

# 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  $\text{inv}(G) - \text{inv}(A22)$
  - 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  $MAF \neq 0$

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

```
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 1350951 Hapmap53766-rs46526150
```

# Saving 'clean' files

- SNP excluded from QC are set as missing (i.e. Code=5)
- Excluded Individuals are treated as unrelated 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 files will be:
    - $gt\_clean$
    - $gt\_clean\_XrefID$
  - Removed will be output in files:
    - $gt\_SNPs\_removed$
    - $gt\_Animals\_removed$

# 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
```

```
' i-j number of sample , i-j renumber Id, G(i,j), G(i,i), G(j,j), r(i,j) '
```

- 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 and pedigree files
- 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  $\geq .02000$

Estimating Regression Coefficients  $G = b_0 11' + b_1 A + e$   
Regression coefficients  $b_0 \ b_1 = \quad 0.514 \quad -0.022$

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 *
*****
```

# 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

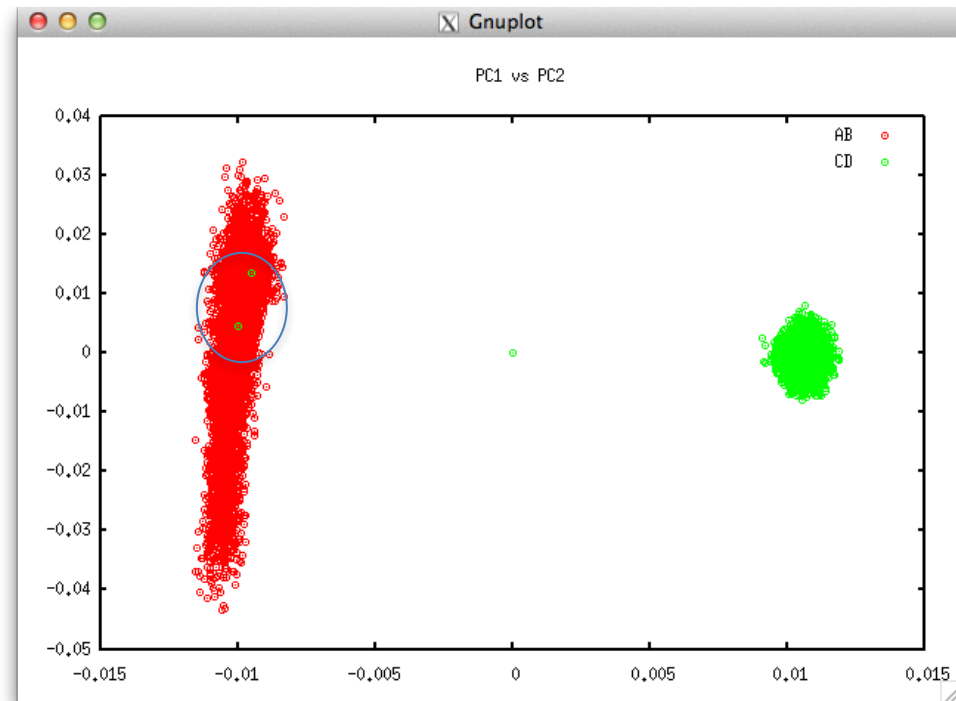
Calculating PCA

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 R<sub>sq</sub>

```
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 R<sub>sq</sub> > threshold. Optional parameter x define the threshold. default value 0.8

```
OPTION thr_output_LD [x]
```

Threshold to print out R<sub>sq</sub> between pair of SNP Optional parameter x define the threshold. default value 0.1



# preGSf90 -Only Quality control

## Atalho...

OPTION SNP\_file snp.dat

OPTION chrinfo angus\_map

OPTION excludeCHR 30 31 32

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 no\_quality\_control

# Memory requirement

- Slow operations for quality control in PREGSF90
  - All data stored in memory as double precision
  - Designed for the computation of G-matrix
  - Required memory for 60k SNPs and 500k genotyped animals = 224GB

# Comparison preGSf90 and QCF90

- Holstein genotypes
  - Number of genotypes: 569,404
  - Number of SNP markers: 60,671
  - Number of Pedigree animals: 10,710,380
- Programs
  - QCF90: with pre-renumbered files
  - PREGSF90: with post-renumbered files

# QCF90: benchmark results

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