UNIVERSITY OF GEORGIA

# PreGSf90
# for Quality Control of SNP data

Andres Legarra
INRA France

Ignacio Aguilar
INIA Uruguay

Daniela Lourenco
UGA USA

UGA TEAM, 08/2019

# Call rate

- Is the percentage of observed genotypes:
  - per animal (per row)
  - per marker (per column)
- In other words, the number of “5”s

- If call rate of an animal <95%
  - genotype of the animal is rejected (delete line)
- If call rate for a marker <95%
  - marker is deleted

SNP

ANIMAL

```
025   1101011115111101111100100012211512051221250225111102501220102010210002211210250001220100
036   2110110102201212222201210122201012022211111202122211111210202010110102011111120110121101
050   1210100211120211112000212122221000211221221221100000202200002110221222121220200011120201
054   1200012002201212111001210022221102112211020112122212002200212121211112021120220020221002
066   2000020202210212211200220012222111012202021102022220202200012221210112010210220100110101
097   1011021202201211221110210011111001022112120221111110202210012012220121110210210210120002
101   1210021202200112211000111122201001011201121212111212012210021020020212112220220100221102
151   1110010202212202102010110122202001212211112212211211112220022011120112121110220000220122
172   2110120202111121012110211022201010012212122211022202012210202121120102111220221120110101
224   2200011102210122101010211025202011121202221222122201101210111022200502101210220100221252
277   2101022001212212112120210122220020122102121102011210212210022110110202110211120210120101
314   1220111201222202102100100021210011201202020012100200212100112010220212121110220101011002
419   2211122101211202222210221021102010211212111220000001112200022111220202221121200121211102
439   2002021001221212101010210122211011122202020221100101112100112010220122202110210100110202
456   1200010202211122001010210022110002022212122222001011022111021201201121221111110211201012
501   1110000212211212012121210022211012022221010221122221102200112021100202011020221000210202
571   1100001202022002212120220012102000111221101102222212002200202120010102121210221020101102
579   1121002102100101011110220022212000222111120202222221102221012020121112221111120110110202
581   2110020215210012212020110022002011251212150225222222502210112011210512022221121110121102
657   1100111202201112111010200122210001122212120212111212002200122022200222122121120011120112
660   2100021202211202211210210122210110122212221212111202012210122011211112111120220000121012
730   2100020202200202222200120022200012202222200211022522001220012021111510010120220010120250
732   2121021215210022012000121011212012151102151225211211502200111021110502022211220110220101
764   1111021215200122122110200012202012252221150215222211502201102021200502020220221111121102
780   1211010211222202101010220022212012011212210121111101112210202020010101122121210020210212
800   2210001202212222102020211022211010121120221202222222002210022111210212020110220101110102
816   1100012202201212201100220111211000110211221212200201122220022221110211112120220110220102
832   1210100111200112111100211122201111121222212102011110202210021122210012121111210121111102
900   2101001102201221212110211021210121202212121211011111102210012021211102110110211000220201
901   1210010202211212122100100021202011112211121220011111102210022010220122121210210000120202
```

# Allele Frequency

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

```
025  1101011115111101111100100012211512 05
036  2110110102201212222201210122201012 02
050  1210100211120211112000212122221000 21
054  1200012002201212111001210022221102 11
066  2000020202210212211200220012222111 01
097  1011021202201211221110210011111001 02
101  1210021202200112211000111122201001 01
151  1110010202212202102010110122202001 21
172  2110120202111210121102110222010100 1
224  2200011102210122101010211025202011 12
277  2101022001212212112120210122220020 12
314  1220111201222202102100100021210011 20
419  2211122101211202222210221021102010 21
439  2002021001221212101010210122211011 12
456  1200010202211122001010210022110002 02
501  1110000212211212012121210022211012 02
571  1100001202022002212120220012102000 11
579  1121002102100101011110220022212000 22
581  2110020215210012212020110022002011 25
657  1100111202201112111010200122210001 12
660  2100021202211202211210210122210110 12
730  2100020202200202222200120022200012 20
732  2121021215210022012000121011212012 15
764  1111021215200122122110200012202012 25
780  1211010211222202101010220022212012 01
800  2210001202212222102020211022211010 12
816  1100012202201212201100220111211000 11
832  1210100111200112111100211122201111 12
900  2101001102201221211021102121012120
901  1210010202211212122100100021202011 11
```

- 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 frequecies
- $p = freq(A)$
- $q = 1 - p = freq(B)$
- $MAF = min(p, q)$
- 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 $\left|\frac{n\ of\ heterozygotes}{n} - 2pq\right| > 0.15$

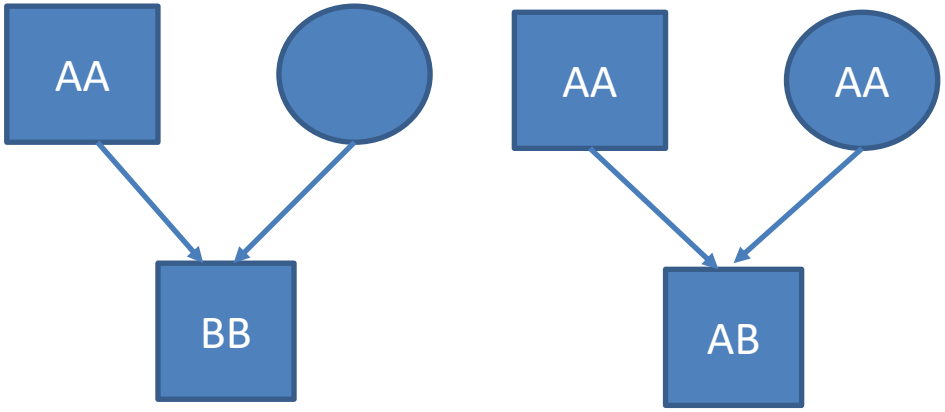
# Non-mapped SNP

- 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:
  - seekparent.f90

# Duplicate genotypes

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

# 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**

# Linkage disequilibrium

- Is a *statistical* concept
- Describes not-random association of two loci
  - Nothing more, so, why is it useful?
- Two loci in LD *most often* are (very) close
  - This is because LD breaks down with recombination
- Linkage disequilibrium of two loci decays *on average* with the distance
- Where does it come from?

- Because chromosomes are transmitted together

  Within known families (« linkage analysis »)

  Within the history of a population (« populational linkage disequilibrium » or « linkage disequilibrium » in short)

# preGSf90

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

- Performs Quality Control of SNP information
- Efficient methods
  - Creation of the genomic relationship matrix and relationship based on pedigree
  - Inverse of relationship matrices

# BLUPF90 programs using Genomic

- Genomic programs
  - controlled by adding OPTION to the parameter file
  - `OPTION SNP_file` *`marker.geno`*
  - Read 2 files:
    - `marker.geno`
    - `marker.geno_XrefID` (created by renumf90)

# Output Files

- 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
  - Store the content of the inv(G) – inv(A22)
  - 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 ONLY when MAF <> 0

# Quality control default exclusion

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

# Control default values

- For MAF
  - OPTION minfreq x
- Call rate
  - OPTION callrate x
  - OPTION callrateAnim x
- Mendelian conflicts
  - OPTION exclusion_threshold x
  - OPTION exclusion_threshold_snp x

# Parent-progeny conflicts

- Presence of these conflicts results in a negative **H**
- Problems in estimation of variance components by REML, programs does 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)

# 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 # > n will be excluded only for HWE and parent-progeny checks, but not in calculations

# 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)

```
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
```

- First column corresponds to first row SNP in genotype file !!!

# SNP map file – new option

- 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
  - SNP_ID 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)

- OPTION saveCleanSNPs
- Save clean genotype data with excluded SNP and individuals
  - For example for a SNP_file named *gt*
  - Clean fles will be:
    - *gt*_clean
    - *gt*_clean_XrefID
  - Removed SNP/animals will be output in files:
    - *gt*_SNPs_removed
    - *gt*_Animals_removed

# 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 > 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

# Only QC in preGSf90

- Quality control
- Genomic relationship matrix (tomorrow)
- How to do only QC avoiding extra steps:
  - `OPTION SNP_file` *`marker.geno`*
  - `OPTION saveCleanSNPs`
  - `OPTION createG 0`
  - `OPTION createGInverse 0`
  - `OPTION createA22 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
  - airemlf90, blupf90, …

# 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 assocication study (GWAS).

Ignacio Aguilar and Ignacy Misztal, University of Georgia
email: iaguilar at inia.org.uy; ignacy at 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
GimA22i=inv(**G**)-inv($\mathbf{A}_{22}$),
where **G** is a genomic relationship matrix and
$\mathbf{A}_{22}$ 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