

# Quality control of SNP data with preGSf90 or qcf90

Daniela Lourenco

Ignacio Aguilar

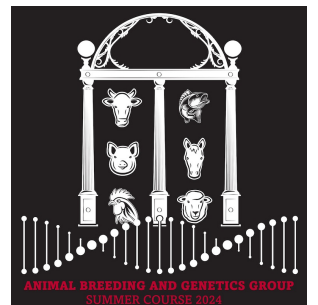
BLUPF90 TEAM – 05/2024



UNIVERSITY OF  
**GEORGIA**

College of Agricultural &  
Environmental Sciences

*Animal Breeding and  
Genetics Group*



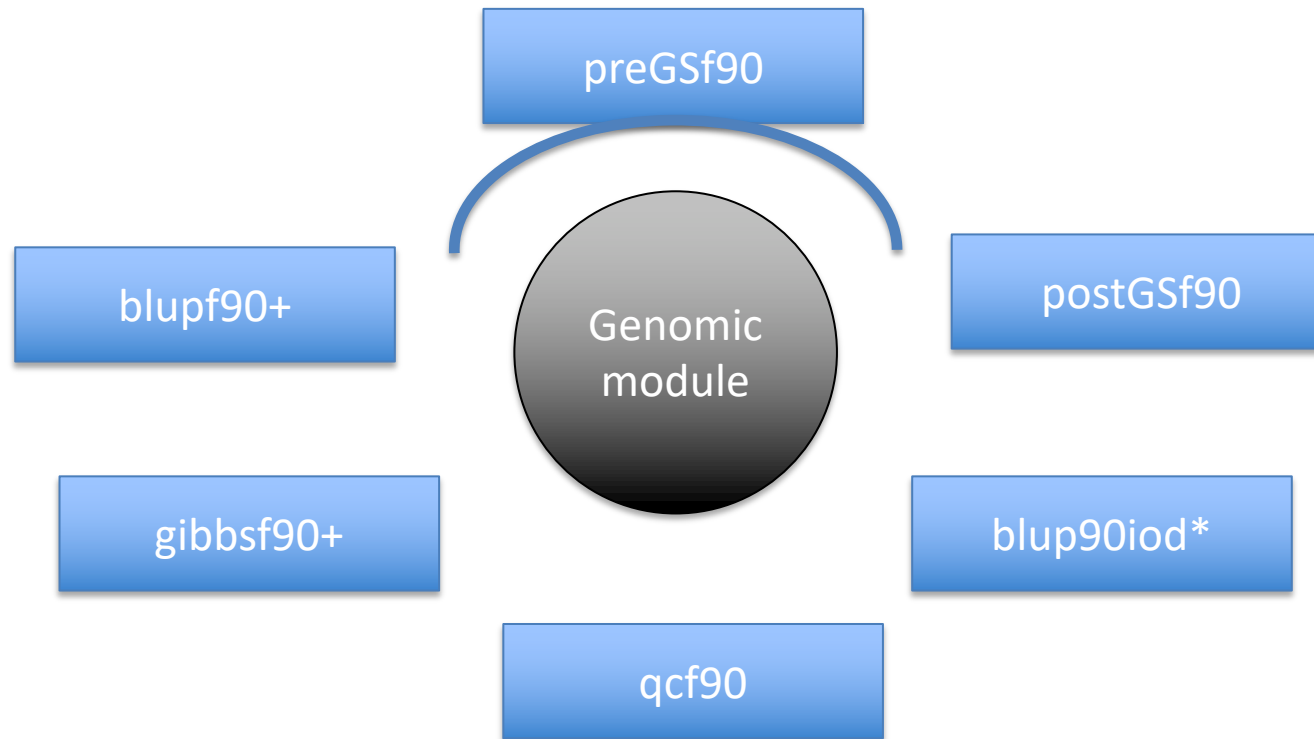
# Quality control

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
  - and relationships based on pedigree
  - 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)

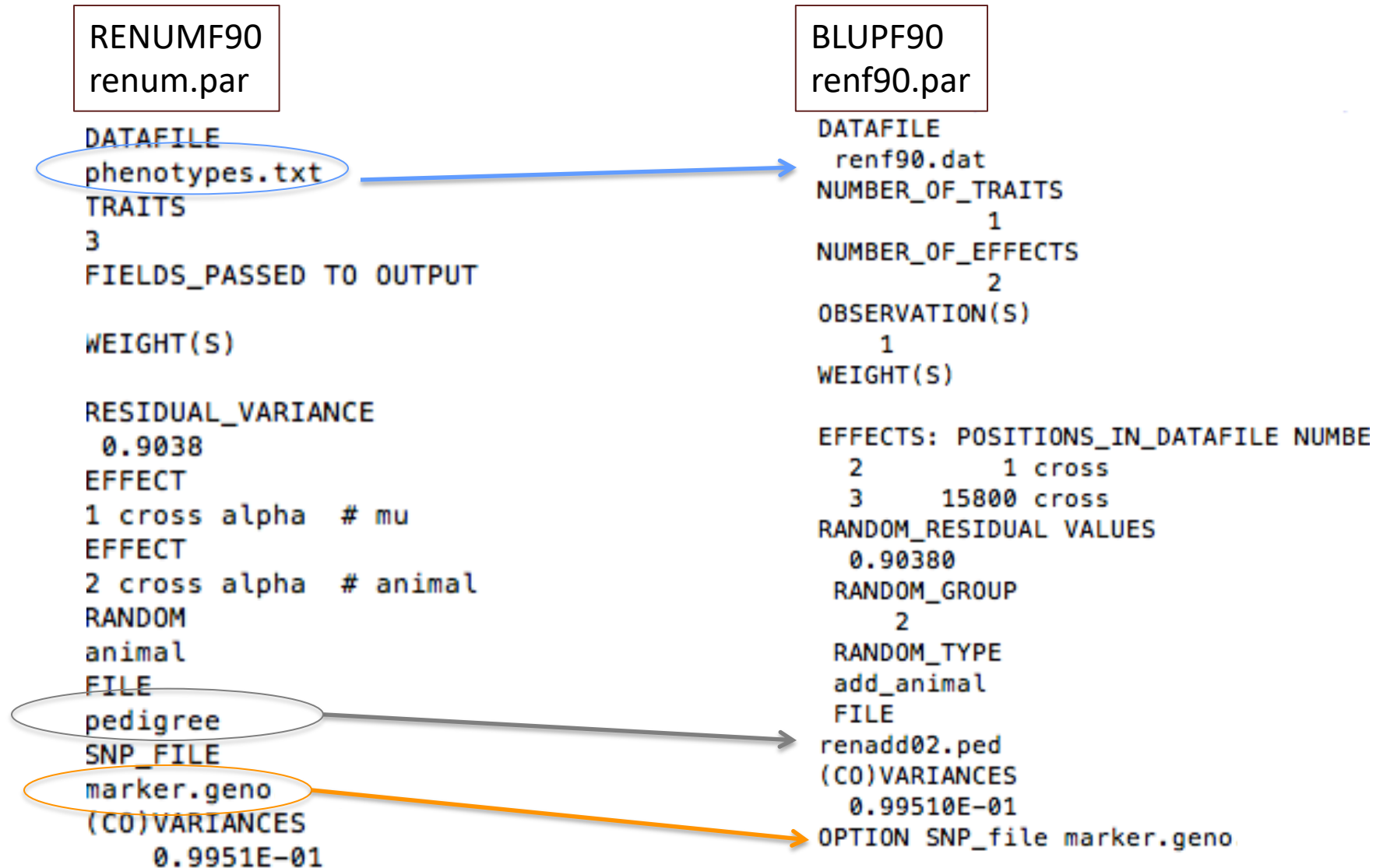
`_XrefID` has 2 columns: Renumbered\_ID Original\_ID

# 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



# renaddXX.ped 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 after running 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

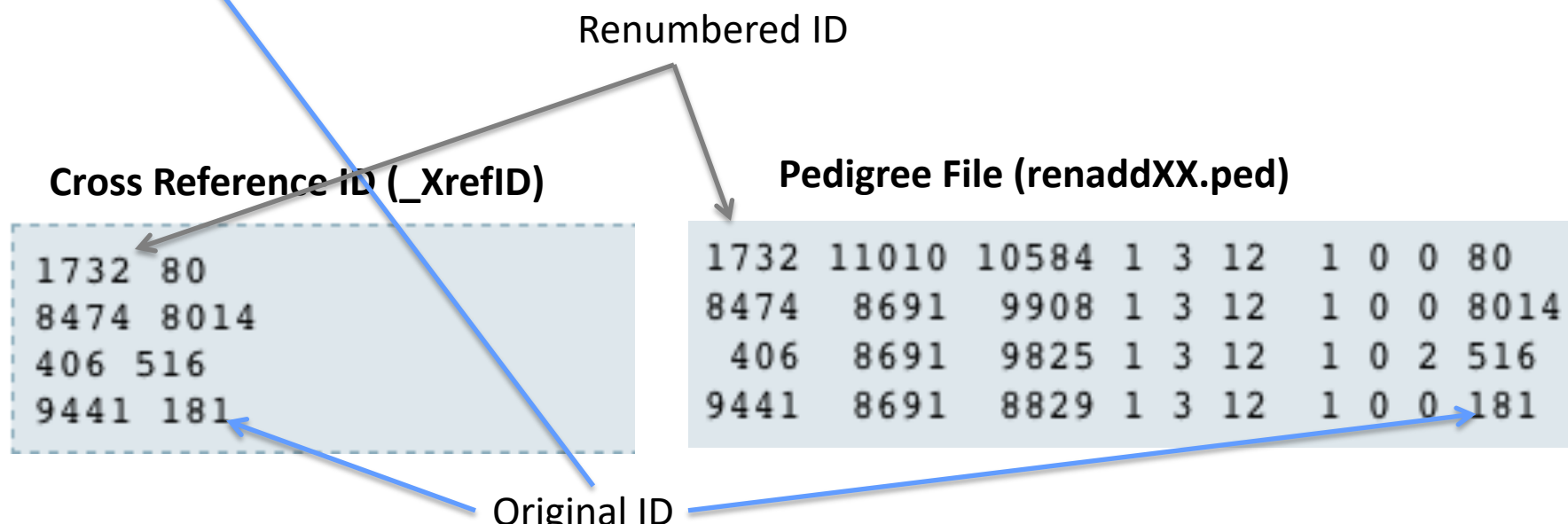
Cross Reference ID (\_XrefID)

```
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

# 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  $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$
  - Only if preGSf90 is used, not in the other programs

# Quality control default exclusion

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

# 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	1200010202211122001010210022110002022212122222001011022111021201201121221111102112010:
501	111000021221121201212121002221101202222101022112222110220011202110020201102022100021020:
571	110000120202200221212022001210200011122110110222221200220020212001010212121022102010110:
579	1121002102100101011110220022212000222111120202222211022210120201211122211112011011020:
581	2110020252100122120201100220020112512121502252222250221011201121051202222112111012110:
657	110011120220111211101020012221000112221212021211121200220012202220022212212112001112011:
660	210002120221120221121021012221011012221222121211120201221012201121111211112022000012101:
730	210002020220020222220012002220001220222220021102252200122001202111151001012022001012025:
732	21210212521002201200012101121201215110215122521211150220011102111050202221122011022010:
764	11110212520012212211020001220201225222115021522221150220110202120050202022022111112110:
780	12110102112220210101022002221201201121221012111110111221020202001010112212121002021021:
800	22100012022122221020202110222110101211202212022222200221002211121021202011022010111010:
816	11000122022012122011002201112110001102112212122002011222200222111021111212022011022010:
832	12101001112001121111002111222011111212222121020111102022100211222100121211112101211110:
900	210100110220122121211021102121012120221212121101111110221001202121110211011021100022020:
901	121001020221121212210010002120201111221112122001111110221002201022012212121021000012020:

# Parent-progeny conflicts

- `OPTION verify_parentage x`
  - 0: no action
  - 1: only detect
  - 2: detect and search for an alternate parent. Not implemented!!!
    - implemented in **seekparentf90** program
  - 3: detect and eliminate progeny 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` (10%)

- `OPTION exclusion_threshold x` (1%)

# Control default values

```
OPTION exclusion_threshold x
```

Number of parent-progeny exclusions as percentage all SNP to determine the wrong relationship.  
default value 1

```
OPTION exclusion_threshold_snp x
```

Number of parent-progeny exclusions for each locus as a percentage, of pair of genotyped animals evaluated, to exclude an SNP from the analysis  
default value 10

```
OPTION number_parent_progeny_evaluations x
```

Set the number of minimum pair of parent-progeny evaluations to exclude SNPs due to parent-progeny exclusions  
default value 100



# Other Options

- Departure of heterozygous from Hardy-Weinberg Equilibrium

OPTION hwe *x*

- Exclusion of selected chromosomes:

OPTION excludeCHR *n1 n2 n3...*

- Inclusion of selected chromosomes:

OPTION includeCHR *n1 n2 n3...*

- Exclude samples from analyses

OPTION excludeSample *i1 i2 i3...*

- Inform which are the sex chromosomes:

OPTION sex\_chr *n*

– Chromosome  $\geq n$  will be excluded for HWE and parent-progeny checks, not for calculations

# Heritability of gene content

```
OPTION h2_gene_content
```

It checks that the heritability of gene content is equal or close to 1 as described in Forneris et al. Genetics 199.3 (2015): 675-681. Markers with estimated  $h^2 < 0.98$  **and** significant p-values of the LRT ( $p < 0.01$ ) are discarded. In addition, heritability and status of each marker are written in file `h2_gc_test`.

The test is useful for homogenous populations (breeds) but theory does not hold for crossbred animals. This test uses explicitly  $\text{inv}(A22)$  so it is not suitable for very large populations.

# 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 – new default

- `OPTION chrinfo <file>`
- `OPTION map_file <file>`

- For QC and GWAS

- 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 the application programs

- Use `renumf90` for renumbering and creating 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](#).

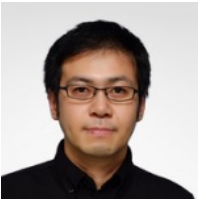
Program `PreGSF90` could be run after `RENUMF90`.

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](#)

# qcf90

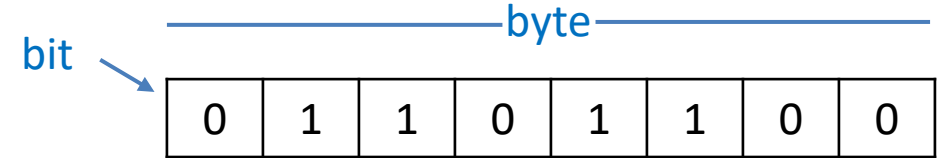


Yutaka  
Masuda

- Quality control tool for large genomic data
  - What is an efficient way to detect genomically identical animals?
  - It implies we should compare all pairs of genotyped animals
- Huge data and slow operations
  - More than 5 million genotyped Holsteins!
  - $80\text{K SNPs} \times [5\text{M} \times 5\text{M}] / 2 \sim 1 \times 10^{18}$  comparisons needed
- The other checks are also needed...
  - Call rate, low MAF, Mendelian conflicts, etc.

# qcf90

- Four states for a biallelic SNP



Genotype	Character	ASCII (8bits)	Re-coded (2bits)
Homozygote (AA)	"0"	00110000	01
Heterozygote (Aa)	"1"	00110001	11
Another Homozygote (aa)	"2"	00110010	10
Missing	"5"	00110101	00

- Task: read and keep 5M genotypes in memory
  - Regular format: 3 TB RAM
  - Efficient format (packed): 93 GB RAM

# qcf90

- Logical manipulation of bit pattern
  - Fortran has functions for bitwise operations
    - Logical manipulation on bit pattern

– Typical operations:

	1100		1100		1100		
<b>AND</b>	1010	<b>OR</b>	1010	<b>XOR</b>	1010	<b>NOT</b>	1010
	----		----		----		----
	1000		1110		0110		0101

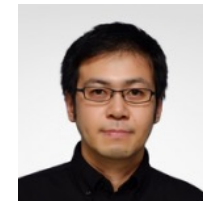
– Population count: the number of 1's

```
popcnt(0000) is 0
popcnt(0010) is 1
popcnt(1010) is 2
```

# qcf90

- qcf90 supports raw files
  - No need to run renumf90 before
- qcf90 was designed for QC
  - preGSf90 was designed for QC and constructing **G** and **A<sub>22</sub>**
- `qcf90 --snpfile snpdata.txt --pedfile pedigree.txt`
  - No parameter file but same output as preGSf90
- `qcf90 --help`                      or                      `qcf90 --long-help`
  - For all the options

# qcf90



Yutaka  
Masuda

- Benchmark test:
  - Holstein genotypes: 569,404
  - Number of SNP: 60,671
  - Number of animals in pedigree: 10,710,380

3x faster  
28x less memory

Step	QCF90 (sec.)	PREGSF90 (sec.)
Reading a SNP file	420	1407
MAF and call rate	150	245
HWE test	84	24
Call rate for animals	3	307
Mendelian tests for SNP	62	316
Mendelian tests for animals	62	248
Recalculation of MAF	136	161
<b>Total</b>	<b>917</b>	<b>2708</b>
<b>Memory usage</b>	<b>9 GB</b>	<b>257 GB</b>

# Pipeline with qcf90

## **qcf90**

- Use statement to save clean files: `--save-clean`

## **renumf90**

- Use clean SNP and map (if present) files

## **blupf90+** or other application program

- Use clean SNP and map (if present) files
- Use renumbered files from renumf90

# Pipeline with preGSf90

## **renumf90**

- Use SNP and map (if present) files

## **preGSf90**

- Use option to save clean files: `OPTION saveCleanSNPs`

## **blupf90+** or other application program

- Use clean SNP and map (if present) files
- Use renumbered files from renumf90