FORMING SINGLE-STEP MIXED MODEL EQUATION AND QUALITY CONTROL

I. Aguilar

Instituto Nacional de Investigación Agropecuaria
INIA Las Brujas, Uruguay

Single-Step to genomic evaluation

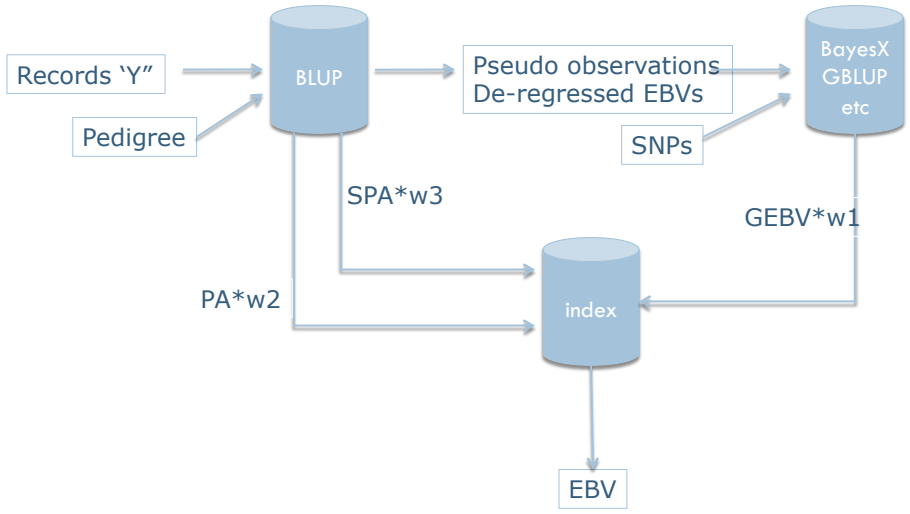
- Traditional genetic evaluation

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z+\alpha A^{-1} \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix}=\begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

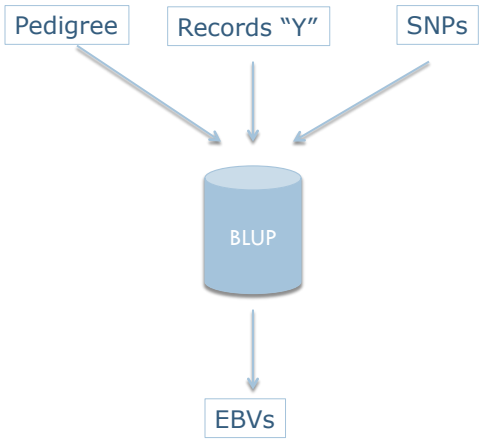
- Single-step genomic evaluation

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z+\alpha H^{-1} \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix}=\begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Multiple-step Genomic Selection



Single-Step Genomic Selection



Single-Step evaluation

- Unified approach with pedigree, phenotypic and genomic markers information considered simultaneously
- Pedigree-based relationships augmented by genomic relationship matrix (Misztal et al. 2009)

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z+\alpha H^{-1} \end{bmatrix}\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix}=\begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

$$H = A + A_{\Delta}$$

A - conventional numerator relationship matrix

A_{Δ} - matrix modified to account for genomic relationships

Single step genomic evaluation

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z+H^{-1}\alpha \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix}=\begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

- Inverses

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

Aguilar et al., 2010
Christensen & Lund, 2010

  - Numerator relationship matrix
  - Pedigree relationships between genotyped animals
  - Genomic relationships

Matrix-vector operations in PCG with genomic information

$$LHS * p = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + H^{-1}\alpha \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$$

$$= \begin{bmatrix} X'Xp_1 + X'Zp_2 \\ Z'Xp_1 + Z'Zp_2 \end{bmatrix}$$ → Contributions due to records

$$+ \begin{bmatrix} 0 \\ A^{-1}\alpha p_2 \end{bmatrix}$$ → Contributions due to relationships

$$+ \begin{bmatrix} 0 \\ 0 \\ (G^{-1} - A_{22}^{-1})\alpha p_{2g} \end{bmatrix}$$ → Contributions due to genomics

Extra matrices required for single-step

- Inverses

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

  - Pedigree relationships between genotyped animals
  - Genomic relationships

OPTIONS – BLUPF90 parameter file

- 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`

Printout: Same heading as other programs

```
Options read from parameter file:

 * SNP file: marker.geno.clean
 * SNP Xref file: marker.geno.clean_XrefID
 * Matrix in Ascii format(default=binary)

 *-------------------------------------------------------------*
 *                Genomic Library: Version 1.110               *
 *                                                             *
 *  Modified relationship matrix (H) created for effect:   2   *
 *-------------------------------------------------------------*

 Read 18600 animals from pedigree file: "renadd02.ped"
 Number of Genotyped Animals: 1500
```

All options that were enter in the parameter file should be here !!. IF not check that keywords are correct (upper and lower case)

Check number of animals and individuals with genotypes

Printout

```
Creating A22
   Extracting subset of: 4634 pedigrees from: 18600 elapsed time:      0.0019
   Calculating A22 Matrix by Colleau ...elapsed time    1.250464

Reading SNP file
   Column position in file for the first marker: 7
   Format to read SNP file: (6x,400000i1)
   Number of SNPs: 3000
   Number of Genotyped animals: 1500
   Reading SNP file elapsed time: .41

Statistics of alleles frequencies in the
   N:            3000
   Mean:        0.500
   Min:         0.101
   Max:         0.898
   Var:         0.016
```

Information from genotype file. The format is detected from the first line !!!

So all genotypes should start in the same column !!!

Number of SNP is also determined by the first line!!

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
- Parent-progeny conflicts (SNP & Individuals)
  - Exclusion -> oposite homozigous
  - For SNP: >10 % of parent-progeny exclusion from the total of pairs evaluated
  - For Individuals: > 2% 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 yet implemented
  - 3: detect and eliminate progenies with conflicts (default)

SNP map file (optional)

- OPTION chrinfo xxxx
- For some genomic analyses (GWAS) or checks
- Format:
  - snp number
    - Index number of SNP in the sorted map by chr and position
  - chromosome number
  - position
- First row corresponds to first column SNP in genotype file !!!

Other Options

- IF OPTION chrinfo is provided, we can exclude selected crhomosomes:
  - OPTION excludeCHR n1 n2 n3 ...
- or inform which are sex chromosomes:
  - OPTION sex_chr n
  - Chr > n will be excluded only for check or parent-progeny, but not in calculations

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

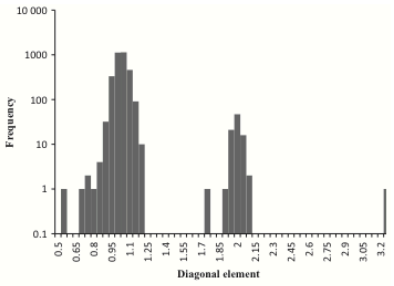


Figure 4 Distribution of the diagonal elements of G for field data. Results are shown in a logarithmic scale.

Simeone et al., 2011 JABG

Potential duplicate samples

- All samples are checked with each other
  - 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
```

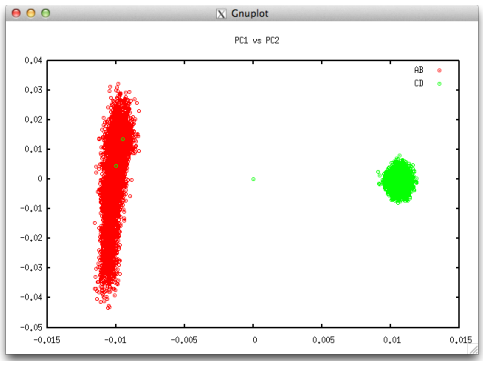
Correlation off-diagonal G vs A

- 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 (see mice example)

```
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 *
*************************************************************************
```

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
```

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, REMLF90
- Generally all steps can be in a script file to facilitate running programs