Constructing A^{-1}

- Pedigree relationships: $u_i = 0.5(u_{s_i} + u_{d_i}) + \varphi_i$

$$\mathbf{A^{-1} = (I - P)'M^{-1}(I - P)}$$

Henderson (1976)
Quaas (1988)

```
ImP=(/1., -.5, -.5/)
Minv=(/2., 4/3., 1., 0./)
ainv=0.0

  do
   read(2,*,iostat=io) animal, sire, dam,par_stat
   if (io /= 0) exit
   p(1)=animal
   p(2)=sire
   p(3)=dam
    do i=0,nanimal
      do k=1,3
        do l=1,3
          ainv(k,l)=ainv(k,l)+ImP(k)*ImP(l)*Minv(par_stat)
        enddo
      enddo
    enddo
```

par_stat:
3 - # of known parents

Theory of GBLUP and single-step GBLUP

Daniela Lourenco

BLUPF90 TEAM – 08/2024

UNIVERSITY OF GEORGIA

College of Agricultural & Environmental Sciences

Animal Breeding and Genetics Group

UNIVERSITÀ DEGLI STUDI FIRENZE

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Statistical methods before genomics

- BLUP (Henderson, 1949 - 1976)
  - Best: minimizes MSE
  - Linear: linear function of the data
  - Unbiased: $E(u) = E(\hat{u})$
  - Prediction: for random effects

Statistical Science
1991, Vol. 6, No. 1, 15–51

That BLUP Is a Good Thing: The Estimation of Random Effects

G. K. Robinson

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Henderson's MME

- Model

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Wu} + \mathbf{e}$$

- Joint probability of phenotypes and EBV

$$p(\mathbf{y}, \mathbf{u}) = p(\mathbf{u}|\mathbf{y})\, p(\mathbf{y}) = p(\mathbf{y}|\mathbf{u})\, p(\mathbf{u})$$

- Joint probability density function of phenotypes and EBV

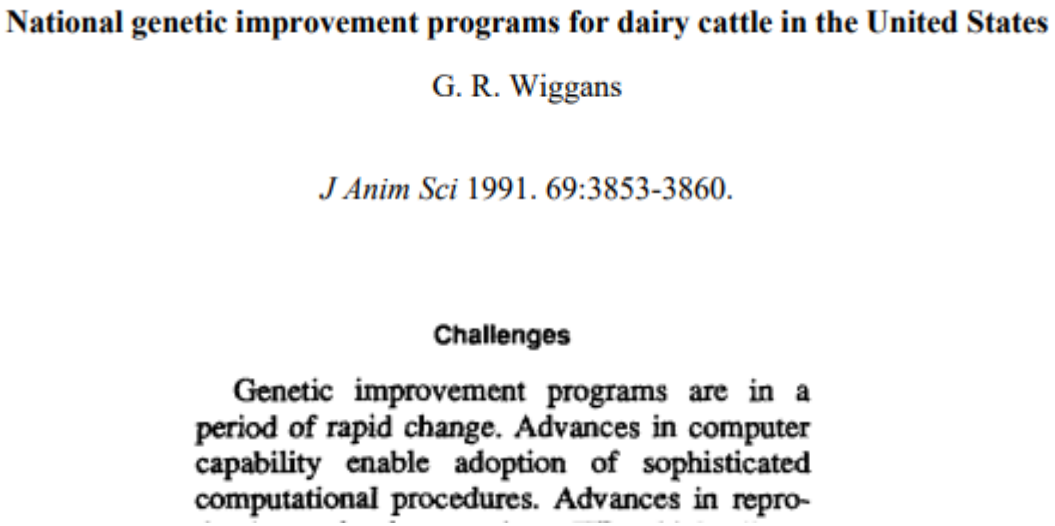
$$p(\mathbf{y}, \mathbf{u}) = p(\mathbf{y}|\mathbf{u})\, p(\mathbf{u}) = \frac{1}{\sqrt{2\pi|\mathbf{R}|}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Wu})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Wu})} \frac{1}{\sqrt{2\pi|\mathbf{G}|}} e^{-\frac{1}{2}(\mathbf{u}-\mathbf{0})'\mathbf{G}^{-1}(\mathbf{u}-\mathbf{0})}$$

$$\begin{cases} \mathbf{X'R^{-1}X\beta} + \mathbf{X'R^{-1}Wu} = \mathbf{X'R^{-1}y} \\ \mathbf{W'R^{-1}X\beta} + (\mathbf{W'R^{-1}W} + \mathbf{G^{-1}})\mathbf{u} = \mathbf{W'R^{-1}y} \end{cases}$$

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} + \mathbf{A^{-1}}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

Henderson's MME for dairy in 1989

- BLUP (Henderson, 1949 - 1976)
- Implementation for dairy in 1989



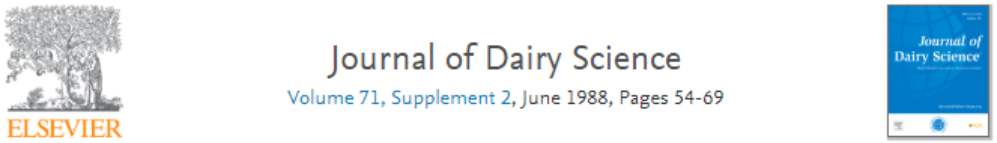
National genetic improvement programs for dairy cattle in the United States

G. R. Wiggans

J Anim Sci 1991. 69:3853-3860.

Challenges

Genetic improvement programs are in a period of rapid change. Advances in computer capability enable adoption of sophisticated computational procedures. Advances in repro-



ELSEVIER

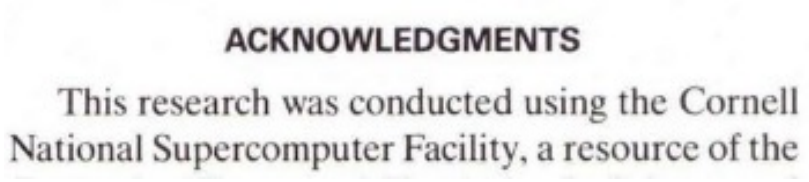
Journal of Dairy Science

Volume 71, Supplement 2, June 1988, Pages 54-69

Implementation of an Animal Model for Genetic Evaluation of Dairy Cattle in the United States

G.R. Wiggans, I. Misztal, L.D. Van Vleck

- 9.5 M animals
- 11 M lactations
- 23.5 M equations to solve
- 7.5 hours

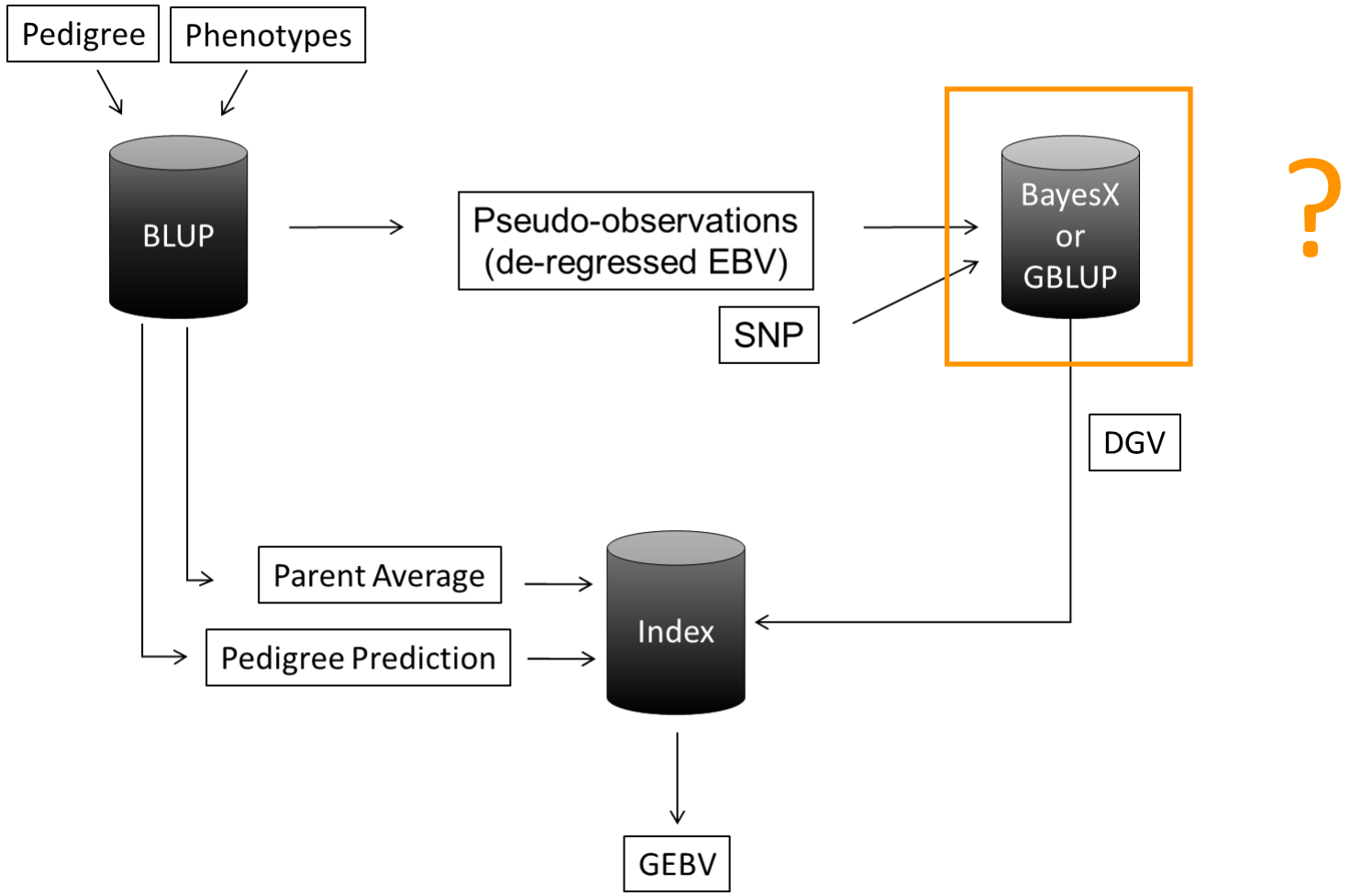


ACKNOWLEDGMENTS

This research was conducted using the Cornell National Supercomputer Facility, a resource of the

Moving from 1989 to 2009

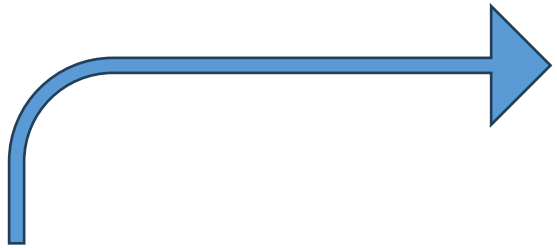
- How to add genomic information to the evaluation system in 2009?



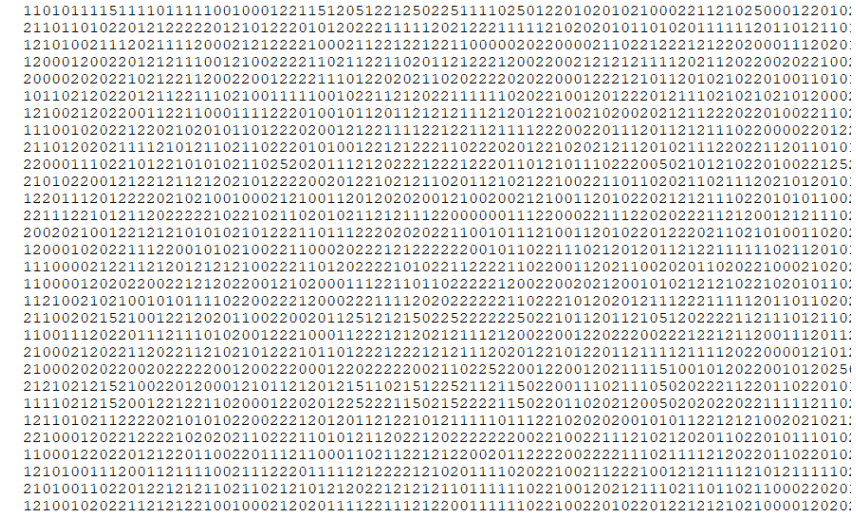
Multistep

SNP-BLUP (ridge regression)

- SNP effect model = outputs SNP effects
- $a \sim N(0, \sigma_a^2)$



$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \boldsymbol{e}$$



$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{I}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

$$\mathbf{GEBV} = \mathbf{Z\hat{a}}$$

- All SNP explain the same proportion of variance on the trait

SNP-BLUP (ridge regression)

- SNP effect model = outputs SNP effects
- All SNP explain the same proportion of variance on the trait

$$\mathbf{GEBV} = \mathbf{Z}\hat{\mathbf{a}}$$

$$\mathbf{u} = \mathbf{Z}\hat{\mathbf{a}}$$

$$\mathrm{Var}(\mathbf{u}) = \mathrm{Var}(\mathbf{Za})$$

$$\mathrm{Var}(\mathbf{u}) = \mathbf{Z}\,\mathrm{Var}(\mathbf{a})\,\mathbf{Z}'$$

$$\mathrm{Var}(\mathbf{u}) = \mathbf{ZZ}'\sigma_a^2$$

$$\sigma_a^2 = \frac{\sigma_u^2}{2\sum_{i=1}^{SNP} p_i(1-p_i)}$$

$$\mathrm{Var}(\mathbf{u}) = \mathbf{ZZ}'\frac{\sigma_u^2}{2\sum_{i=1}^{SNP} p_i(1-p_i)}$$

$$\mathrm{Var}(\mathbf{u}) = \frac{\mathbf{ZZ}'}{2\sum_{i=1}^{SNP} p_i(1-p_i)}\sigma_u^2$$

Genomic relationship matrix VanRaden (2008)

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum_{i=1}^{SNP} p_i(1-p_i)}$$

$$\mathrm{Var}(\mathbf{u}) = \mathbf{G}\sigma_u^2$$

GBLUP assumption!!!

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

GBLUP: Genomic BLUP

- GEBV-based model = outputs genomic breeding values (GEBV)
- $\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \boldsymbol{e}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

Fernando & Grossman (1989)
Bernardo (1994)
Nejati-Javaremi et al. (1997)

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum p_i(1-p_i)}$$

VanRaden (2008)

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

GBLUP: Genomic BLUP

$$\mathbf{G} = \frac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)}$$

```
025   1101011115111101111100100012211512051221250225111102501220102010210002211210250001220102
036   2110110102201212222201210122201012022211111202122211111210202010110102011111201101211012
050   1210100211120211112000212122221000211221221221100000202200002110221222121220200011202012
054   1200012002201212111001210022221102112211020112122212002200212121211112021120220020221002
066   2000020202210212211200220012222111012202021102022220202200012221210112010210220100110102
097   1011021202201211221110210011111001022112120221111102022100120122201211102102102101200021
101   1210021202200112211000111122201001011201121212111212012210021020020212112222022010022110
151   1110010202212202102010110122202001212211112212211211112220022011120112121110220000220122
172   2110120202111121012110211022201010012212122211022202012210202121120102111220221120110101
224   2200011102210122101010211025202011121202221222122201101210111022200502101210220100221252
277   2101022001212212112120210122220020122102121102011210212210022110110202110211120210120101
314   1220111201222202102100100021210011201202020012100200212100112010220212121110220101011002
419   2211122101211202222210221021102010211212111220000001112200022111220202221121200121211102
439   2002021001221212101021012221101112220202022110010111210011201022012220211021010011020
456   1200010202211122001010210022110002022212122222001011022111021201201121221111102112010
501   1110000212211212012121210022211012022221010221122221102200112021100202011020221000210202
571   1100001202022002212120220012102000111221101102222120022002021200101021212102210201011
579   1121002102100101011110220022212000222111120202222221102221012020121112221111201101102
581   2110020215210012212020110022002011251212150225222225022101120112105120222211211101211
657   1100111202201112111010200122210001122212120212111212002200122022200222122121120011201112
660   2100021202211202211210210122210110122212221212111202012210122011211112111112022000012101
730   2100020202200202222200120022200012202222200211022522001220012021111510010120220010120250
732   2121021215210022012000121011212012151102151225211211502200111021110502022211220110220101
764   1111021215200122122110200012202012252221150215222115022011020212005020202202111112110
780   1211010211222202101010220022212012011212210121111101112210202020010101122121210020210212
800   2210001202212222102020211022211010121120221202222220022100221112102120201102201011010
816   1100012202201212201100220111211000110211221212200201122220022221110211112120220110220102
832   1210100111200112111100211122201111121222212102011110202210021122210012121111210121111102
900   2101001102201221212110211021210121202212121211011111102210012021211102110110211000220201
901   1210010202211212122100100021202011112211121220011111102210022010220122121210210000120202
```

Genomic relationship matrix

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)} = \frac{(\mathbf{M}-\mathbf{2P})(\mathbf{M}-\mathbf{2P})'}{2\sum p_i(1-p_i)}$$

Genotypes {0,1,2}

Shifted to refer to the average of a population with allele frequencies *p*

Scaled to refer to the genetic variance of a population with allele frequencies *p*

Each SNP contributes $2pqa^2$ to the genetic variance

http://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf

https://doi.org/10.3390/genes11070790

Genomic relationship matrix

Genotypes {0,1,2}

Shifted to refer to the average of a population with allele frequencies *p*

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)} = \frac{(\mathbf{M}-\mathbf{2P})(\mathbf{M}-\mathbf{2P})'}{2\sum p_i(1-p_i)}$$

Scaled to refer to the genetic variance of a population with allele frequencies *p*

If base allelic frequencies are used, **G** is an unbiased and efficient estimator of IBD realized relationships

Pedigree vs. Genomic relationships

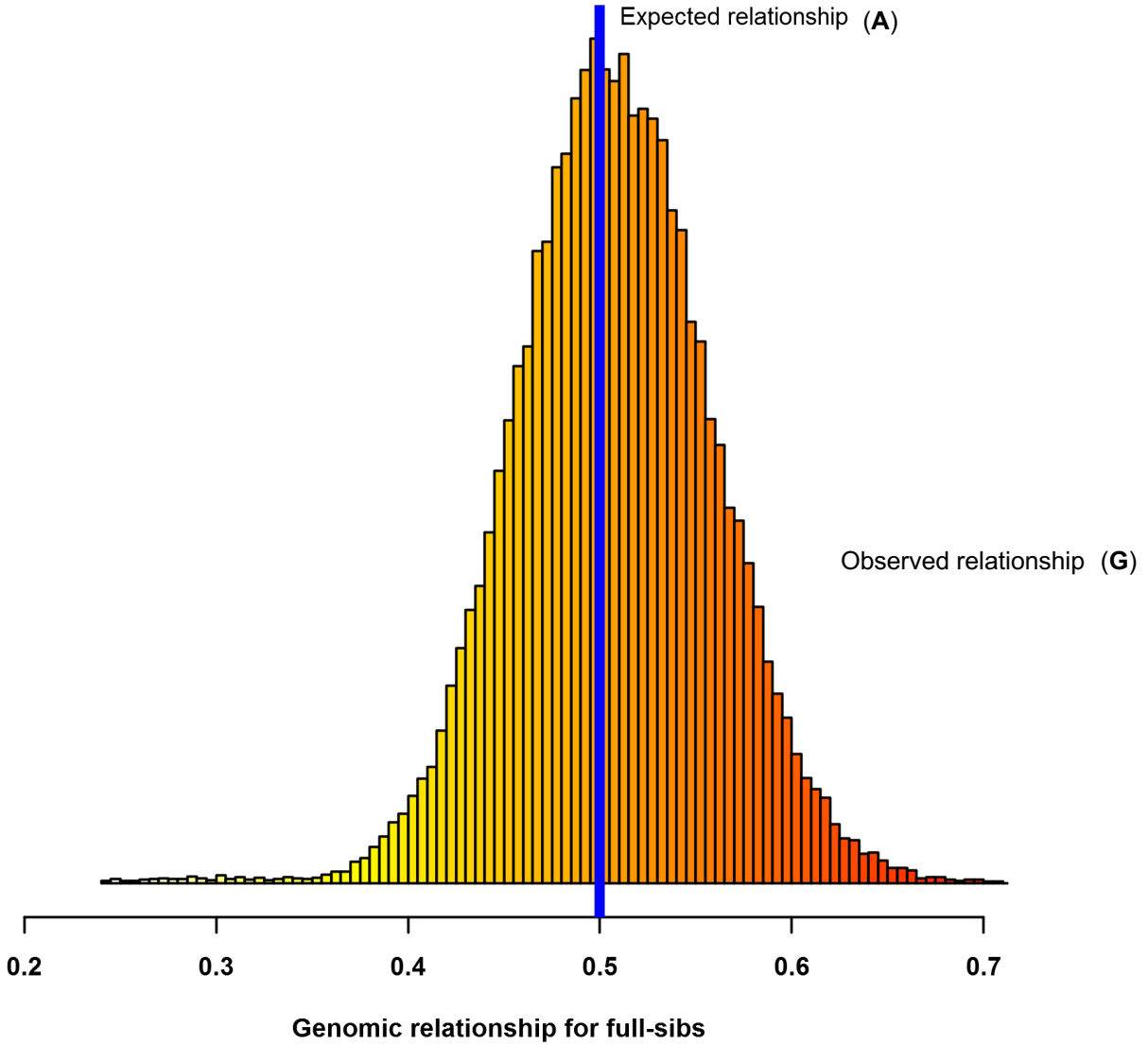
- **A** provides Identical By Descent relationships (IBD)
- **G** provides Identical by State relationships (IBS)

- **A** is the expectation of realized or observed relationships
- **G** contains realized or observed relationships

- SNPs more informative than **A**
  - Two full sibs might have a correlation of 0.4 or 0.6

- Many markers needed to better estimate relationships
  - IBS as on estimator of IBD

Pedigree vs. Genomic relationships



Adapted from Lourenco et al. (2015)

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Some “interesting” properties of **G**

- VanRaden (2008)
  - **G** can be singular if few SNP or identical genotypes (twins)
  - **G** must be singular if number of individuals > number of SNP

$$\mathbf{G} = 0.95\frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)} + 0.05\mathbf{I} \quad \text{OR} \quad \mathbf{G} = 0.95\frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)} + 0.05\mathbf{A} \quad \rightarrow \quad \mathbf{G} = \alpha\mathbf{G}_0 + \beta\mathbf{A}$$

- Blending ≈ Adding a residual polygenic effect

Some “interesting” properties of **G**

- For all matrices of the kind $$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)} = \frac{(\mathbf{M}-\mathbf{2P})(\mathbf{M}-\mathbf{2P})'}{2\sum p_i(1-p_i)}$$
  - We don’t need to put the same *p*’s in the upper and and in the lower part
- Changing allele frequencies in $\boldsymbol{P}$ shifts EBV’s by a constant
  - Irrelevant if there is an overall mean or fixed effect in the model (Stranden and Christensen, 2011)
- Changing allele frequencies in $\frac{1}{2\sum p_i q_i}$ “scales”

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

GBLUP

- GEBV-based model = outputs genomic breeding values (GEBV)
- $\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \boldsymbol{e}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

Only for genotyped individuals!!!

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum p_i(1-p_i)}$$

VanRaden (2008)

Not all individuals are genotyped

- Genomic evaluation would be simpler if all individuals were genotyped
- What to do when there are genotyped and non-genotyped individuals?
  - SNPs are capturing relationships
  - Pedigrees give information about relationships
  - Genomic and pedigree relationships can be combined in a single matrix!

Non-genotyped

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

Genotyped

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G} \end{bmatrix}$$

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

Misztal et al., 2009

Not all animals are genotyped

- Genomic info can be extended to non-genotyped animals
  - joint distribution of EBV for non-genotyped (u_1) and genotyped (u_2)

$$p(u_1, u_2) = p(u_2)p(u_1|u_2)$$

Legarra et al., 2009

$$\mathbf{H} = \begin{pmatrix} var(u_1) & cov(u_1, u_2) \\ cov(u_2, u_1) & var(u_2) \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{pmatrix}$$

Error in the prediction

Variance of predicting genotypes for non-genotyped animals

Prediction generates a covariance

$$\mathbf{H} = \left[\begin{array}{c|c} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \hline \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{array}\right]$$

Relationships from genotypes

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Some properties of **H**

- Always semi-positive definite
  - eigenvalues are always positive or zero
- Positive definite & invertible if **G** is invertible
- In practice, if **G** is too different from $\mathbf{A}_{22}$ (wrong pedigree or genotyping)
  - Numerical problems
- If no one is genotyped, Single-step is BLUP
- If everyone is genotyped, Single-step is GBLUP

Realized relationship matrix (**H**)

Animal	Sire	Dam
1	0	0
2	0	0
3	1	2
4	1	2

Pedigree Relationship Matrix (**A**)

$$\begin{bmatrix} 1.0 & 0.0 & 0.5 & 0.5 \\ . & 1.0 & 0.5 & 0.5 \\ . & . & 1.0 & 0.5 \\ . & . & . & 1.0 \end{bmatrix}$$

Genomic Relationship Matrix (**G**) for animals 3 and 4

$$\begin{bmatrix} 1.0 & 0.52 \\ . & 1.0 \end{bmatrix}$$

Realized Relationship Matrix (**H**)

$$\begin{bmatrix} 1.004 & 0.0 & 0.507 & 0.507 \\ . & 1.004 & 0.507 & 0.507 \\ . & . & 1.0 & 0.52 \\ . & . & . & 1.0 \end{bmatrix}$$

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Understanding **H**

- It is a projection of **G** on the rest of the individuals so that **G** makes sense
  - e.g., parents of two animals related in **G** should be related in **A**

- It is a Bayesian update of the pedigree matrix based on new information from genotypes

- Typically
  - **A** in the millions
  - **G** and $\mathbf{A}_{22}$ in the thousands
  - Leads to a very efficient method of genomic evaluation:
    - **Single Step GBLUP**

Single-step Genomic BLUP (ssGBLUP)

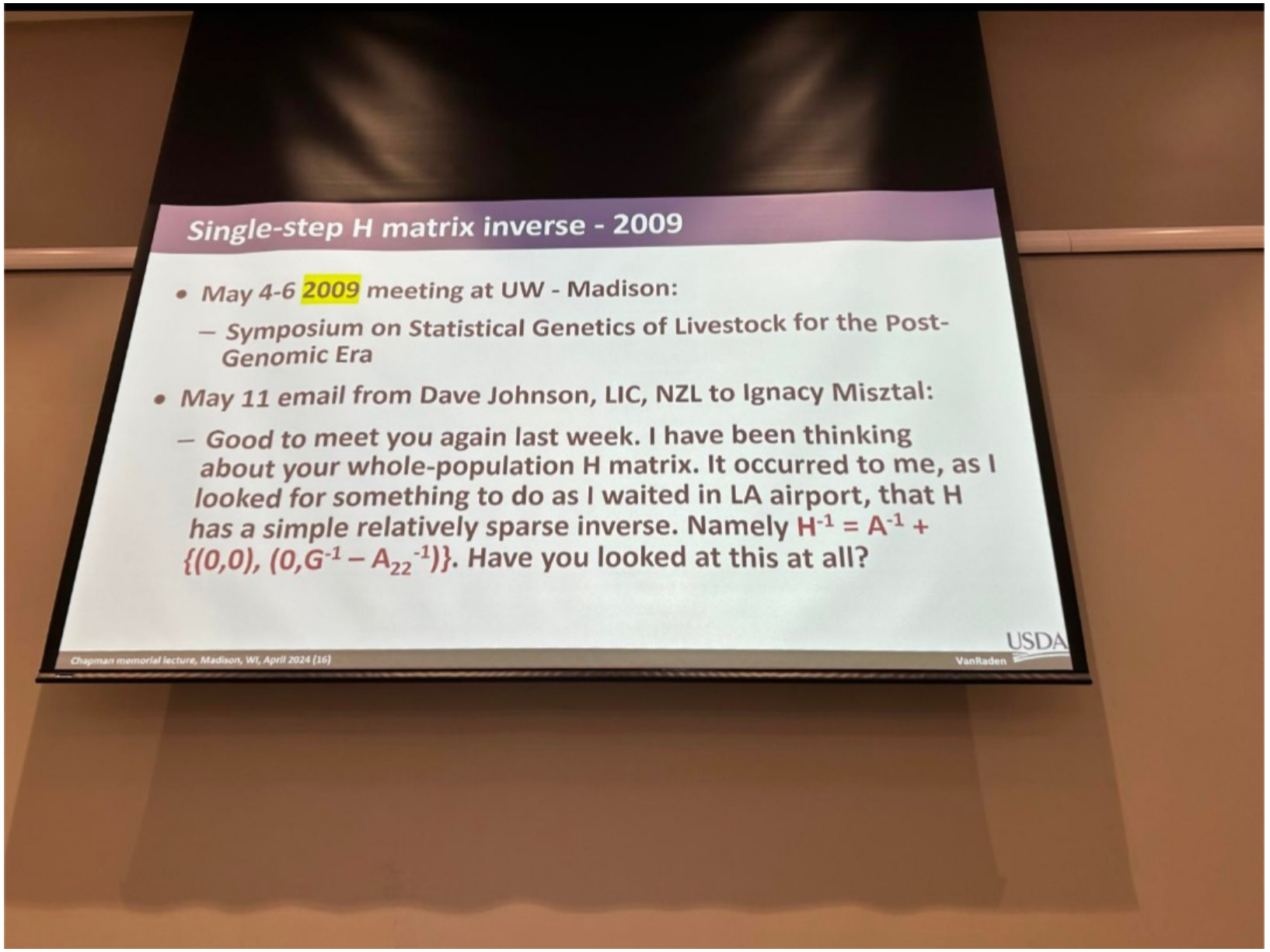
- Because not all animals are genotyped
  - 5% to 15% in large populations

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{H^{-1}}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

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

Aguilar et al., 2010
Christensen and Lund, 2010

Single-step Genomic BLUP (ssGBLUP)



ELSEVIER

Journal of Dairy Science
Volume 93, Issue 2, February 2010, Pages 743-752

Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score [1]

I. Aguilar [*] [†], I. Misztal [*], D.L. Johnson [‡], A. Legarra [§], S. Tsuruta [*], T.J. Lawlor [#]

Combining two sources of relationships

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G}-\mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G}-\mathbf{A}_{22}) \\ (\mathbf{G}-\mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}-\mathbf{A}_{22} \end{bmatrix}$$

- **A**
  - Contains expected relationships
  - Is limited by the pedigree depth and completeness
  - Depends on accuracy of recording pedigrees

- **G**
  - Contains number of alleles shared between animals weighted by heterozygosity
  - No limitations regarding to the number of past generations
  - Depends on allele frequency and quality of genomic data

Combining two sources of relationships

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

Computed using Henderson-Quaas' algorithm with inbreeding

Computed using VanRaden's formula, which considers inbreeding

Computed using Colleau's algorithm, which considers inbreeding

- Tuning
  - Base of **G** is *genotyped* animals
  - Base of **A** is *founders of the pedigree*
  - For SSGBLUP, Vitezica et al. 2011 modeled a mean in genotyped animals:

$$p(\boldsymbol{u}_2) = N(\mathbf{1}\mu, \mathbf{G})$$

Integrate μ : $\mathbf{G}^* = a + b\mathbf{G}$

μ = (Pedigree base) – (Genomic base)

Tries to put **G** and **A** in the same scale

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Single-step

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{H}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

ssGBLUP

Misztal et al. (2009)
Legarra et al. (2009)
Aguilar et al. (2010)
Christensen & Lund (2010)

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'ZM} & \mathbf{X}_n'\mathbf{Z}_n \\ \mathbf{M'Z'X} & \mathbf{M'Z'ZM}+\mathbf{I}\frac{\sigma_e^2}{\sigma_\alpha^2} & \mathbf{M}_n'\mathbf{Z}_n'\mathbf{Z}_n \\ \mathbf{Z}_n'\mathbf{X}_n & \mathbf{Z}_n'\mathbf{Z}_n\mathbf{M}_n & \mathbf{Z}_n'\mathbf{Z}_n+\mathbf{A}^{nn}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{M'Z'y} \\ \mathbf{Z}_n'\mathbf{y}_n \end{bmatrix}$$

ssSNPBLUP or **ssBR**

Fernando et al. (2014)
Liu et al. (2014)
Mantysaari & Stranden (2016)

Fernando *et al. Genetics Selection Evolution* 2014, **46**:50
http://www.gsejournal.org/content/46/50

equation (3) results in the usual non-genomic MME for the BVM.

Theory underlying SSBV-BLUP

Legarra et al. [11] proposed an ingenious strategy to combine information from genotyped and non-genotyped animals in a single BLUP analysis based on a BVM, which we refer to as SSBV-BLUP. Suppose **g** is partitioned as:

$$\mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{T}_2\boldsymbol{\alpha} \end{bmatrix},$$

J. Dairy Sci. 101:10082–10088
https://doi.org/10.3168/jds.2018-14913

Short communication: **Genomic prediction using different single-step methods in the Finnish red dairy cattle population**

H. Gao,*†[1] M. Koivula,‡ J. Jensen,* I. Strandén,‡ P. Madsen,* T. Pitkänen,‡ G. P. Aamand,† and E. A. Mäntysaari‡
*Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark
†Nordic Cattle Genetic Evaluation, DK-8200 Aarhus, Denmark
‡Natural Resources Institute Finland (Luke), FIN-31600 Jokioinen, Finland

We confirmed that regular ssGBLUP and ssBR with an extra polygenic effect led to the same predictions.

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Bases for Genomic Predictions

Bases for Genomic Prediction

Andres Legarra Daniela A.L. Lourenco Zulma G. Vitezica

2022-05-11

http://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

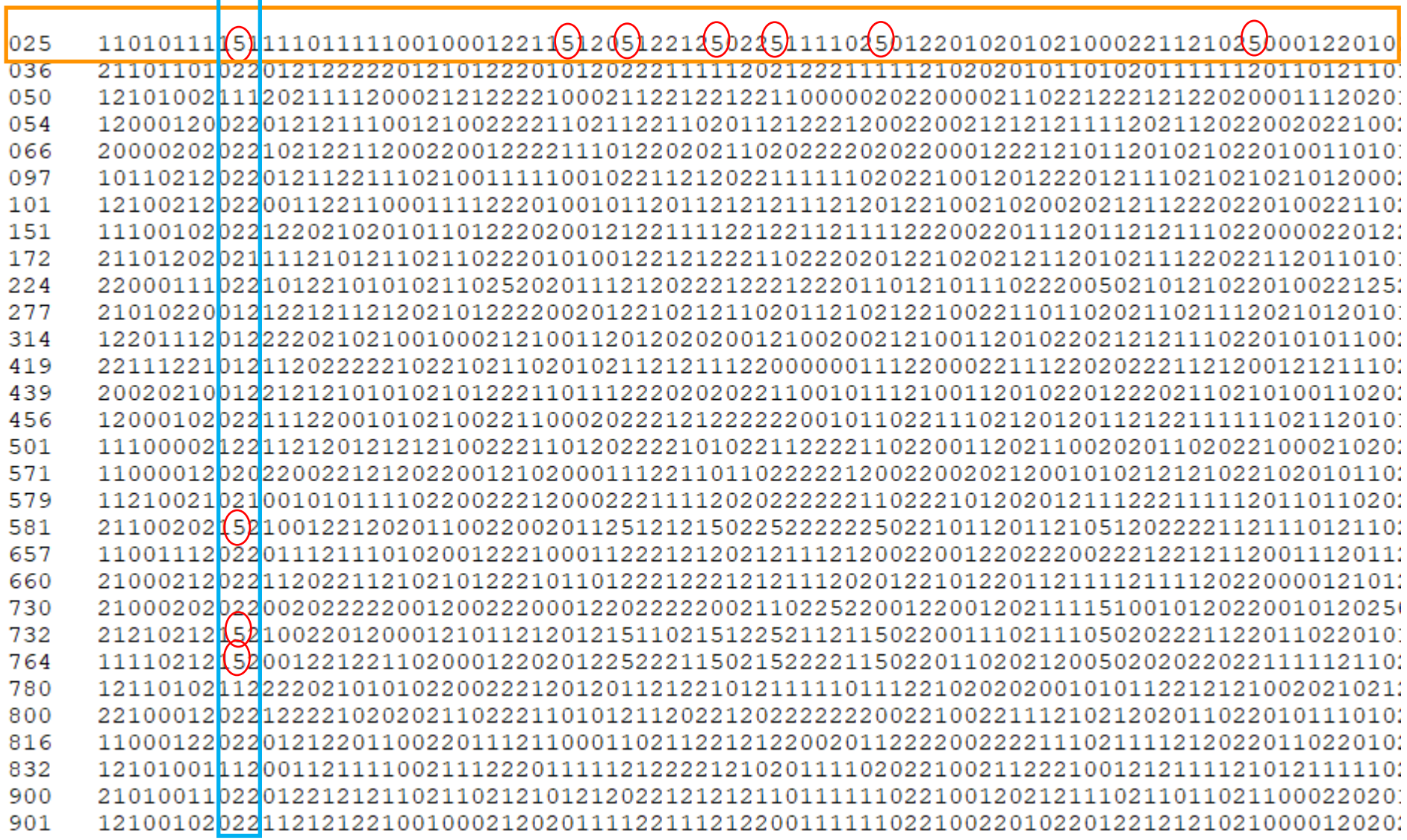
Quality Control of SNP data and creation of genomic matrices with BLUPF90 software

BLUPF90 TEAM, 08/2024

SNP file

SNP

ANIMAL



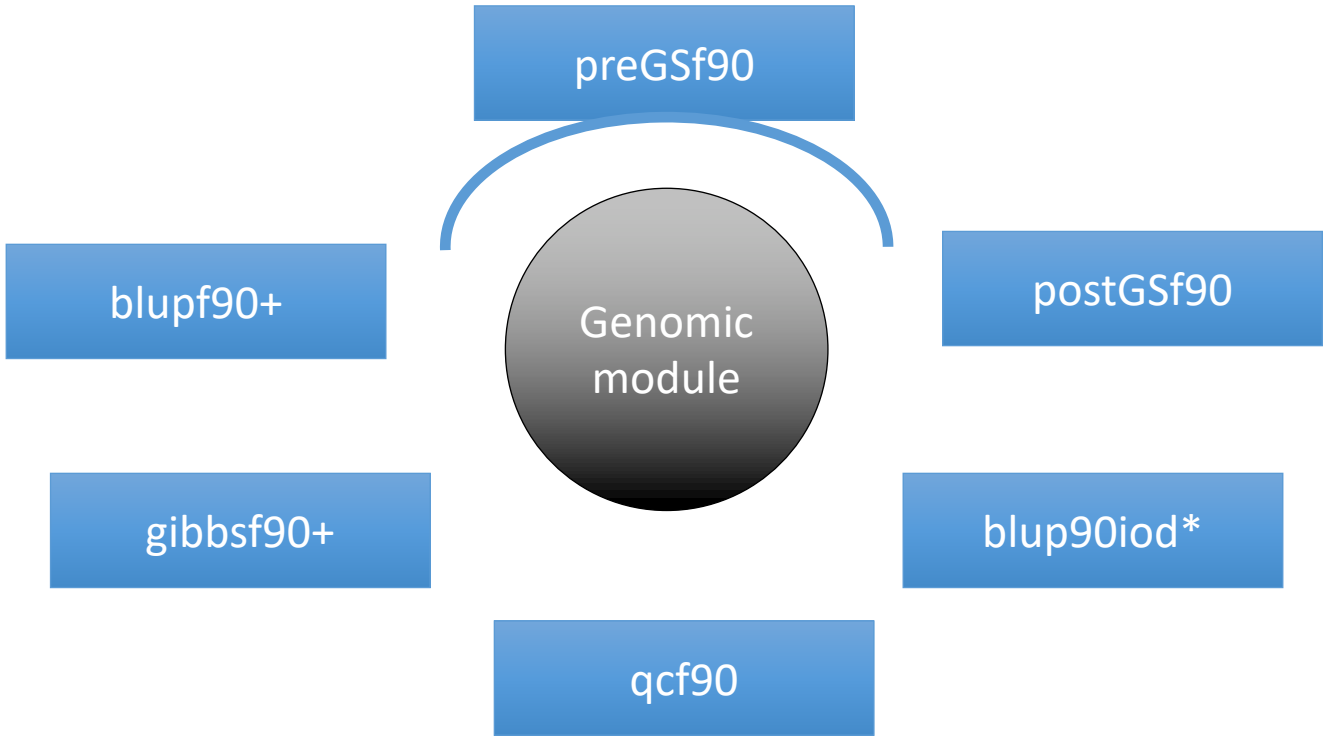
Quality control

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 (**G**)
  - and relationships based on pedigree ($\mathbf{A}_{22}$)
  - 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

SNP map file

- OPTION map_file *<file>*
  - For GWAS and QC
- Format:
  - A header must be 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

- OPTION saveCleanSNPs
- Save clean genotype data without excluded SNP and individuals
  - For example, for a SNP_file named *marker.geno*
  - Clean files 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

- preGSf90 does:
  - 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 stop_after_quality_control`

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` to renumber and create XrefID and files

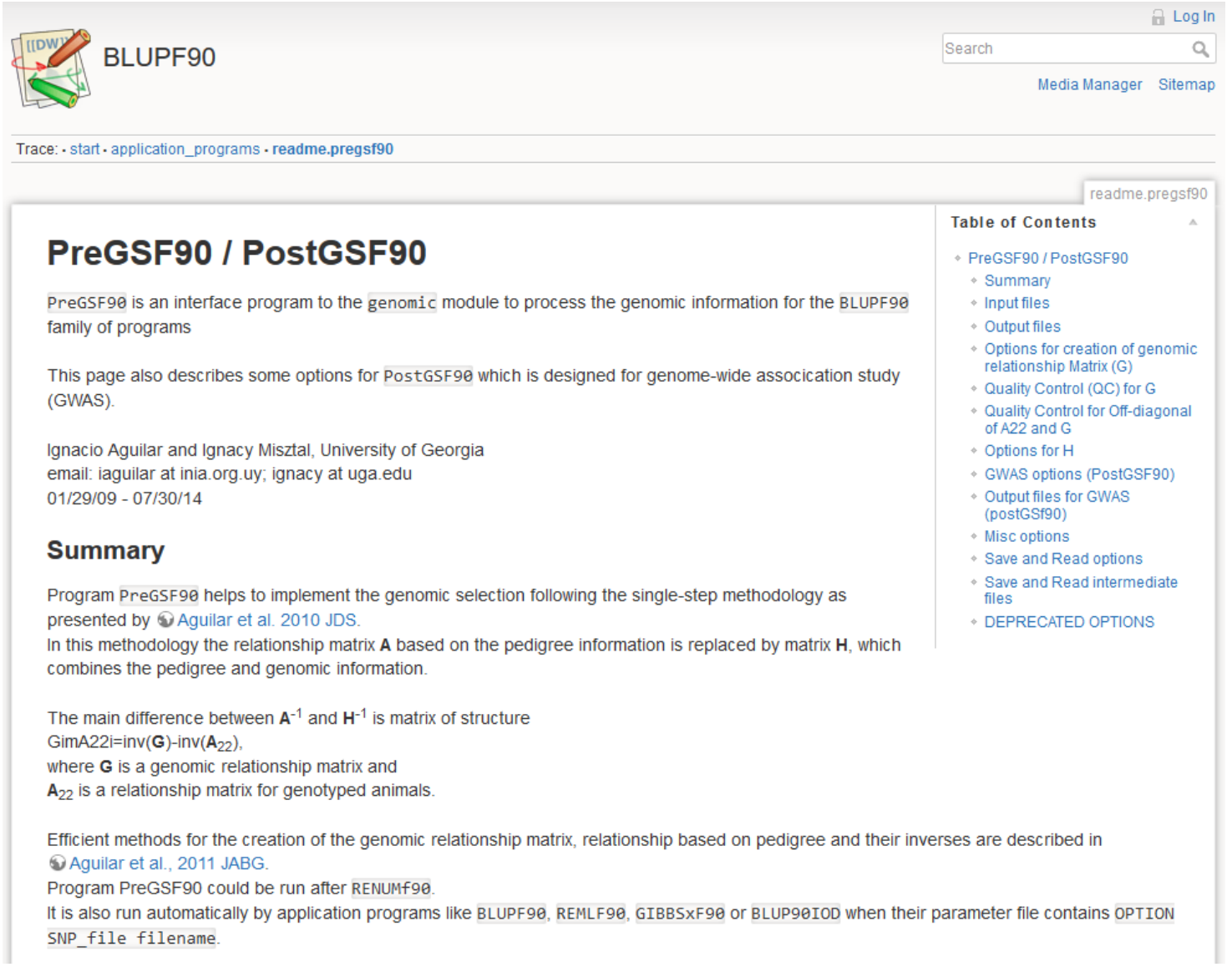
  `SNP_FILE`
  `marker.geno`

```
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



BLUPF90

Log In

Search

Media Manager Sitemap

Trace: • start • application_programs • readme.pregsf90

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

preGSf90

- Performs Quality Control of SNP information
- Creates the genomic relationship matrix (**G**)
  - and relationships based on pedigree ($\mathbf{A}_{22}$)
  - Inverse of relationship matrices

PreGSf90

- Created to construct the matrices using in ssGBLUP

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

$\mathbf{G}$ $\mathbf{G}^{-1}$

$\mathbf{A}_{22}$ $\mathbf{A}_{22}^{-1}$

$\mathbf{G}^{-1}-\mathbf{A}_{22}^{-1}$

Genomic Relationship Matrix - **G**

- $\mathbf{G} = \frac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)}$ (VanRaden, 2008)
  - Z = matrix for SNP marker
  - Dimension of Z = *n***i*
  - *n* animals
  - *i* markers

Genotype Codes
0 – Homozygous
1 – Heterozygous
2 – Homozygous
5 – No Call (Missing)

SNP file

```
80    211010110020120110110101101111112111112101 00
8014  211101015111011202211101115111121011122101 00
516   211001012022520211202101211021112022121111 01
181   211101111122011205502000201010222122111111 00
```

PreGSf90

- Efficient methods
  - create the genomic relationship matrix and the relationship matrix based on pedigree
  - Invert the relationship matrices

    - Computes statistics for the matrices
      - Means, Var, Min, Max
      - Correlations between diagonals
      - Correlations for off-diagonals
      - Correlations for the full matrices
      - Regression coefficients

Genomic Matrix default options

$$\mathbf{G_0} = \frac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)} = \frac{(\mathbf{M}-\mathbf{2P})(\mathbf{M}-\mathbf{2P})'}{2\sum p_i(1-p_i)}$$

(VanRaden, 2008)

- With:
  Z centered using current allele frequencies:
    Current genotyped animals

Genomic Matrix Options

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

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

Genomic Matrix default options

- **Blending** - to avoid singularity problems

$$\mathbf{G} = 0.95*\mathbf{G}_0 + 0.05*\mathbf{A}_{22}$$

  - OPTION AlphaBeta 0.95 0.05 #(default)
  - Beta may vary from 0.2 to 0.01

Genomic Matrix default options

- **Tuning**
  - Adjust **G** to have mean of diagonals and off-diagonals equal to $\mathbf{A}_{22}$
  - OPTION tunedG 2 #(default) Chen et al. (2011)

    - Base of GBLUP is *genotyped* animals
    - Base of pedigree is *founders of the pedigree*
    - For SSGBLUP modelled as a mean for genotyped animals
      - $p(\boldsymbol{u}_2) = N(\mathbf{1}\mu, \mathbf{G})$
      - Integrate μ : $\mathbf{G}^* = 11'\boldsymbol{\lambda} + (1 - \lambda/_2)\mathbf{G}$
      - μ = (Genomic base) – (Pedigree base)
      - Vitezica et al. 2011

Options for matching **G** to $\mathbf{A}_{22}$

- OPTION tunedG *x*
  - 0: no adjustment
  - 1: mean(diag(G))=1, mean(offdiag(G))=0
  - 2: mean(diag(G))=mean(diag(A_{22})), mean(offdiag(G))=mean(offdiag(A_{22})) (default)
  - 3: mean(G)=mean(A_{22})
  - 4: Use Fst adjustment. Powell et al. (2010) & Vitezica et al. (2011)

$$\lambda = \frac{1}{n^2}\left(\sum_i \sum_j \mathbf{A}_{22_{ij}} - \sum_i \sum_j \mathbf{G}_{ij}\right) \qquad \mathbf{G}^* = 11'\lambda + (1 - \lambda/2)\mathbf{G}$$

Storing and Reading Matrices

- preGSf90 saves $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$ by default (file: GimA22i)

To save 'raw' **G**:

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

To save $\mathbf{G}^{-1}$

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

To save $\mathbf{A}_{22}$ and inverse

- OPTION saveA22 and OPTION saveA22Inverse

Saves in binary format!!!

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

Genomic Matrix - Population structure

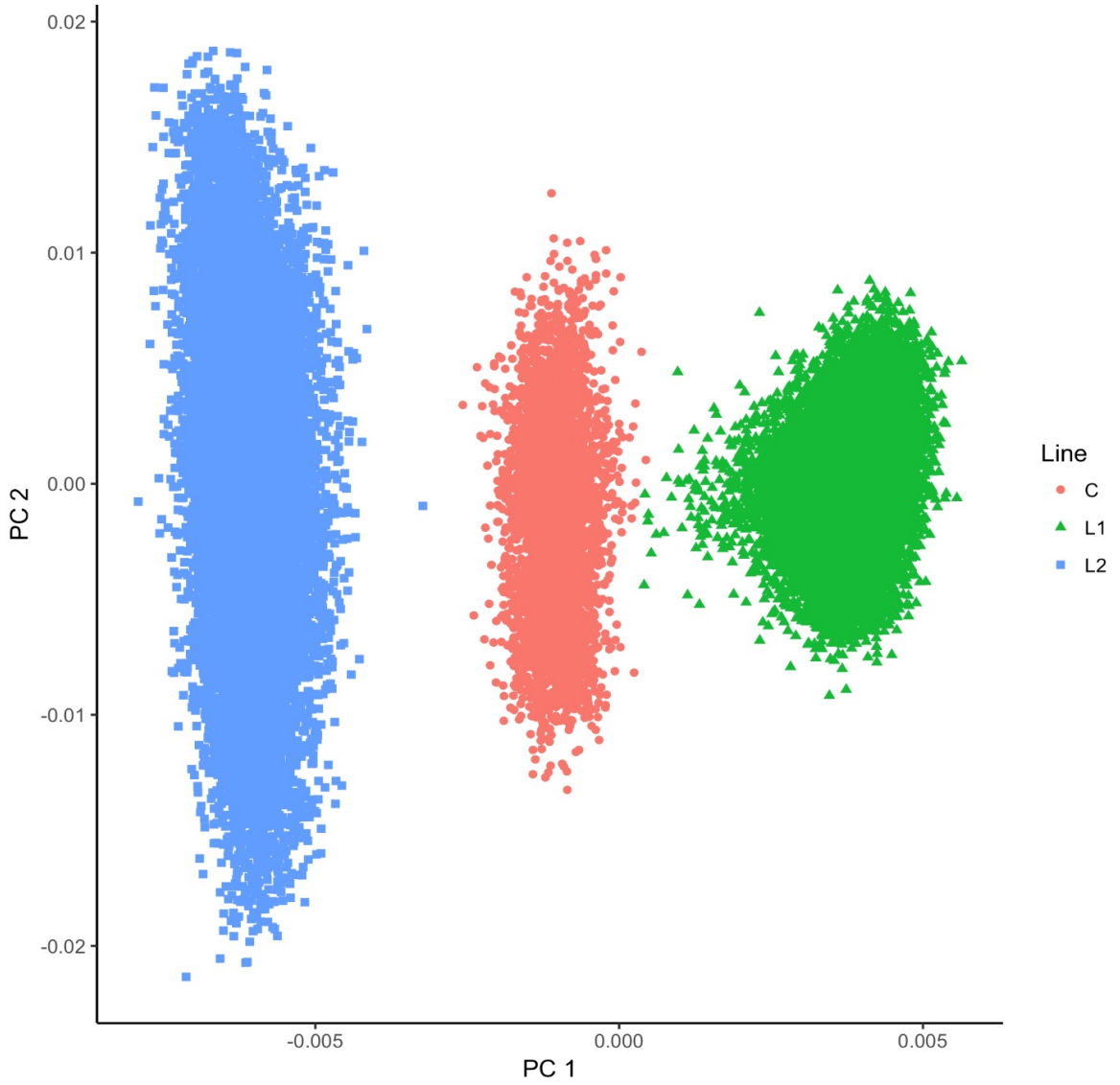
```
OPTION plotpca
```

Plot first two principal components to look for stratification in the population.

```
OPTION extra_info_pca file col
```

Reads from *file* the column *col* to plot with different colors for different classes.

Genomic Matrix - Population structure



Tricks to setup **G** for GBLUP #1

- Tricks are needed because preGSf90 is set up for ssGBLUP

1) Use a dummy pedigree

1 0 0
2 0 0
…

2) Use PED_DEPTH 1 in renumf90

3) Change blending parameters

- OPTION AlphaBeta 1.00 0.00 → G = 1.00***G** + 0.00***I**
- OPTION AlphaBeta 0.95 0.05 → G = 0.95***G** + 0.05***I**

4) No adjustment for compatibility with $\mathbf{A}_{22}$

- OPTION tunedG 0

Tricks to setup **G** for GBLUP #2

1) In renum.par, remove any information about the pedigree. Example:

```
FILE
pedigree.txt
FILE_POS
1 2 3 0 0
PED_DEPTH
3
```

3) Change blending parameters

- `OPTION AlphaBeta 1.00 0.00` → G = 1.00***G** + 0.00***I**
- `OPTION AlphaBeta 0.95 0.05` → G = 0.95***G** + 0.05***I**

4) No adjustment for compatibility with $\mathbf{A}_{22}$

- `OPTION tunedG 0`

PreGSf90 inside BLUPF90 ??

- Almost all programs from BLUPF90 support creating genomic relationship matrices
- OPTION SNP_file xxxx

- Why preGSF90 ?
  - Same genomic relationship matrix for several models, traits, etc.
    - Just do it once and store GimA22i or Gi and A22i separate

Use in application programs

- Use renumf90 for renumbering and creating of XrefID and files

```
SNP_FILE
marker.geno
```

- Option 1:

  run blupf90+ right after renumf90

- Option 2:

  run preGSf90 with quality control, saving clean files

  run blupf90+ with clean files