structure

# Closing remark



# Closing remark



Thank you  
for coming

