



UNIVERSITY OF  
**GEORGIA**

# PreGSf90 for Quality Control of SNP data

Andres Legarra  
INRA France

Ignacio Aguilar  
INIA Uruguay

Daniela Lourenco  
UGA USA

# SNP data

025 110101111511110111110010001221151205122125022511110250122010201021000221121025000122010:  
036 211011010220121222220121012220101202221111202122211112102020101101020111112011012110:  
050 121010021112021111200021212222100021122122122110000020220000211022122212122020001112020:  
054 120001200220121211100121002222110211221102011212221200220021212121111202112022002022100:  
066 200002020221021221120022001222211101220202110202222020220001222121011201021022010011010:  
097 10110212022012112211102100111110010221121202211111020221001201222012111021021021012000:  
101 121002120220011221100011112220100101120112121211121201221002102002021211222022010022110:  
151 111001020221220210201011012220200121221111221221121111222002201112011212111022000022012:  
172 211012020211112101211021102220101001221212221102220201221020212112010211122022112011010:  
224 22000111022101221010102110252020112120222122212220110121011102220050210121022010022125:  
277 210102200121221211212021012222002012210212110201121021221002211011020211021112021012010:  
314 122011120122220210210010002121001120120202001210020021210011201022021212111022010101100:  
419 2211122101211120222221022102110201021121211122000000111220002211122020222112120012121110:  
439 200202100122121210101021012221101112220202022110010111210011201022012220211021010011020:  
456 1200010202211122001010210022110002022212122222001011022111021201201121221111102112010:  
501 111000021221121201212121002221101202222101022112222110220011202110020201102022100021020:  
571 110000120202200221212022001210200011122110110222221200220020212001010212121022102010110:  
579 1121002102100101011110220022212000222111120202222211022210120201211122211112011011020:  
581 211002021521001221202011002200201125121215022522222250221011201121051202222112111012110:  
657 110011120220111211101020012221000112221212021211121200220012202220022212212112001112011:  
660 210002120221120221121021012221011012221222121211120201221012201121111211112022000012101:  
730 2100020202200202222200120022200012202222200211022522001220012021111510010120220010120250:  
732 212102121521002201200012101121201215110215122521121150220011102111050202221122011022010:  
764 111102121520012212211020001220201225222115021522221150220110202120050202022022111112110:  
780 121101021122220210101022002221201201121221012111110111221020202001010112212121002021021:  
800 22100012022122221020202110222110101211202212022222200221002211121021202011022010111010:  
816 110001220220121220110022011121100011021122121220020112222002222111021111212022011022010:  
832 12101001112001121111002111222011112122221210201111020221002112221001212111121012111110:  
900 210100110220122121211021102121012120221212121101111110221001202121110211011021100022020:  
901 121001020221121212210010002120201111221112122001111110221002201022012212121021000012020:

# Call rate

- Is the percentage of observed (non-missing) genotypes:
  - per animal (per row)
  - per SNP marker (per column)
- In other words, the number of “5”s
- If call rate of an animal <90%
  - genotype of the animal is rejected (delete line)
- If call rate for a marker <90%
  - marker is deleted (delete column)

SNP

ANIMAL

025	110101111(5)11101111100100012211(5)120(5)12212(5)22(5)111102(5)012201020102100022112102(5)000122010
036	211011010220121222220121012220101202221111120212221111121020201011010201111112011012110:
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	1200010202211122001010210022110002022212122222001011022111021201201121221111102112010:
501	111000021221121201212121002221101202222101022112222110220011202110020201102022100021020:
571	110000120202200221212022001210200011122110110222221200220020212001010212121022102010110:
579	11210021021001010111102200222120002221111202022222110222101202012111222111112011011020:
581	21100202(5)100122120201100220020112512121502252222250221011201121051202222112111012110:
657	110011120220111211101020012221000112221212021211121200220012202220022212212112001112011:
660	210002120221120221121021012221011012221222121211120201221012201121111211112022000012101:
730	210002020220020222220012002220001220222220021102252200122001202111151001012022001012025:
732	21210212(5)1002201200012101121201215110215122521211150220011102111050202221122011022010:
764	11110212(5)0012212211020001220201225222115021522221150220110202120050202022022111112110:
780	121101021122220210101022002221201201121221012111110111221020202001010112212121002021021:
800	22100012022122221020202110222110101211202212022222200221002211121021202011022010111010:
816	11000122022012122011002201112110001102112212122002011222200222211021111212022011022010:
832	121010011120011211110021112220111112122221210201111020221002112221001212111121012111110:
900	210100110220122121211021102121012120221212121101111110221001202121110211011021100022020:
901	121001020221121212210010002120201111221112122001111110221002201022012212121021000012020:

# Allele Frequency

- The allele frequency  $p$  is simply the frequency of the reference allele

```

025 1 0101111511110111110010001221151205
036 2 11011010220121222220121012220101202
050 1 2101002111202111120002121222100021
054 1 20001200220121211100121002222110211
066 2 00002020221021221120022001222211101
097 1 01102120220121122111021001111100102
101 1 21002120220011221100011112220100101
151 1 11001020221220210201011012220200121
172 2 11012020211112101211021102220101001
224 2 20001110221012210101021102520201112
277 2 10102200121221211212021012222002012
314 1 22011120122220210210010002121001120
419 2 21112210121120222221022102110201021
439 2 00202100122121210101021012221101112
456 1 20001020221112200101021002211000202
501 1 11000021221121201212121002221101202
571 1 10000120202200221212022001210200011
579 1 12100210210010101111022002221200022
581 2 11002021521001221202011002200201125
657 1 10011120220111211101020012221000112
660 2 10002120221120221121021012221011012
730 2 10002020220020222220012002220001220
732 2 12102121521002201200012101121201215
764 1 11102121520012212211020001220201225
780 1 21101021122220210101022002221201201
800 2 21000120221222210202021102221101012
816 1 1000122022012220110022011121100011
832 1 21010011120011211110021112220111112
900 2 10100110220122121211021102121012120
901 1 21001020221121212210010002120201111
    
```

- 30 animals = 60 alleles
  - 0 = AA
  - 1 = AB
  - 2 = BB
- How many copies of B:
  - $(1+2+1+1+1+...+1)/60$
  - or
  - Average/2
- Allele frequency of B = 0.7167
- Allele frequency of A = 0.2833

# Minor allele Frequency

- MAF is the lowest of the two allele frequencies
- $p = \text{freq}(A)$
- $q = 1 - p = \text{freq}(B)$
- $MAF = \min(p, q)$
- Why is MAF important?
  - A fixed marker ( $p = 0$  or  $p = 1$ ) gives no information
  - An almost-fixed marker ( $p = 0.0001$  or  $p = 0.9999$ ) gives almost no info
  - Common sense: delete markers with  $MAF < 0.01$  or  $< 0.05$
  - For prediction and GWAS it does not make much difference
  - For sequence analysis with *de novo* variants it makes a difference

# Hardy-Weinberg Equilibrium

- If animals reproduce at random we expect to find HW proportions of genotypes:

$$p^2, 2pq, q^2$$

- We can use a Chi-square test to test this, but
  - Does HWE equilibrium this hold?
  - Only approximately
  - At each generation  $p$  changes a little bit, so it does not hold across all generations
  - Also, animals do not mate at random

# Hardy-Weinberg Equilibrium

Rule of thumb used by AIPL (Wiggans 2011):

- Number of heterozygotes should not deviate too much
- Delete marker if  $\left| \frac{n \text{ of heterozygotes}}{n} - 2pq \right| > 0.15$



# Non-mapped SNP

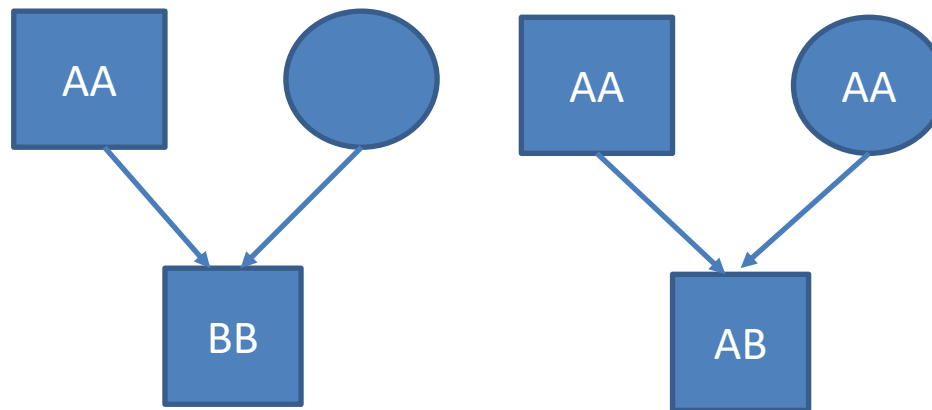
- SNP markers are in chromosomes
- The position of some SNP is still unknown!
- This is reported as “chromosome 0”
- It is better to remove these markers



```
GGaluGA360484 0 0
GGaluGA360493 0 0
GGaluGA360494 0 0
GGaluGA360497 0 0
GGaluGA360501 0 0
GGaluGA360505 0 0
GGaluGA001820 1 34388
Gga_rs16686671 1 67781
GGaluGA001841 1 80477
Gga_rs15995401 1 111556
```

# Mendelian conflicts

- In absence of mutation (which is rare) this kind of inheritance is not possible:



# Mendelian conflicts

- If a marker is seen in many Mendelian conflicts,
  - possibly the genotyping of the marker is wrong and the marker is deleted
- If an animal is seen in many Mendelian conflicts,
  - Possibly there is a misidentification in animal or in pedigree
- You may try to find this animals' parent:
  - seekparentf90

# Duplicate genotypes

- Two animals should not have identical SNPs unless they are clones or monozygotic twins
- Duplicated genotypes come from mislabeling: the DNA sample of the same animal has been given two different IDs

# Linkage disequilibrium

- « Gametic phase disequilibrium »

Statistical association between alleles at two loci in the same chromosome

- Loci : places
- Alleles: alternative forms of a gene (A,B,0)
- Phase: notion of being in the same chromosome (of a pair) or coming from same origin (sire or dam)

# Linkage disequilibrium

$$p(A) = 0.6$$

$$p(B) = 0.5$$

if independent,  $p(AB) = 0.3$ ,  $p(ab) = 0.2$

The expected proportions are:

A	a	
B	0.3	0.2
b	0.3	0.2

# Linkage disequilibrium

$$p(A) = 0.6$$

$$p(B) = 0.5$$

**in reality:**

A	a	
B	0.4	0.2
b	0.1	0.3

vs. **expected**

A	a	
B	0.3	0.2
b	0.3	0.2

More AB & ab than expected !!

This is **linkage disequilibrium** (statistical concept)

# preGSf90

- Interface program to the genomic module to process the genomic information for the BLUPF90 family of programs
- Performs Quality Control of SNP information
- Creates the genomic relationship matrix
  - and relationships based on pedigree
  - Inverse of relationship matrices



# preGSf90

- Controlled by adding OPTION to the parameter file
  - OPTION SNP\_file *marker.geno*
  - Read 2 extra files (besides data and pedigree):
    - *marker.geno*
    - *marker.geno\_XrefID* (created by renumf90)

# Run renumf90 before preGSf90

- Use renumf90 for renumbering and creation of 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
```

# 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
```

## 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
```

Renumber ID

Original ID

# Output Files from preGSf90

- freqdata.count
  - Contains the estimated allele frequency before QC
- freqdata.count.after.clean
  - Contains allele frequencies as used in calculations, remove code
  - For removed SNP these will be zero
- Gen\_call\_rate
  - List of animals removed by low call rate
- Gen\_conflicts
  - Report of animals with Mendelian conflicts
- GimA22i
  - Stores the content of the  $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$
  - Only if preGSf90 is used, not in applications programs

# Quality control default exclusion

- MAF
  - SNP with  $MAF < 0.05$
- Call rate
  - SNP with call rate  $< 0.90$
  - Individuals with call rate  $< 0.90$
- Monomorphic
  - Excludes monomorphic SNP

# Quality control default exclusion

- Parent-progeny conflicts (SNP & Individuals)
  - Exclusion -> opposite homozygous
  - For SNP: Number of parent-progeny exclusion from the total of pairs evaluated (>10 % )
  - For Individuals: Number of parent-progeny exclusions as percentage of all SNP (> 1%)



# Parent-progeny conflicts

- Presence of these conflicts results in a negative **H**
- Problems in estimation of variance components by REML, programs do not converge, etc.
- Solution:
  - Report all conflicts, with counts for each individual as parent or progeny to trace the conflicts
  - Remove progeny genotype
    - maybe not the best option
    - But results in a positive-definite **H**

# Parent-progeny conflicts

- OPTION verify\_parentage x
  - 0: no action
  - 1: only detect
  - 2: detect and search for an alternate parent; no change to any file. Not implemented
    - implemented in **seekparentf90** program
  - 3: detect and eliminate progenies with conflicts (default)

# Control default values

- For MAF
  - OPTION minfreq x
- Call rate
  - OPTION callrate x
  - OPTION callrateAnim x
- Mendelian conflicts
  - OPTION exclusion\_threshold\_snp x
  - OPTION exclusion\_threshold x

# Other Options

- Exclusion of selected chromosomes:
  - OPTION excludeCHR *n1 n2 n3 ...*
- Inclusion of selected chromosomes:
  - OPTION includeCHR *n1 n2 n3 ...*
- Exclude samples from analyses
  - OPTION excludeSample *n1 n2 n3 ...*
- Inform which are sex chromosomes:
  - OPTION sex\_chr *n*
  - Chromosome  $\geq n$  will be excluded only for HWE and parent-progeny checks, but not in calculations

# LD calculation and options

```
OPTION calculate_LD
```

Calculate LD as the squared correlation of allele counts for two SNP

Results are stored in "ld\_results", columns: snp\_i, chr\_i, pos\_i, freq\_i, snp\_j, chr\_j, pos\_j, freq\_j, dist\_ij, Rsq\_ij

```
OPTION LD_by_chr
```

Calculate LD within chromosome

```
OPTION LD_by_pos x
```

Calculate LD within chromosome and windows of SNP based on position optional parameter x define with windows size in Bp, default value 200000

```
OPTION filter_by_LD x
```

Filter SNP with  $Rsq > \text{threshold}$ . Optional parameter x define the threshold. default value 0.8

```
OPTION thr_output_LD x
```

Threshold to print out  $Rsq$  between pair of SNP Optional parameter x define the threshold. default value 0.1

# SNP map file

- OPTION chrinfo <file>
- For some genomic analyses (GWAS) or QC
- Format:
  - SNP number
    - Index number of SNP in the sorted map by chromosome and position
  - Chromosome number
  - Position
  - SNP name (Optional)
- First column is like a header for the SNP file!

```
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
```

# SNP map file – new default

- OPTION chrinfo <file>
- OPTION map\_info <file>
- Format:
  - No defined position if a header is 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 as missing (i.e. Code=5)
  - 5 is replaced by 0 in calculations
- OPTION saveCleanSNPs
- Save clean genotype data with excluded SNP and individuals
  - For example for a SNP\_file named *gt.snp*
  - Clean files will be:
    - *gt.snp\_clean*
    - *gt.snp\_clean\_XrefID*
  - Removed SNP/animals will be output in files:
    - *gt.snp\_SNPs\_removed*
    - *gt.snp\_Animals\_removed*



# Only QC in preGSf90

- Quality control
- Genomic relationship matrix
- 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 Quality control

- 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 as needed
  - blupf90, airemlf90, 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](#)