BLUPf90 & PreGS and Quality Control

PreGSf90

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

- Efficient methods
  - creation of the genomic relationship matrix, relationship based on pedigree
  - Inverse of relationship matrices

- Performs Quality Control of SNP information

BLUPF90 programs using Genomic

- Genomic programs
  - controled by adding OPTIONS commands to the parameter file
  - `OPTION SNP_file` *`marker.geno.clean`*
  - Read 2 files:
    - `marker.geno.clean`
    - `marker.geno.clean_XrefID`

Output Files

- GimA22i
  - Store the content of the inv(G) – inv(A22)
  - Only if preGSf90 for runs, not in applications programs

- 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

Quality control
By default exclude

- MAF
  - SNP with MAF < 0.05
- Call rate
  - SNP with call rate < 0.90
  - Individuals with call rate < 0.90
- Monomorphic
  - Exclude monomorphic SNP. ONLY when MAF <> 0

Quality control
By default exclude (cont)

- 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 matrix !!!
- Problems in estimation of variance component 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 matrix !!!

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)

```
1 1 135098 Hapmap43437-BTA-101873
2 1 267940 ARS-BFGL-NGS-16466
3 1 393248 Hapmap34944-BES1_Contig627_
4 1 471078 ARS-BFGL-NGS-98142
5 1 516404 Hapmap53946-rs29015852
6 1 571340 ARS-BFGL-NGS-66449
7 1 845494 ARS-BFGL-BAC-32770
8 1 883895 ARS-BFGL-NGS-65067
9 1 950841 ARS-BFGL-BAC-34682
10 1 974586 ARS-BFGL-NGS-3964
11 1 1009504 ARS-BFGL-NGS-98203
12 1 1189382 ARS-BFGL-BAC-31722
13 1 1234172 ARS-BFGL-BAC-6557
14 1 1264369 ARS-BFGL-BAC-7196
15 1 1359951 Hapmap53766-ss46526150
```

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

Saving ‘clean’ files

- SNP excluded from QC are set as missing (i.e. Code=5)
- Excluded Individuals are treated as unrealated in G and A22
  - For individual *i*

    G[i,:] = 0; G[:,i]=0; G[i,i]=1 ; Same for A22

    so G-A22 will cancel out

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

Inspection of Diagonal of G

- High diagonal elements from G
  - Mislabed samples , individuals from other populations/lines
  - Problems with sample, low call rate
  - By default values >1.6 are excluded from analysis, Threshold can be changed with:

OPTION threshold_diagonal_g x

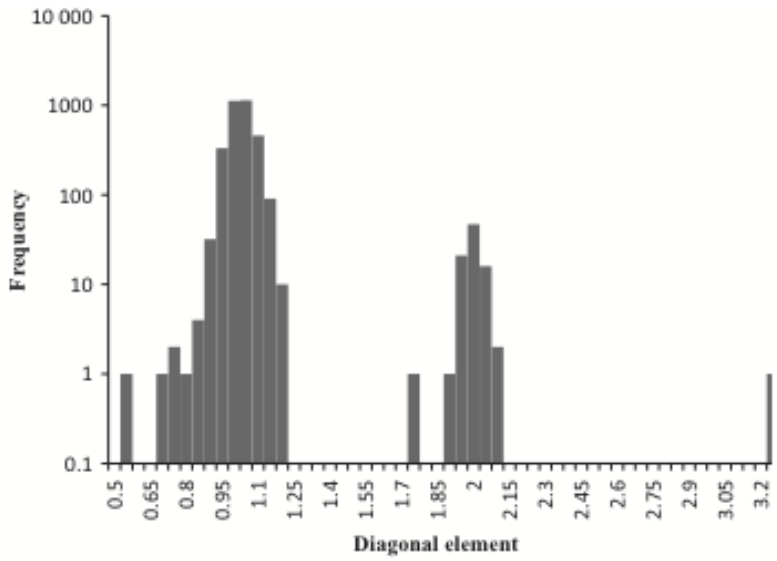


Figure 4 Distribution of the diagonal elements of G for field data. Results are shown in a logarithmic scale.

Simeone et al., 2011 JABG

Potential duplicate samples

- All samples are checked with each other using values from genomic relationship matrix
  - x = G(i,j)/sqrt(G(i,i),G(j,j))
  - Values of x > 0.90 are printed in the output

```
*****************************************
*  Possible genotype samples duplicates *
*****************************************

** i-j sample #, i-j Id, G coeff      174     167      82     860  0.9719  0.9728  0.9723  0.9993
** i-j sample #, i-j Id, G coeff      317     249     203    1144  1.0866  1.0883  1.0875  0.9988
** i-j sample #, i-j Id, G coeff      646     532     535    1398  0.9483  0.9494  0.9496  0.9987
** i-j sample #, i-j Id, G coeff     1400    1362    1652    1310  1.0108  1.0151  1.0154  0.9957
```

- Threshold to identify potential duplicates
  - OPTION threshold_duplicate_samples x
- Exclude specific samples
  - OPTION excludeSample n1 n2….

Correlation off-diagonal G vs A

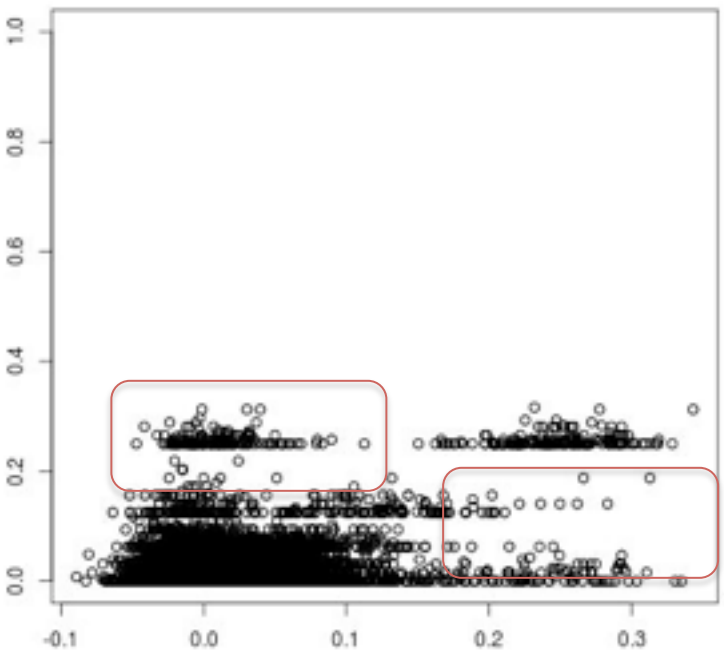
- Compute correlation for all elements of A > 0.02
- Potential problems with matching genotype file and pedigree file
- For low values (<0.5) => print a warning !!!!
- For low values (<0.3) => program stop !!!
- If still you want to go …
  - OPTION thrStopCorAG -1

```
Off-Diagonal
   Using 29494 elements from A22 >= .02000

   Estimating Regression Coefficients G = b0 11' + b1 A + e
   Regression coefficients b0 b1 =      0.514    -0.022

   Correlation Off-Diagonal elements G & A    -0.004

*************************************************************************
* CORRELATION FOR OFF-DIAGONALS G & A22 IS LOW THAN  0.50  !!!!!  *
* MISIDENTIFIED GENOMIC SAMPLES OR POOR QUALITY GENOMIC DATA *
*************************************************************************
```

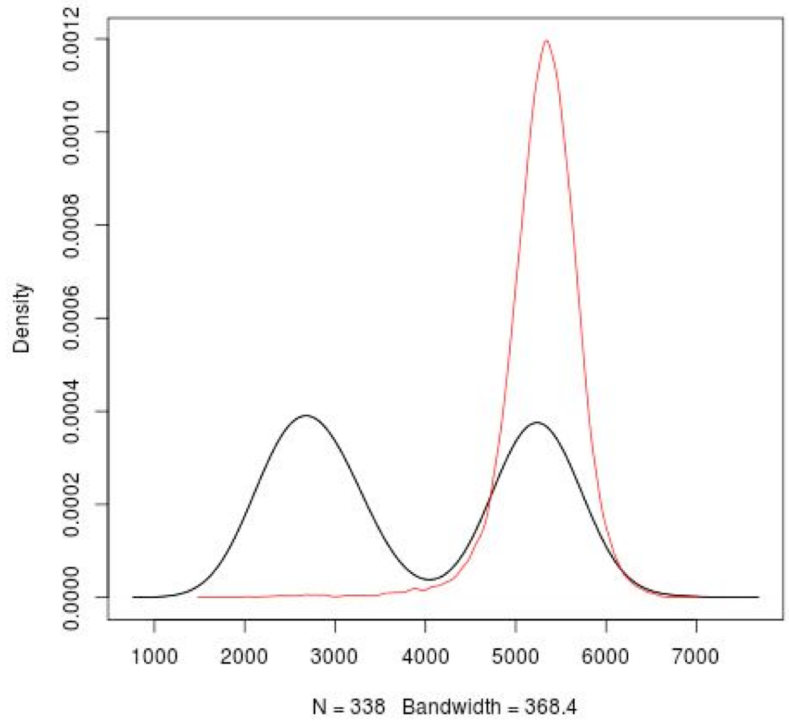
Low off-diagonal correlation Half-sibs contemporary group

Correlation off-diagonal 0.47

Off-diagonal G vs A22

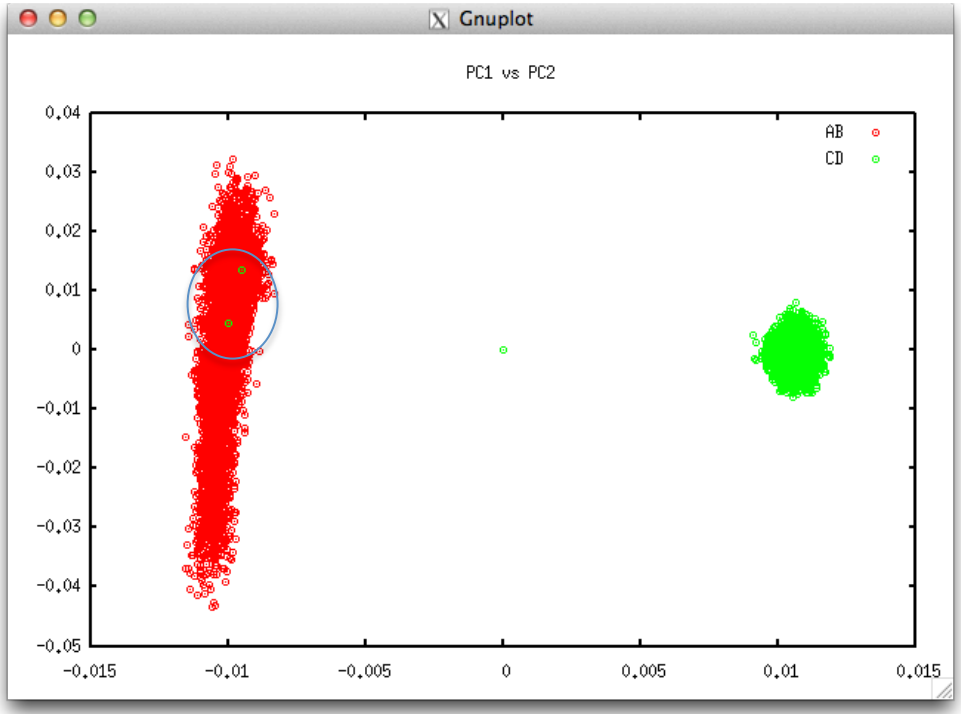


Opposite Homozygous



Looking for stratification in populations

- OPTION plotpca
  - (only preGSf90 not in application programs)
  - Plot the first 2 PC
- OPTION extra_info_pca *filename col*
  - File with variables (alphanumeric) to plot PC with different colors for different classes
  - Same order as genotype file




```
Calculating PCA
Eigenvalue Decomposition DSYEV LAPACK
Sum Eigenvalues    9672.00
First 6 PC
          Eigenvalue   % Explained
  PC:      1  2227.       23.02
  PC:      2  71.32      0.7374
  PC:      3  57.34      0.5929
  PC:      4  48.34      0.4998
  PC:      5  46.11      0.4768
  PC:      6  44.93      0.4646
```

LD calculation and options

```
OPTION calculate_LD
```

Calculate LD as Rsq

```
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

No Quality control

- ONLY use:
  - If QC was performed in a previous run
  - and “clean” genotype file is used

- OPTION no_quality_control

Options for Blending G and A

- OPTION AlphaBeta alpha beta
  - G = alpha*G^r + beta*A

- OPTION tunedG
  - 0: no adjustment
  - 1: mean(diag(G))=1, mean(offdiag(G))=0
  - 2: **mean(diag(G))=mean(diag(A)), mean(offdiag(G))=mean(offdiag(A)) (default)**
  - 3: mean(G)=mean(A)
  - 4: Use Fst adjustment. Powell et al. (2010) & Vitezica et al. (2011)

Use in application programs

- Use renumf90 for proper renumering and creation of cross reference id and parameter file
- If large number of genotypes
  - Precompute inv(G)-inv(A22) (PreGSF90)
  - Modify parameter file to read GimA22i
  - BLUPF90, AIREMLF90, GIBBSxF90….
- Generally all steps can be in a script file to facilitate running programs

PreGSf90 wiki

readme.pregsf90 [BLUPF90]

[[readme.pregsf90]] BLUPF90

Search

Trace: • readme.pregsf90

PreGSF90

Table of Contents

- PreGSF90
  - 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

`PreGSF90` is an interface program to the `genomic` module to process the genomic information for the `BLUPF90` family of programs

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`.

Input files

* Parameter file (ie `renf90.par`) as created by `RENUMF90` with genotype file specified with keyword `SNP_FILE`

* Genotype file

- Field 1 - animal ID with format as in pedigree file
- Field 2 - genotype with 0,1,2 and 5 (missing) or real values for gene content 0.12 ...

seekparentf90

- Program to check and assign paternity using genomic information

- Detect parent-offspring incompatibilities based on counts of conflicts (opposite homzygous OHM)
  - Hayes 2010 JAS
  - Wiggans et al 2010 JDS

Paternity check

Missing dam genotype
3 putatives sires

Assumption that 1 conflict is genotyping error

| | | |

seekparentf90

Usage

```
seekparentf90 --pedfile <pedigree_file_name> --snpfile <snp_file> [ ... ]
```

Optional arguments

```
--yob
```

Indicate that year of birth should be read in the 4th column.
If yob information is present, it will be used to validate a putative parent.

```
--seeksire <sire_file>
```

Indicate a list of sires that will be used to search for a parent

```
--seekdam <sire_file>
```

Indicate a list of dams that will be used to search for a parent

```
--seeksire_in_ped
```

Create a list of genotyped sires from the pedigree and use it as a list to search for a parent.

```
--seekdam_in_ped
```

Create a list of genotyped dams from the pedigree and use it as a list to search for a parent

```
--seektype n
```

Set the which animals will be used to search for a parent

Codes:

- 1: search only non-match parent (default)
- 2: search all genotyped individuals

seekparentf90

- Multiple SNP Chips

  --chips *<map file>*

SNP_ID	Chr	pos	chip1	chip2	chip3	chip4
Hapmap43437-BTA-101873	1	135098	1	1	1	1
ARS-BFGL-NGS-16466	1	267940	2	2	2	2
ARS-BFGL-NGS-19289	1	305793	3	0	3	3
ARS-BFGL-NGS-105096	1	353745	4	0	4	4
Hapmap34944-ig627_1906	1	393248	5	3	5	5
BTA-07251-no-rs	1	434180	6	0	0	0
ARS-BFGL-NGS-98142	1	471078	7	4	6	6
Hapmap53946-rs29015852	1	516404	8	5	7	7
ARS-BFGL-NGS-114208	1	533815	9	0	8	8

- SNP File

```
individual1          2 12211021111022211020200022011221211122 21
individual2          4 11122120220022222020200022000220
individual3          1 111121211100221102020002202222020022222022202102211
individual4          3 11122121110022252020 20011
```

seekparentf90

- Exclusion / inclusion SNP
  - List of SNP names to be excluded or included in parentage testing

  --exclude_snp <*file*>

  --include_snp <*file*>

- Call rate

  --thr_call_rate <*value*>

Output files

Parent_Conflicts_Stat

Parent-Progeny Mendelian Conflicts present for 36 pair animals of 1485

OrigId	#_sire	tot_#_sire	#_dam	tot_#_dam	#_ind	tot_#_ind
ccaaaaa	1	3	0	0	0	0
bbasaaa	1	24	0	0	0	1
aaaaaaa	2	4	0	0	0	2

Check_Parent_Pedigree.txt

\nimal-Sire	aaaaaaa	xxxxxx	7	25578	0.0003	145	0	145	.9943	1.0000	.9943	Match
\nimal-Sire	bbbbbbb	zzzzzz	282	25578	0.0110	0	0	0	1.0000	1.0000	1.0000	No-Match
\nimal-Sire	ccccccc	zzzzzz	4	25578	0.0002	111	0	111	.9957	1.0000	.9957	Match

Seek_Sire.txt

Seek for parent: aaaaa	Sire										
No-Match: aaaaa	xxxx	1921	25578	0.0751	779	41	814	0.9695	0.9984	0.9682
Match: aaaaa	zzzz	10	25578	0.0004	779	158	926	0.9695	0.9938	0.9638
No-Match: aaaaa	yyyy	1736	25578	0.0679	779	0	779	0.9695	1.0000	0.9695