UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Quality Control of SNP data
+
Using PreGSf90

Daniela Lourenco
UGA USA

Andres Legarra
INRA France

Ignacio Aguilar
INIA Uruguay

BLUPF90 TEAM, 02/2022

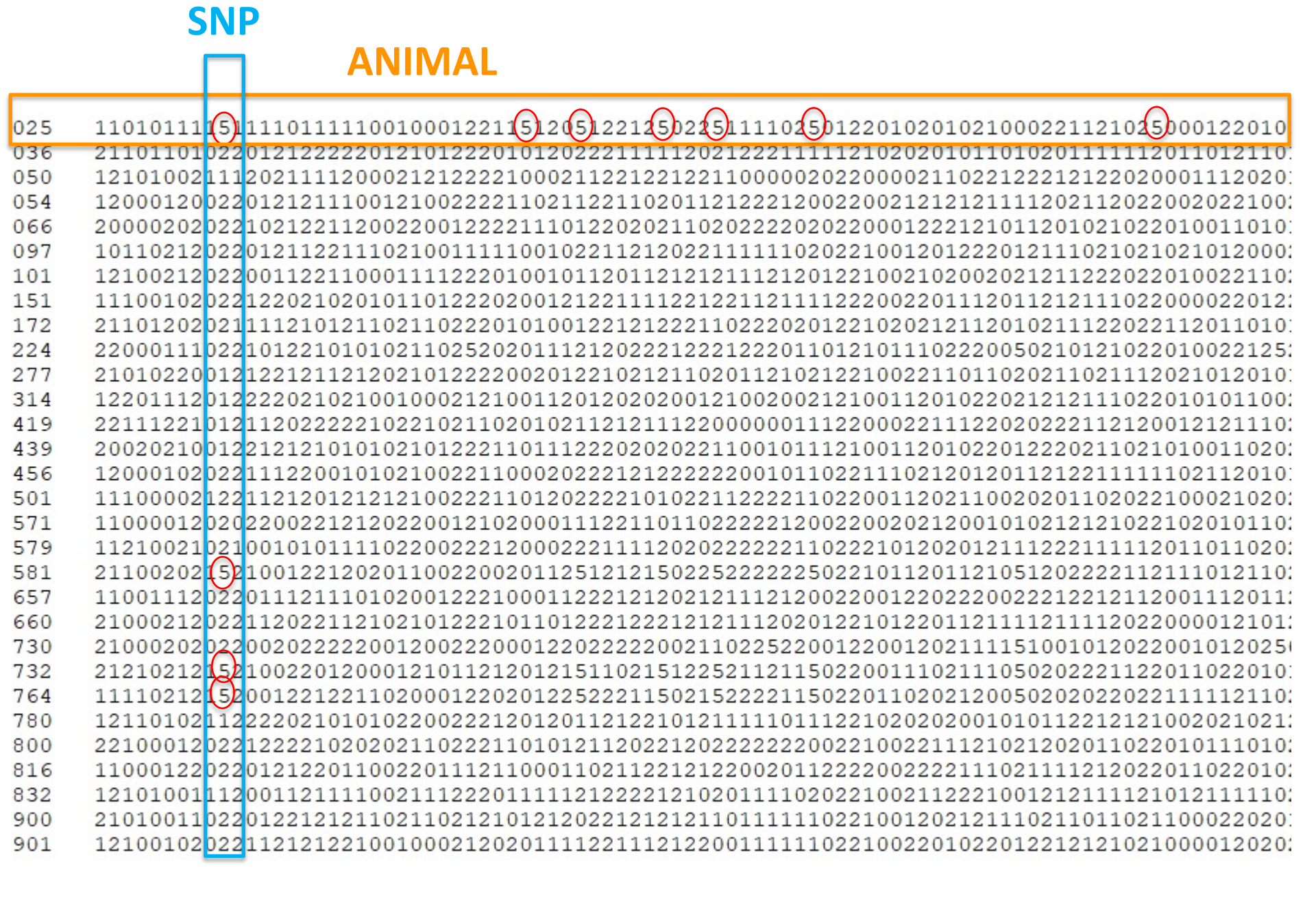
SNP data

```
025   1101011115111101111100100012211512051221250225111102501220102010210002211210250001220102
036   2110110102201212222201210122201012022211111202122211111210202010110102011111201101211012
050   1210100211120211112000212122210002112212212211000002022000021102212221212202000111202012
054   1200012002201212111001210022221102112211020112122212002200212121211112021120220020221002
066   2000020202210212211200220012222111012202021102022220202200012221210112010210220100110102
097   1011021202201211221110210011111001022112120221111102022100120122201211102102102101200021
101   1210021202200112211000111122201001011201121212111212012210021020020212112220220100221102
151   1110010202212202102010110122202001212211112212211211112220022011120112121110220000220122
172   2110120202111121012110211022201010012212122211022202012210202121120102111220221120110101
224   2200011102210122101010211025202011121202221222122201101210111022200502101210220100221252
277   2101022001212212112120210122220020122102121102011210212210022110110202110211120210120101
314   1220111201222202102100100021210011201202020012100200212100112010220212121110220101011002
419   2211122101211202222210221021102010211212111220000001112200022111220202221121200121211102
439   2002021001221212101010210122211011122202020221100101112100112010220122202110210100110203
456   1200010202211122001010210022110002022212122222200101102211102120120112122111110211201012
501   1110000212211212012121210022211012022221010221122211022001120211002020110202210002102023
571   1100001202022002212120220012102000111221101102222212002200202120010102121210221020101102
579   1121002102100101011110220022212000222111120202222221102221012020121112221111120110110203
581   2110020215210012212020110022002011251212150225222225022101120112105120222211211101211023
657   1100111202201112111010200122210001122212120212111212002200122022200222122121120011201132
660   2100021202211202211210210122210110122212221212111202012210122011211112111120220000121013
730   2100020202200202222200120022200012202222200211022522001220012021111510010120220010120250
732   2121021215210022012000121011212012151102151225211211502200111021110502022211220110220101
764   1111021215200122122110200012202012252221150215222211502201102021200502020220221111121102
780   1211010211222202101010220022212012011212210121111101112210202020010101122121210020210213
800   2210001202212222102020211022211010121120221202222220022100221112102120201102201011010302
816   1100012202201212201100220111211000110211221212200201122220022221110211112120220110220102
832   1210100111200112111100211122201111121222212102011110202210021122210012121111210121111102
900   2101001102201221211021102121012120221212121101111110221001202121110211011021100220201
901   1210010202211212122100100021202011112211121220011111102210022010220122121210210000120203
```

Call rate

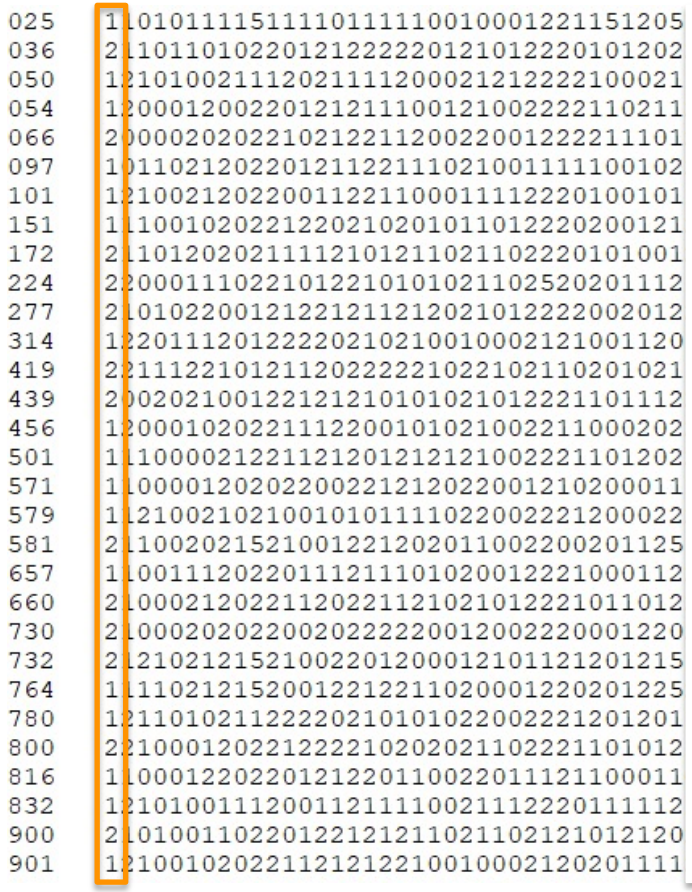
- Is the percentage of observed genotypes (non-missing):
  - per animal (per row)
  - per SNP marker (per column)
- In other words, proportion of SNP ≠ 5

- 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 110101111511110111110010001221151205122125028511110250122010201021000221121025000122010
036 211011010220121222220121012220101202221111120212221111121020201011010201111112011012110
050 121010021112021111200021212222100021122122122110000020220000211022122212122020001112020
054 120001200220121211100121002222110211221102011212221200220021212121111202112022002022100
066 200002020221021221120022001222211101220202110202222020220001222121011201021022010011010
097 101102120220121122111021001111100102211212022111111020221001201222012111021021021012000
101 121002120220011221100011112220100101120112121211121201221002102002021211222022010022110
151 111001020221220210201011012220200121221111221221121111222002201112011212111022000022012
172 211012020211112101211021102220101001221212221102220201221020212112010211122022112011010
224 220001110221012210101021102520201112120222122212220110121011102220050210121022010022125
277 210102200121221121202101222200201221021211020112102122100221101102021102111202101201010
314 122011120122220210210010002121001120120202001210020021210011201022021212111022010101100
419 221112210121120222221022102110201021121211122000000111220002211122020222112120012112110
439 200202100122121210102102122211011122202020221100101112100112010220122202110210100110202
456 120001020221112200101021002211000202221212222200101102211102120120112122111110211201010
501 111000021221121201212121002221101202222101022112222110220011202110020201102022100021020
571 110000120202200221212022001210200011122110110222221200220020212001010212121022102010110
579 112100210210010101111022002221200022211112020222221102221012020121112221111120110110202
581 211002021521001221202011002200201125121215022522222502210112011211051202222112111012110
657 110011120220111211101020012221000112221212021211121200220012202220022212212112001112011
660 210002120221120221121021012221011012221222121211120201221012201121111211112022000012101
730 210002020220020222220012002220000122022222002110225220012200120211115100101202200101202
732 212102121521002201200012101121201215110215122521121150220011102111050202221122011022010
764 111102121520012212211020001220201225222115021522221150220110202120050202022022111112110
780 121101021122220210101022002221201201121221012111110111221020202001010112212121002021021
800 221000120221222210202021102221101012112022120222222200221002211121021202011022010111010
816 110001220220121220110022011121100011021122121220020112222002222111021111212022011022010
832 121010011120011211110021112220111112122221210201111020221002112221001212111121012111110
900 210100110220122121211021102121012120221212121101111102210012021211102110110211000220201
901 121001020221121212210010002120201111221112122001111110221002201022012212121021000012020

Allele Frequency

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



- 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
- $q = freq(B)$
- $p = 1 - q = freq(A)$
- $MAF = min(p, q)$
- 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:

  $f(\text{AA}) = p^2$

  $f(\text{AB}) = 2pq$

  $f(\text{BB}) = q^2$

- We can use a Chi-square test to test this, but
  - Does HWE equilibrium 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
    - many SNP removed

Hardy-Weinberg Equilibrium

Rule of thumb used by AIPL (Wiggans 2011):

- frequency of heterozygotes should not deviate too much

- Delete marker if $\left|\frac{n\ of\ heterozygotes}{n} - 2pq\right| > 0.15$

- Tricky in crossbred populations

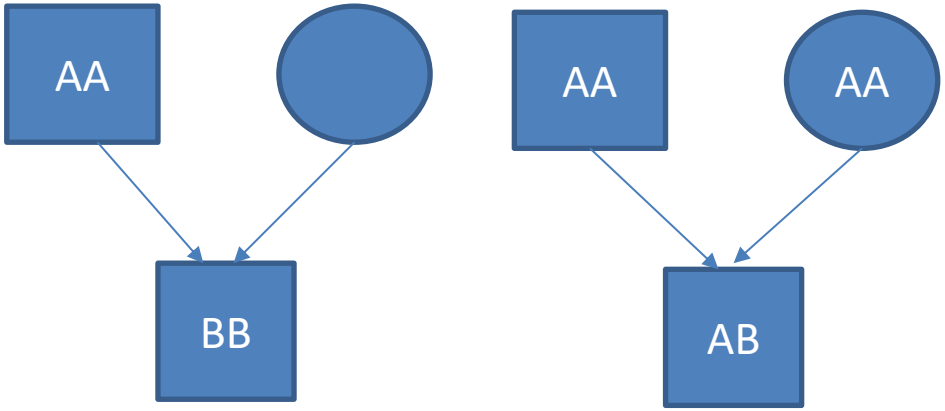
Non-mapped SNP

- 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
  - the genotyping is wrong and the marker is deleted
- If an animal is seen in many Mendelian conflicts
  - there is a misidentification for animal or pedigree
- try to find the possible parents based on SNP
  - 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)
  - Phase: notion of being in the same chromosome (of a pair) or coming from same origin (sire or dam)

Linkage disequilibrium

$f(A)=0.6;\ f(a)=0.4$

$f(B)=0.5;\ f(b)=0.5$

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

$f(A)=0.6$; $f(a)=0.4$
$f(B)=0.5$; $f(b)=0.5$

in reality:

	A	a
B	0.4	0.2
b	0.1	0.3

expected:

	A	a
B	0.3	0.2
b	0.3	0.2

More AB & ab than expected !!

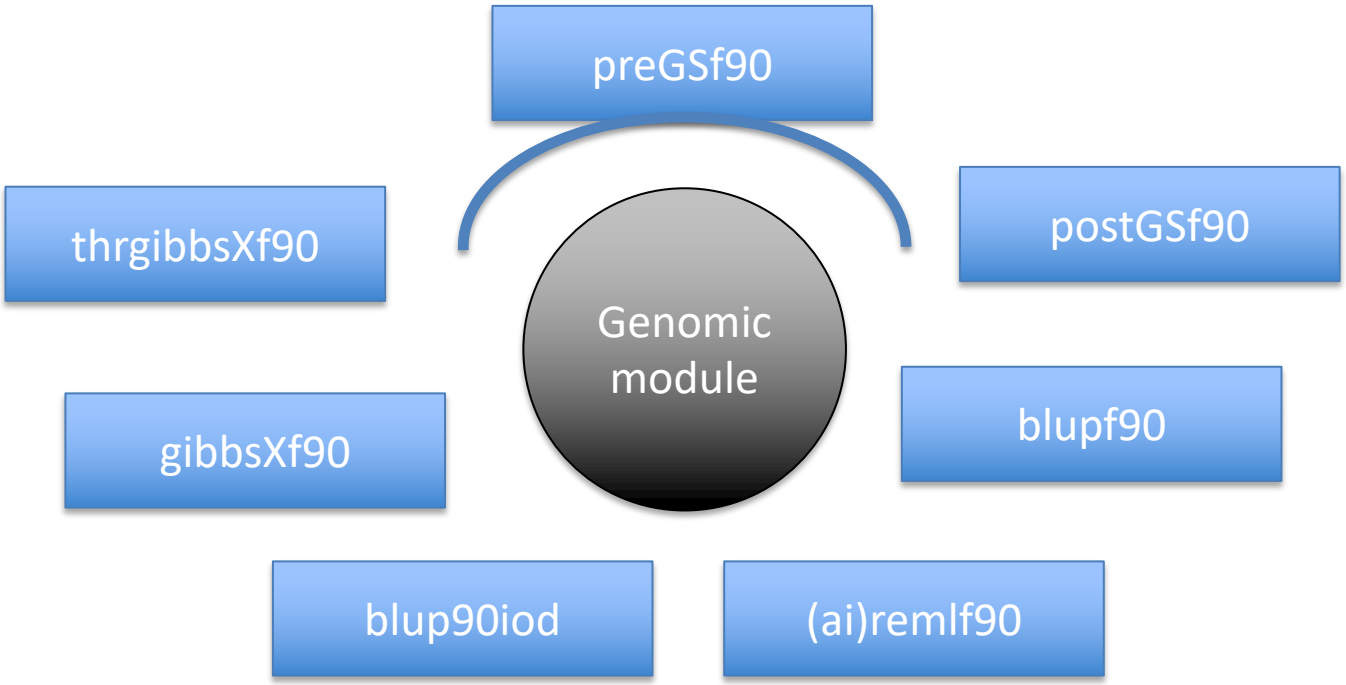
This is **linkage disequilibrium** (statistical concept)

Quality control

- 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
  - and relationships based on pedigree
  - Inverse of relationship matrices

preGSf90

- Same parameter file as for all BLUPF90 programs
- Needs an extra OPTION
  - `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 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
```

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
- 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   21101011002012011011010110111 1
8014 21110101511101120221110111511 1
516  21100101202252021120210121102 1
181  21110111112201120550200020101 0
```

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

Output Files from preGSf90

- freqdata.count
  - Contains the calculated allele frequency before QC
- freqdata.count.after.clean
  - Contains allele frequencies as used in calculations, removal code
  - AF will be zero for removed SNP
- 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 other 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 may 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 (problem may be in the pedigree)
    - 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 >= *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 > 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 – new default

- OPTION chrinfo *<file>*
- OPTION map_info *<file>*
  - For GWAS and QC
- 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 without excluded SNP and individuals
  - For example for a SNP_file named *gt.snp*
  - Clean fles 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 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 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 with clean files as needed
  - blupf90, airemlf90, gibbs2f90, …

PreGSf90 wiki

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 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