Forming Single-step mixed model equation and quality control

Ignacio Aguilar

INIA Las Brujas, Uruguay

05-2018

Single-Step to genomic evaluation

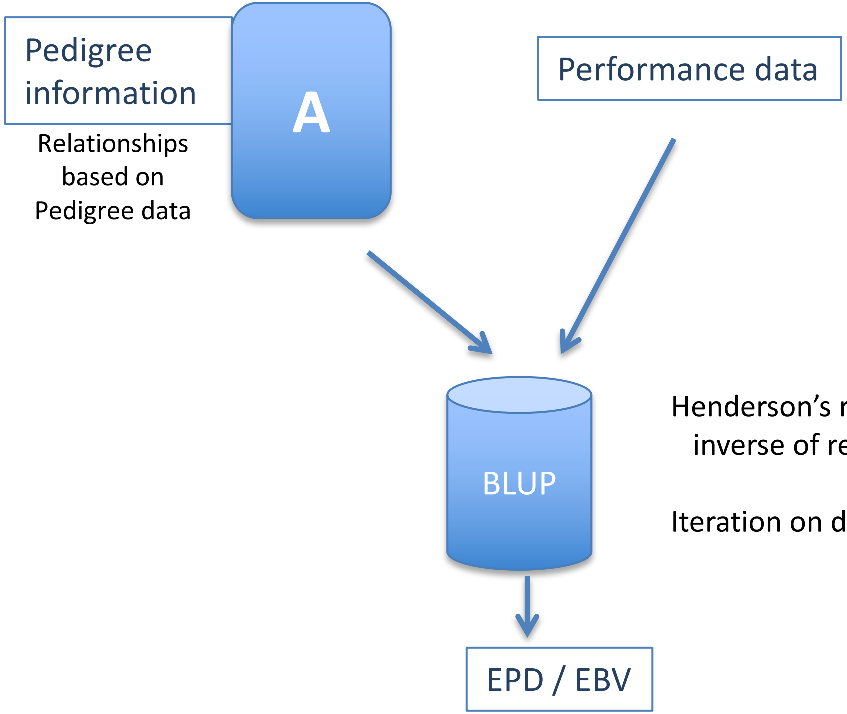
- Traditional genetic evaluation

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

- Single-step genomic evaluation

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

Genetic Evaluation



Pedigree information

A

Relationships based on Pedigree data

Performance data

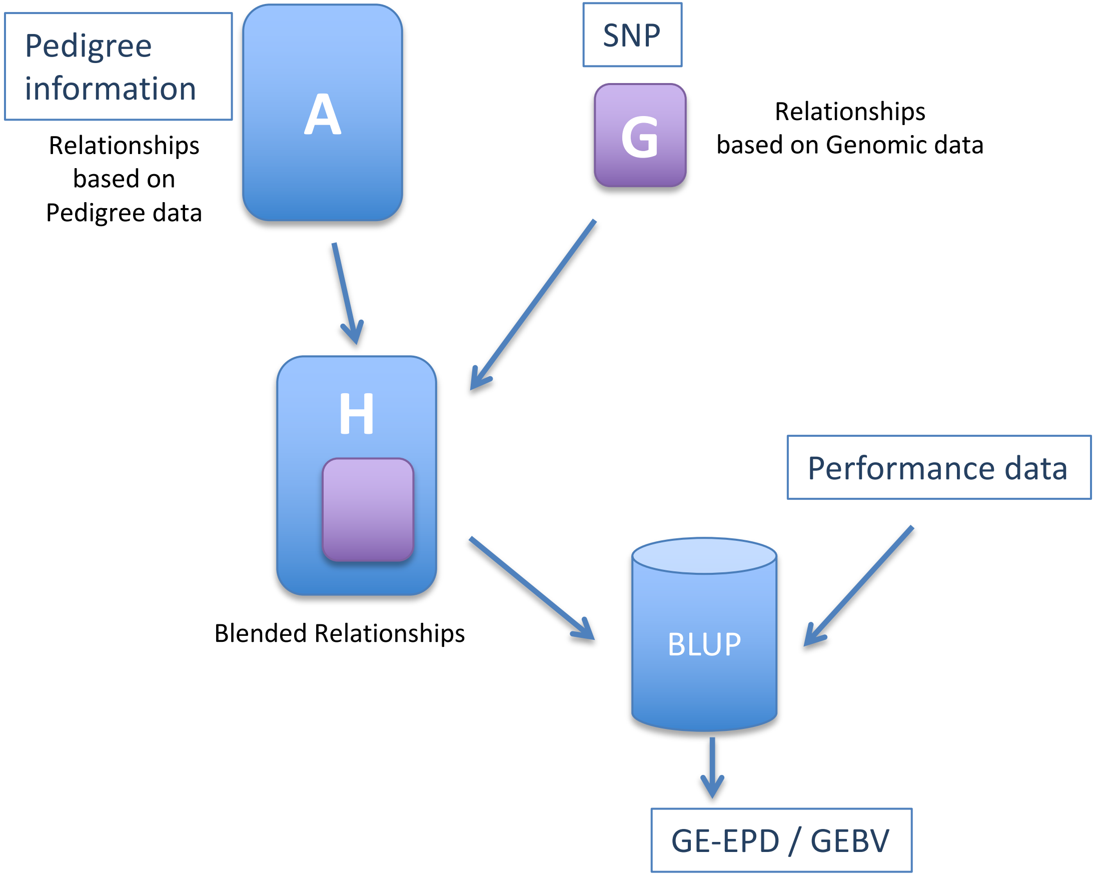
BLUP

EPD / EBV

Henderson’s rules to create inverse of relationship matrix

Iteration on data methods

Single-Step Genetic Evaluation



Pedigree information

Relationships based on Pedigree data

SNP

Relationships based on Genomic data

Blended Relationships

Performance data

GE-EPD / GEBV

Single step genomic evaluation

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

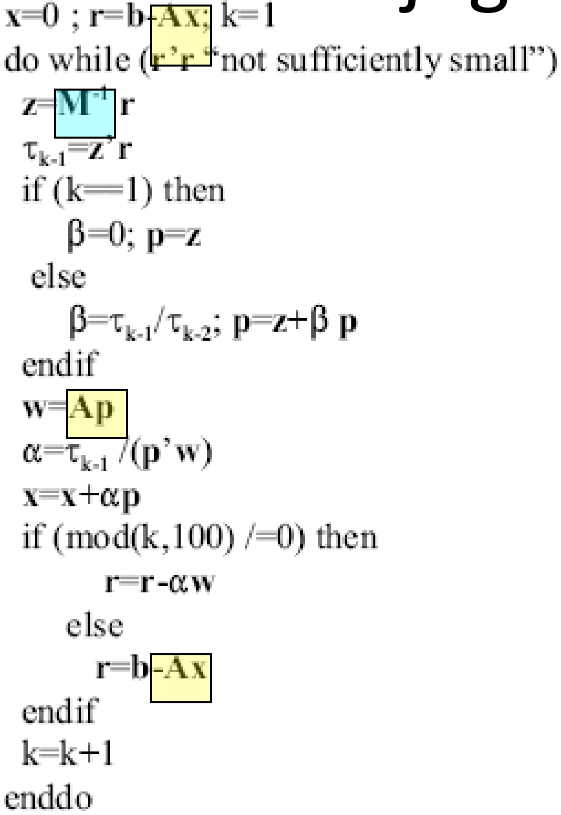
- Inverses

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

Aguilar et al., 2010
Christensen & Lund, 2010

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

Solving Ax=b by Preconditioned Conjugate Gradient



```
x=0 ; r=b-Ax; k=1
do while (r'r "not sufficiently small")
  z=$M^{-1}$r
  $\tau_{k-1}$=z'r
  if (k==1) then
     $\beta$=0; p=z
   else
     $\beta=\tau_{k-1}/\tau_{k-2}$; p=z+$\beta$ p
  endif
  w=Ap
  $\alpha=\tau_{k-1}$ /(p'w)
  x=x+$\alpha$p
  if (mod(k,100) /=0) then
        r=r-$\alpha$w
      else
        r=b-Ax
  endif
  k=k+1
enddo
```

A used only in matrix-vector product

System solved:

$\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}$

M – preconditioner

usually **M**=diag(**A**)

Berger et al (1988?)

Extra matrices required for single-step

- Inverses

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

PREGSF90

  - 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

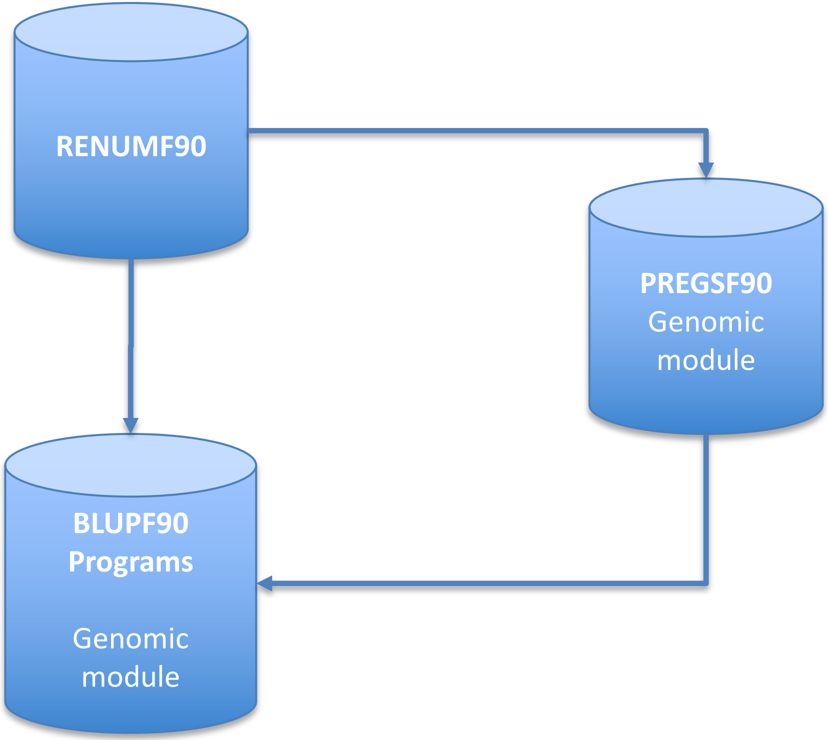
$$+ \begin{bmatrix} 0 \\ A^{-1}\alpha p_2 \end{bmatrix}$$

Contributions due to relationships

$$+ \begin{bmatrix} 0 \\ 0 \\ (G^{-1} - A_{22}^{-1})\alpha p_{2_g} \end{bmatrix}$$

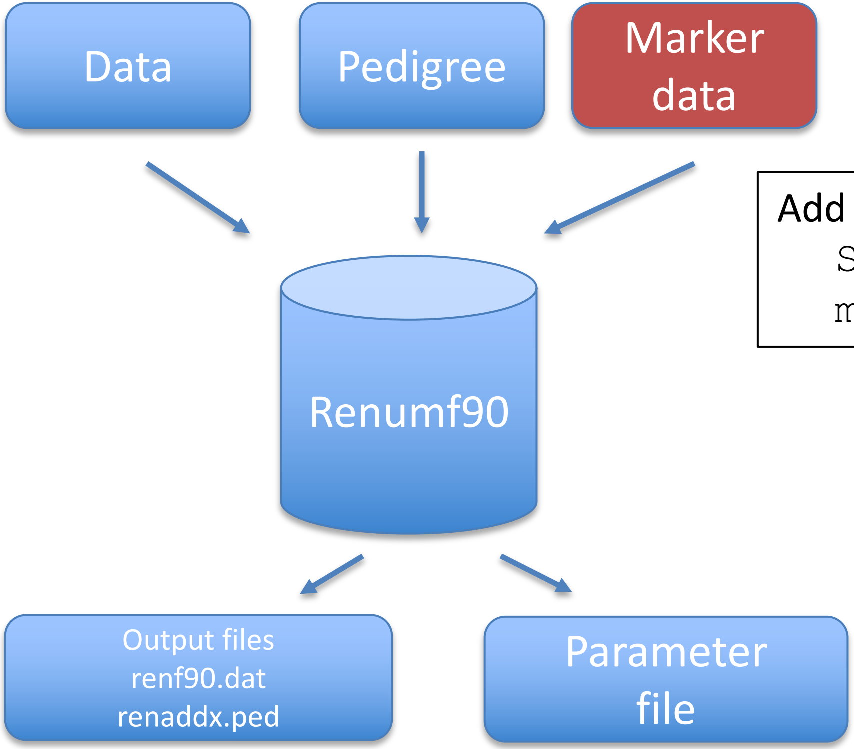
Contributions due to genomics

How BLUPF90 performs Single-Step Genomic



blupf90
(ai)remlf90
Gibbsxf90
etc.

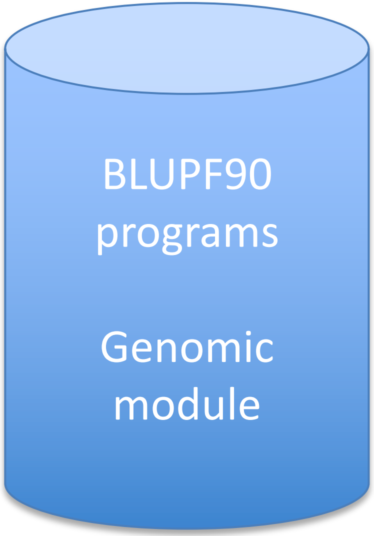
Single Step in BLUPF90 package
RENUMF90



Add keyword to the "animal effect"

```
SNP_FILE
marker_file_name
```

How BLUPF90 performs Single-Step Genomic



BLUPF90 programs

Genomic module

Genomic Module

perform quality control

create extra matrices

genomic relationship

pedigree relationship for genotyped

```
OPTION SNP_file marker.file
```

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

- Former program to performs Quality Control of SNP information

Input file for genomic BLUPf90

- Same parameter file as for all BLUPf90 programs
  - But with “`OPTION SNP_file marker_file_name`”
  - indicate to run genomic subroutines

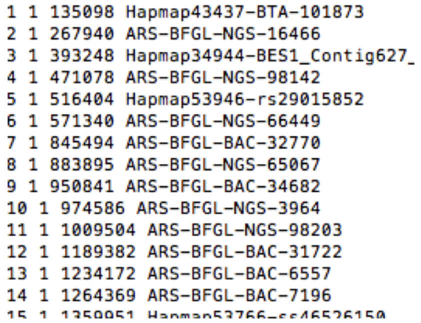
- Pedigree file

- Marker information (SNP file)

- Cross Reference file for renumber ID
  - Links genotypes files with codes in pedigree, etc.
  - Generated by renumf90

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)

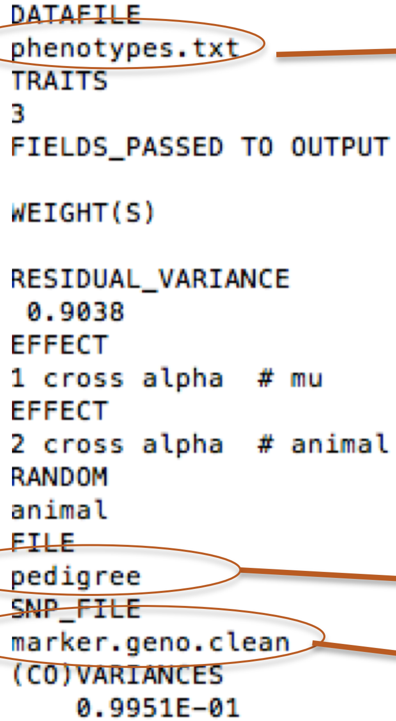


```
1 1 135098 Hapmap43437-BTA-101873
2 1 267940 ARS-BFGL-NGS-16466
3 1 393248 Hapmap34944-BES1_Contig627_
4 1 471078 ARS-BFGL-NGS-98142
5 1 516404 Hapmap53946-rs29015852
6 1 571340 ARS-BFGL-NGS-66449
7 1 845494 ARS-BFGL-BAC-32770
8 1 883895 ARS-BFGL-NGS-65067
9 1 950841 ARS-BFGL-BAC-34682
10 1 974586 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 !!!

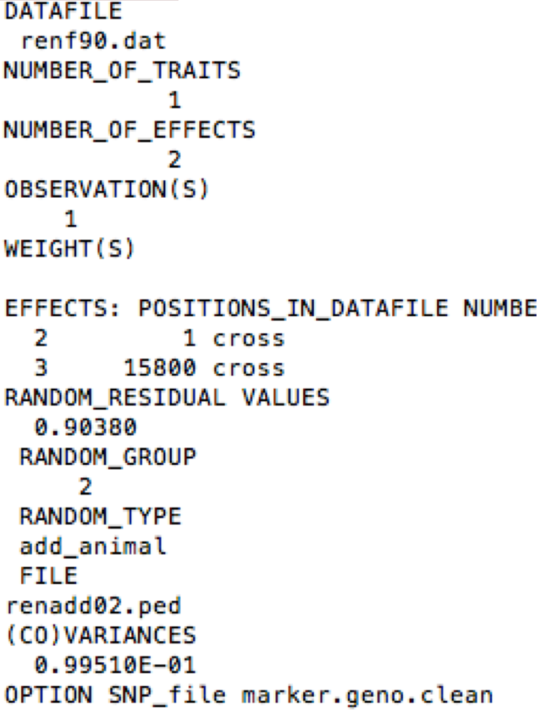
Parameters file

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.clean
(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.clean
```

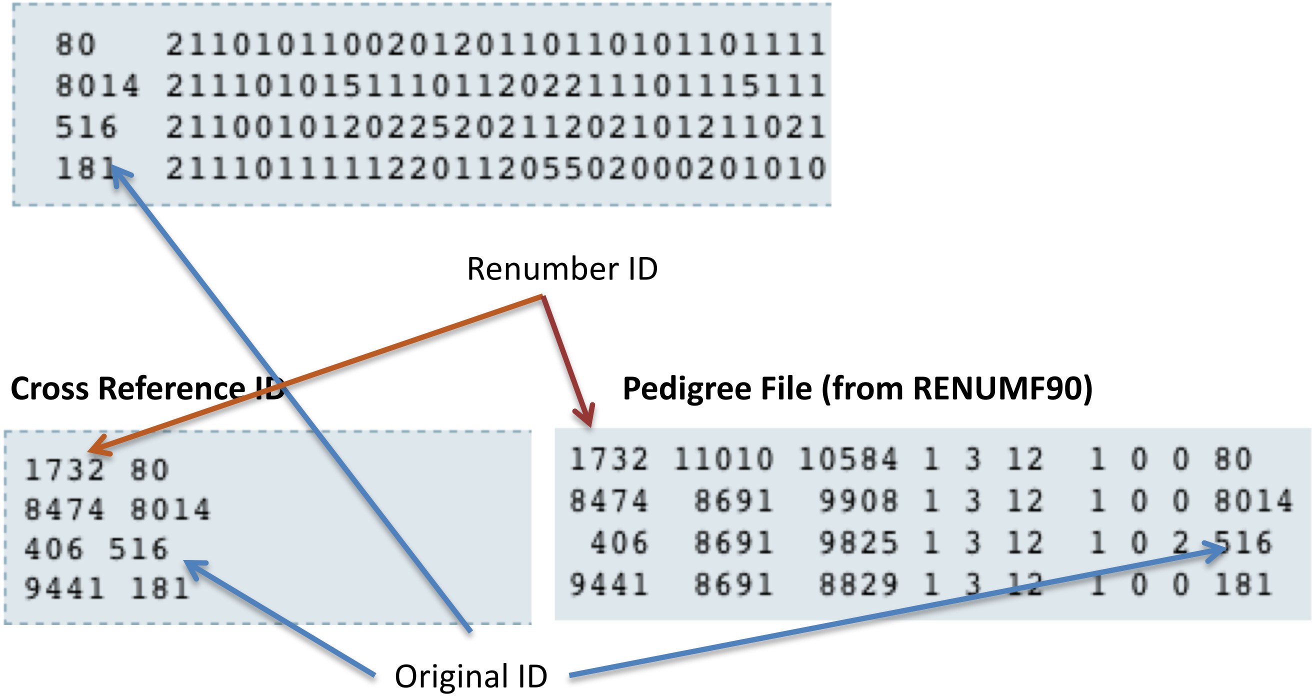
Pedigree file from RENUMF90

- 1 - **animal number**
- 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 & Cross Reference Id

SNP File First col: Identification, could be alphanumeric
Second col: SNP markers {codes: 0,1,2 and 5 for missing}



```
80   211010110020120110110101101111
8014 211101015111011202211101115111
516  211001012022520211202101211021
181  211101111122011205502000201010
```

Renumber ID

Cross Reference ID

```
1732 80
8474 8014
406 516
9441 181
```

Pedigree File (from RENUMF90)

```
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

Genomic Matrix default options

- G* = ZZ’/k as in VanRaden, 2008
- With:
  - Z center using allele frequencies estimated from the genotyped individuals
  - k = 2 sum (p * (1-p))

- G = G*0.95 + A*0.05 (to invert)

- Tunning of G (see Z. Vitezica work)
  - Adjust G to have mean of diagonals and off-diagonals equal to A_{22}

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)

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 !!!

Genomic Matrix Options

- OPTION whichfreq x
  - 0: read from file *freqdata* or other specified
  - 1: 0.5
  - 2: current calculated from genotypes (default)

- OPTION FreqFile *file*
  - Reads allele frequencies from a file

- OPTION maxsnps x
  - Set the maximum length of string for reading marker data from file => BovineHD chip

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
```

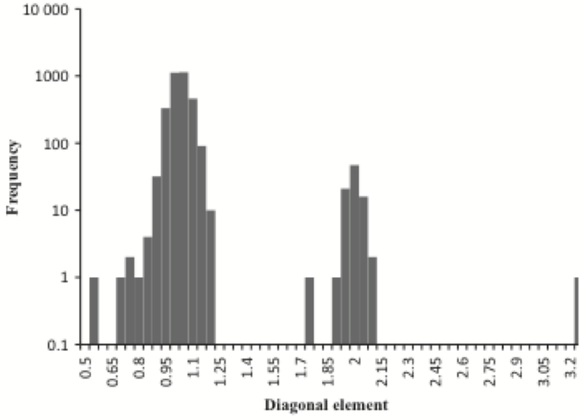
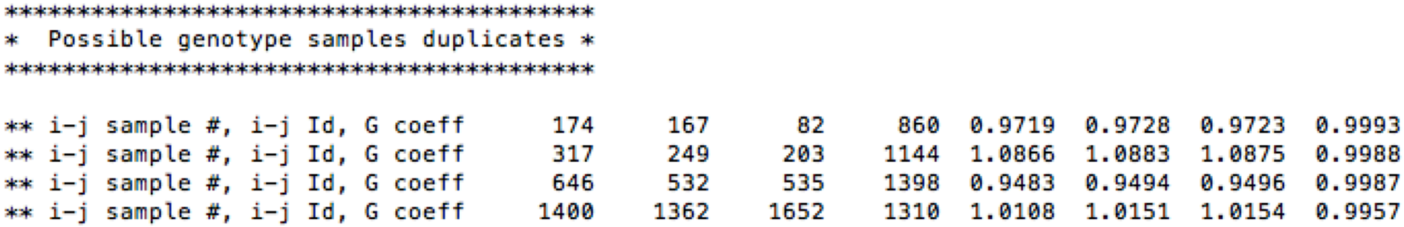


Figure 4 Distribution of the diagonal elements of G for field data. Results are shown in a logarithmic scale.

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

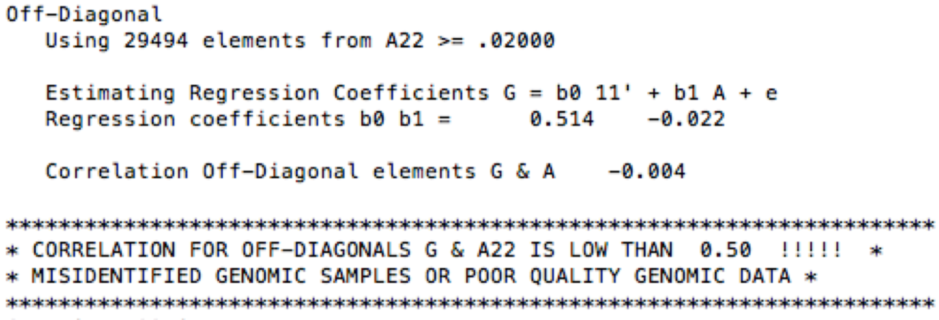


```
*****************************************
*  Possible genotype samples duplicates *
*****************************************

** 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

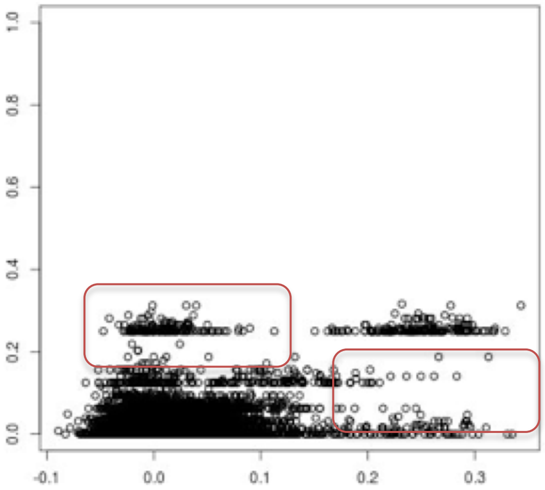


```
Off-Diagonal
   Using 29494 elements from A22 >= .02000

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

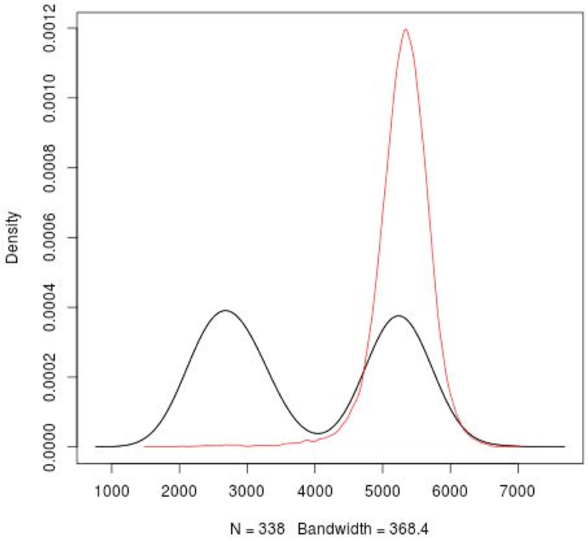
Low off-diagonal correlation
Half-sibs contemporary group

Correlation off-diagonal 0.47

Off-diagonal G vs A22

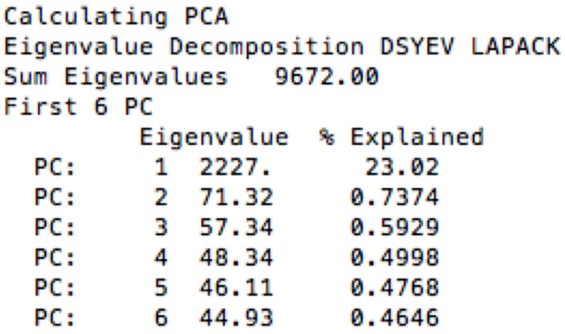


Opposite Homozygous

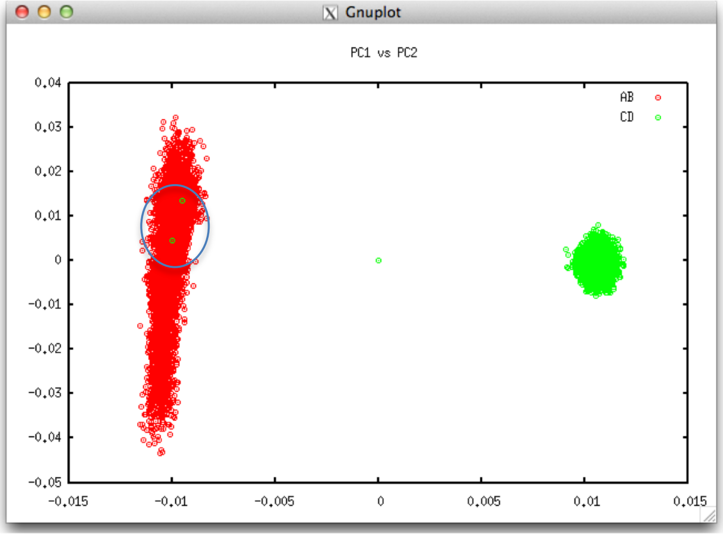


Looking for stratification in populations

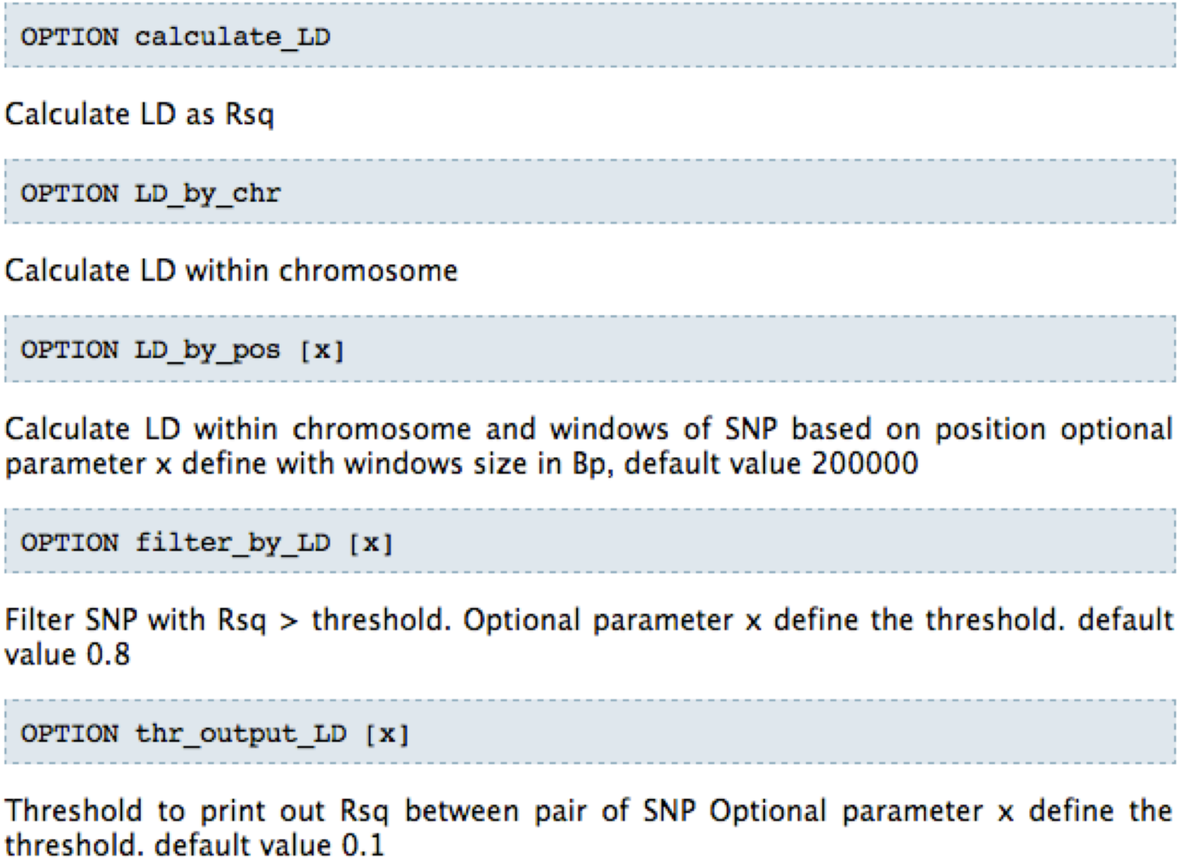
- OPTION plotpca
  - (only preGSf90 not in application programs)
  - Plot the first 2 PC
- OPTION extra_info_pca *filename col*
  - File with variables (alphanumeric) to plot PC with different colors for different classes
  - Same order as genotype file



```
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

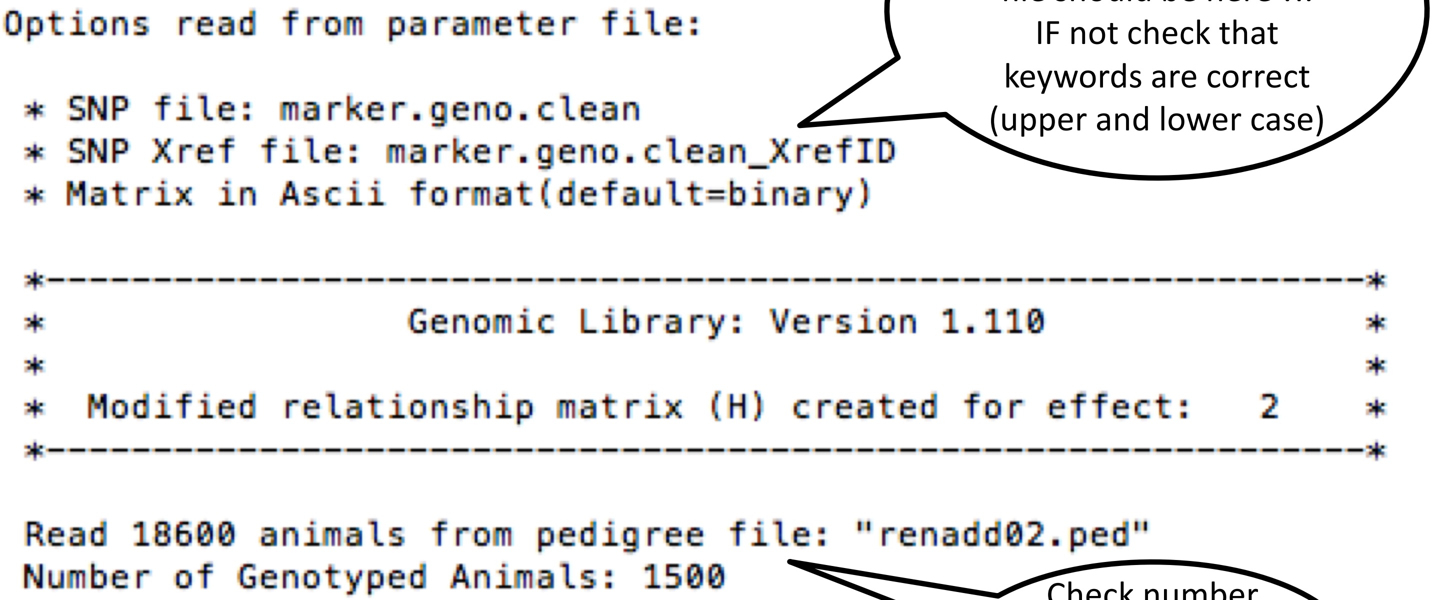
```
OPTION filter_by_LD [x]
```

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

Printout: Same heading as other programs

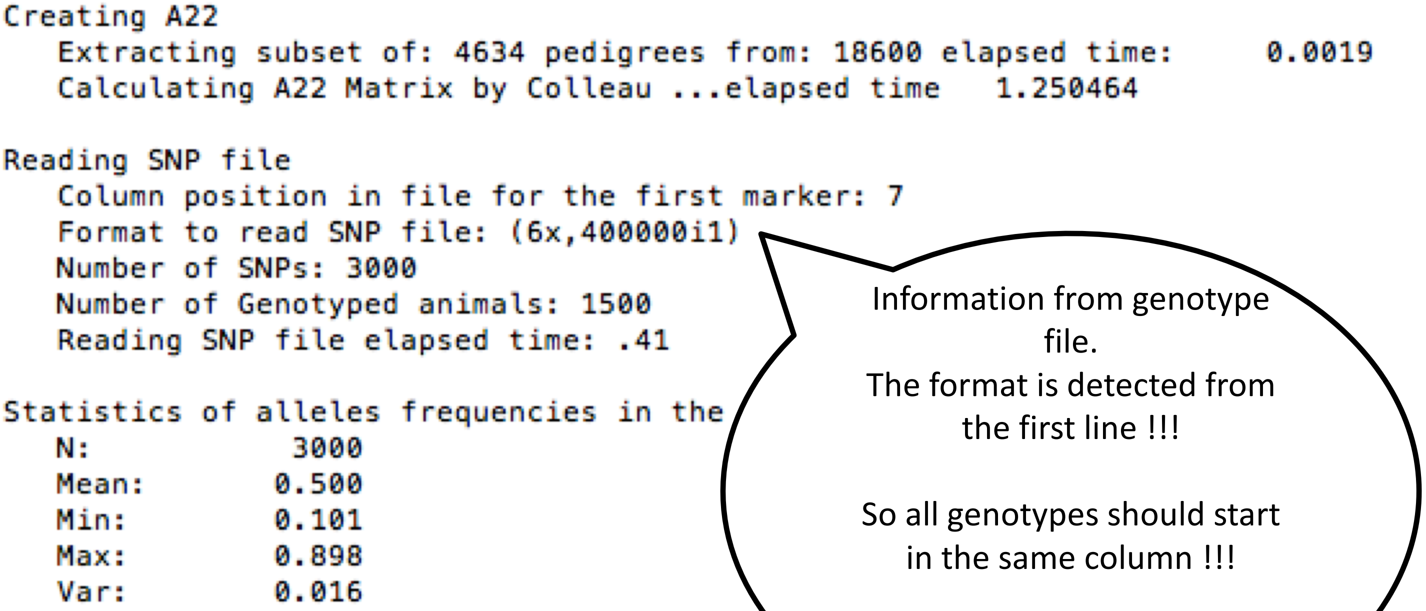


```
Options read from parameter file:

 * 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!!

No Quality control

- ONLY use:
  - If QC was performed in a previous run
    - preGSf90 or qcf90
  - and “clean” genotype file is used

- OPTION no_quality_control

Creation a subset of relationship matrix (A22)

- Create a relationship matrix for only genotyped animals (~ thousands)

- Full pedigree (~millions)

- Trace only ancestors of genotyped
  - reduce but still large number to créate A matrix by tabular method

Relationship Matrix of Genotyped Animals

- Colleau's algorithm to creates A_{22}
- No need to have explicit A matrix
- Method uses "matrix-vector" multiplication with a decomposition of A matrix

$$\mathbf{v} = \mathbf{A}\mathbf{r} = (\mathbf{I} - \mathbf{P})^{-1}\mathbf{D}(\mathbf{I} - \mathbf{P})^{-1\prime}\mathbf{r}$$

Example A times a vector

Pedigree

```
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    2    0    0
[3,]    3    1    2
```

Matrix P

```
     [,1] [,2] [,3]
[1,]  0.0  0.0  0.0
[2,]  0.0  0.0  0.0
[3,]  0.5  0.5  0.0
```

Matrix $(I\text{-}P)^{-1}$

```
     [,1] [,2] [,3]
[1,]  1.0
[2,]  0.0  1.0
[3,]  0.5  0.5  1.0
```

$$\mathbf{v} = \mathbf{Ar} = (\mathbf{I} - \mathbf{P})^{-1}\mathbf{D}(\mathbf{I} - \mathbf{P})^{-1\prime}\mathbf{r}$$

Matrix $(I\text{-}P)^{-1}$

```
     [,1] [,2] [,3]
[1,]  1.0
[2,]  0.0  1.0
[3,]  0.5  0.5    1
```

Matrix D

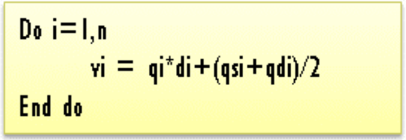
```
     [,1] [,2] [,3]
[1,]    1
[2,]         1
[3,]             0.5
```

Vector q, Matrix $(I\text{-}P)^{-1\prime}$

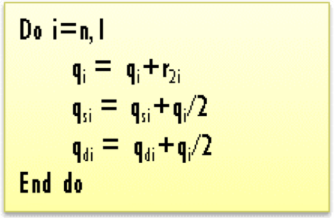
```
     [,1]              [,1] [,2] [,3]
[1,]   25         [1,]    1    0  0.5
[2,]   35 = [3,]  [2,]         1  0.5
30                [3,]            1.0
```

Vector r_2

```
     [,1]
[1,]   10
[2,]   20
[3,]   30
```



```
Do i=1,n
      vi = qi*di+(qsi+qdi)/2
End do
```

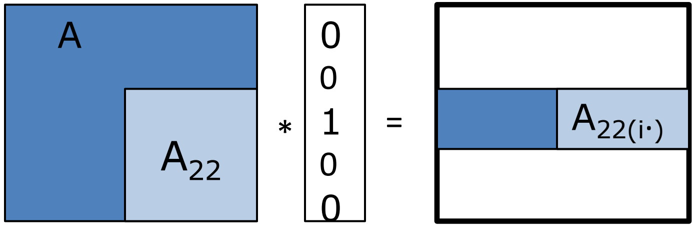


```
Do i=n,1
      $q_i = q_i + r_{2i}$
      $q_{si} = q_{si} + q_i/2$
      $q_{di} = q_{di} + q_i/2$
End do
```

Relationship Matrix of Genotyped Animals

- For each genotyped animal in A_{22}

$$\mathbf{v} = \mathbf{A}\mathbf{r}_2 = \mathbf{(I - P)^{-1}D(I - P)^{-1'}r_2}$$



- Calculation of relationship for each animal can be done in parallel with OpenMP

Tabular method vs. Colleau algorithm

- Testing
  - 6,500 genotyped Holsteins
  - 57,000 pedigrees

	Tabular*	Colleau method
CPU Time	311 s	45 s
Memory	12.1GB	322MB

* Gmatrix.f90 (VanRaden, 2009)

Storing and Reading Matrices

- PreGSF90:
  - Facilitate the implementation of single-step
  - Matrix A is replaced by H with:

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

  - Default output is the matrix GimA22i, to be included in apllication programs (BLUPF90, REMLF90..)

- BUT: intermediate matrices could be stored for examination, use in application programs, etc.

Storing and Reading Matrices

- Matrices that can be stored:
  - A22, inv(A22), G, inv(G), GmA22, inv(GmA22), inv(H)
- All matrices are stored in same format:
  - upper triangle
  - By default in binary format
  - But to store in text (Ascii) format:
    - Use: `OPTION saveAscii`
- Values
  - i j val
  - i & j refers to the row number in the genotype file !!!!!
  - Renumber ID could be obtained from the XrefID file

Storing and Reading Matrices

To save our ‘raw’ genomic matrix:

- OPTION saveG [all]
  - If the optional *all* is present all intermediate G matrices will be saved!!!

or it inverse

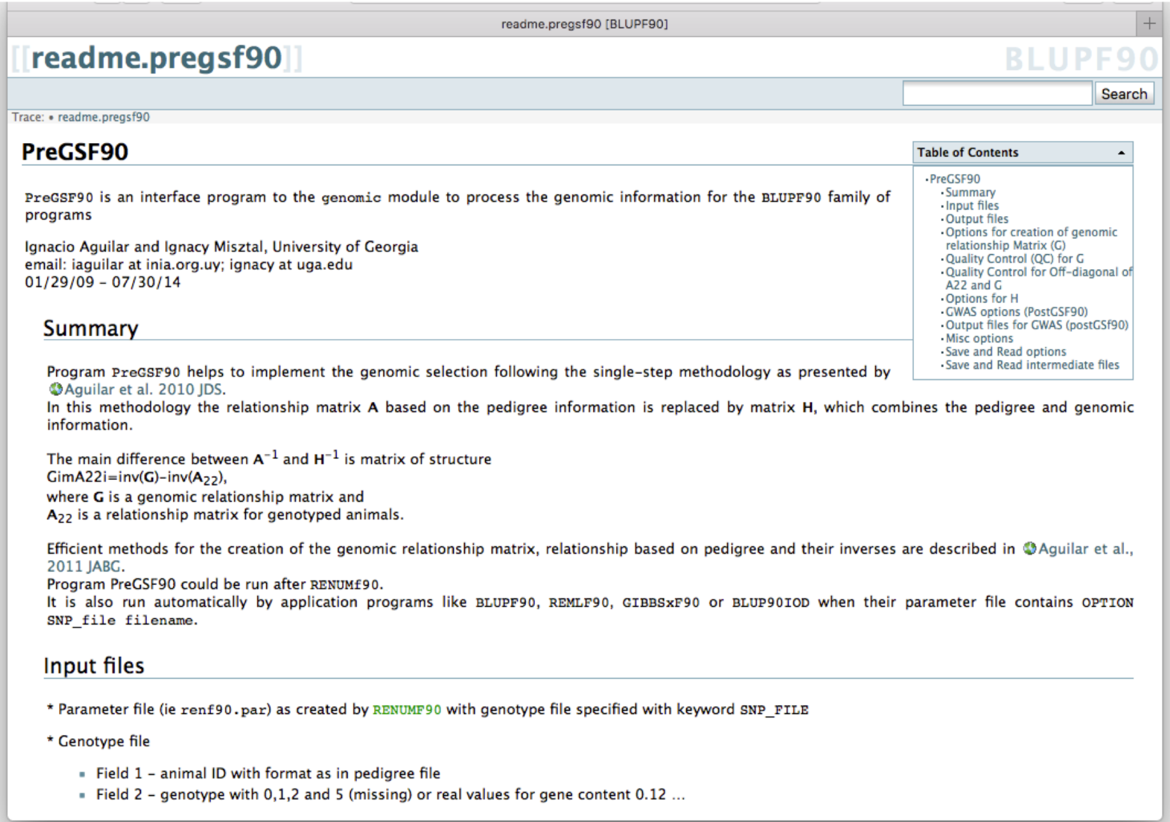
- OPTION saveGInverse
  - Only the final matrix G, after blending, scaling, etc. is inverted !!!

- Look in wiki for keywords for other matrices

Storing with Original IDs

- Some matrices could be stored in text files with the original IDs extracted from *renaddxx.ped* created by the RENUMF90 program (col #10)
- For example:
  - OPTION saveGOrig
  - OPTION saveDiagGOrig
  - OPTION saveHinvOrig
- Values
  - origID_i, origID_j, val

PreGSf90 wiki



PreGSF90

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 …