UNIVERSITY OF
GEORGIA
College of Agricultural &
Environmental Sciences

Quality Control of SNP data with preGSf90 and qcf90

BLUPF90 TEAM, 05/2022

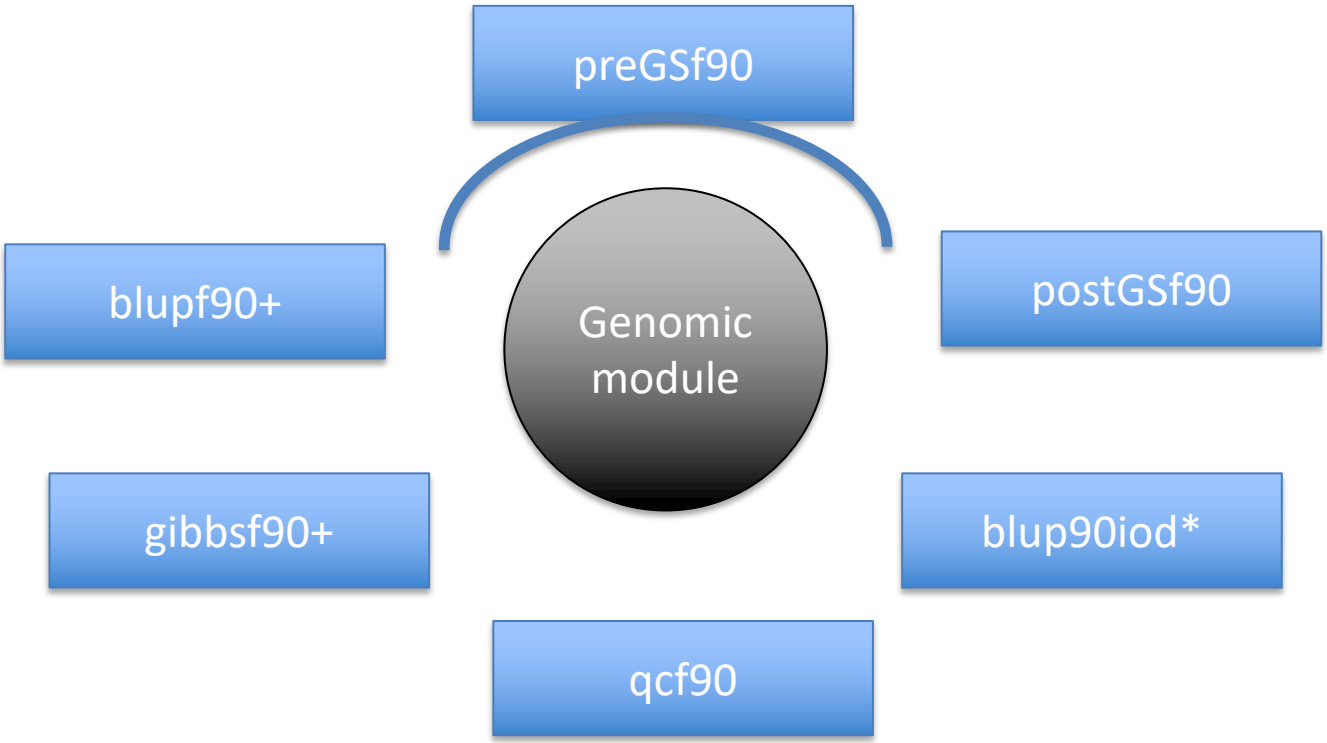
Quality control

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

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

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 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 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
  - OPTION exclusion_threshold x

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` *`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 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 h2<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 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 > 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 *marker.geno*
  - Clean fles 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 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

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

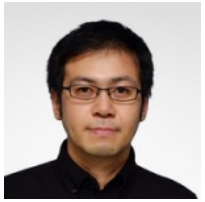
Program `PreGSF90` helps to implement the genomic selection following the single-step methodology as presented by Aguilar et al. 2010 JDS.
In this methodology the relationship matrix **A** based on the pedigree information is replaced by matrix **H**, which combines the pedigree and genomic information.

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

qcf90



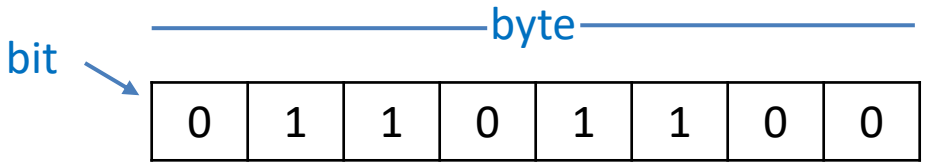
Yutaka
Masuda

- 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
  - Checks for human error or identical twins before GS

- Huge data and slow operations
  - More than 5 million genotyped Holsteins!
  - 80K SNPs x [5M x 5M] / 2 ~ 1×10^{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
AND  1010   OR   1010    XOR  1010        NOT  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

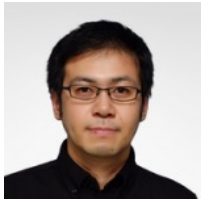
- qcf90 supports raw files
  - No need to run renumf90 before

- qcf90 was designed for QC
  - preGSf90 was designed for QC and constructing **G** and $\mathbf{A}_{22}$

- `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 markers: 60,671
  - Number of Pedigree animals: 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
Total	917	2708
Memory usage	9 GB	257 GB

pipeline

`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