UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Introduction to genomics

Daniela Lourenco
UGA USA

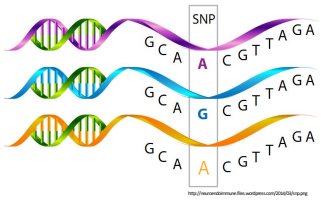
Andres Legarra
INRA France

BLUPF90 TEAM, 02/2023

1

Genomic Information


articles
Initial sequencing and analysis of the human genome
nature
the human genome



Mutation < 1% < SNP

2

What are SNP used for?

Theor Appl Genet (1983) 67:25–33

© Springer-Verlag 1983

Genetic polymorphism in varietal identification and genetic improvement*

M. Soller[1] and J. S. Beckmann[2]
[1] Department of Genetics, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel
[2] Institute of Field and Garden Crops, Agricultural Research Organization, The Volcani Center 50250 Bet Dagan, Israel

Received July 14, 1982; Accepted July 3, 1983
Communicated by A. Robertson

Summary. New sources of genetic polymorphisms promise significant additions to the number of useful genetic markers in agricultural plants and animals, and prompt this review of potential applications of polymorphic genetic markers in plant and animal breeding. Two major areas of application can be distinguished. The first is based on the utilization of genetic markers to determine genetic relationships. These applications include varietal identification, protection of breeder's rights, and parentage determination. The second area of application is based on the use of genetic markers to identify and map loci affecting quantitative traits, and to monitor these loci during introgression or selection programs. A variety of breeding applications based on

Use of DNA polymorphisms as genetic markers
- Construct genetic relationships
- Parentage determination
- Identification of QTL

RFLP
Expensive

3

Excitement about genomics

Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes† and M. E. Goddard†,‡
*Research Institute of Animal Science and Health, 8200 AB Lelystad, The Netherlands, †Victorian Institute of Animal Science, Attwood 3049, Victoria, Australia and ‡Institute of Land and Food Resources, University of Melbourne, Parkville 3052, Victoria, Australia

Manuscript received August 17, 2000
Accepted for publication January 17, 2001

- Genotyping will become cheap
  - Thousands of SNP
- Compute GEBV based on SNP
  - High accuracy
  - Animals with no phenotypes
  - Select the best animals earlier

4

4

Genotyping became cheaper in 2008

- First genomic evaluation for dairy and beef cattle in 2009
  - $300 in 2009 vs. $30 in 2022
  - 50,000 SNP

What about statistical methods able to fit genomic information?

5

5

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

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

G. K. Robinson

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

6

6

Henderson's MME

- Model

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

- 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

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

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

7

7

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-

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

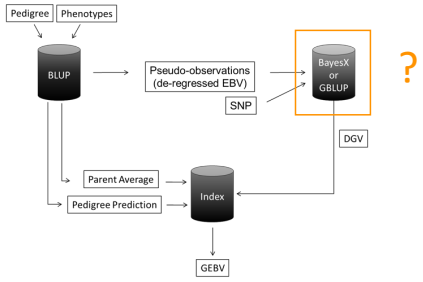
This research was conducted using the Cornell National Supercomputer Facility, a resource of the

8

8

From 1989 to 2009

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



Multistep

9

9

Bayesian Alphabet

- SNP effect models = outputs SNP effects
- BayesA (Meuwissen et al., 2001)
  - All SNPs have effect on the trait (few with large effect) $a_i \sim N(\mu, \sigma^2_{a_i})$
  - Different variances for each SNP
- BayesB (Meuwissen et al., 2001)
  - $p(a_i|\sigma^2_{a_i}, \pi) = \begin{cases} t(0, v, \sigma^2_{a_i}) or\ N(0, \sigma^2_{a_i})\ with\ probability\ (1-\pi) \\ 0\ with\ probability\ \pi \end{cases}$
- When π = 0, BayesB becomes BayesA

10

10

Bayesian Alphabet

- BayesC (Habier et al., 2011)
  - $p(a_i|\sigma^2_a) = \begin{cases} N(0, \sigma^2_a)\ with\ probability\ (1-\pi) \\ 0\ with\ probability\ \pi \end{cases}$
- BayesR (Erbe et al., 2012)
  - $p(a_i|\pi, \sigma^2_a) = \pi_1 \times N(0, 0\times\sigma^2_u) + \pi_2 \times N(0, 10^{-4}\times\sigma^2_u) + \pi_3 \times N(0, 10^{-3}\times\sigma^2_u) + \pi_4 \times N(0, 10^{-2}\times\sigma^2_u)$
- BayesRC (MacLeod et al., 2016)
  - BayesR using biological information to assign SNP to classes
- High computing cost and simple models
- After > 10 years, assumption of normality is good enough!

11

11

SNP-BLUP (ridge regression)

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

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

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{I}\frac{\sigma^2_e}{\sigma^2_a} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \qquad \mathbf{GEBV} = \mathbf{Z\hat{a}}$$

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

12

12

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!!!

13

13

Understanding SNP variance

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

How do we get the variance of SNP effects, σ_a^2 ?

1) You can estimate it (Bayes C, REML)
2) You can « guess » from the genetic variance σ_u^2

SNP 1 contributes $2p_1q_1a_1^2$ to the genetic variance
SNP 2 contributes $2p_2q_2a_2^2$ to the genetic variance
…

$$\sigma_u^2 = 2\sum p_i q_i a_i^2 \approx 2\left(\sum p_i q_i\right)\times\overline{(a_i^2)} \approx 2\left(\sum p_i q_i\right)\sigma_a^2$$

Reversing the expression gives

$$\sigma_a^2 \approx \frac{\sigma_u^2}{2(\sum p_i q_i)}$$

14

14

GBLUP: equivalent to SNP-BLUP

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

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

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

Bernardo (1994)
Nejati-Javaremi et al. (1997)

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

VanRaden (2008)

15

15

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*

16

16

What are genomic relationships?

- Relationships were conceived as standardized covariances (Fisher, Wright)
  - $Cov(u_i, u_j) = R_{ij}\sigma_u^2$
  - R_{ij} "some" relationship
  - σ_u^2 genetic variance
- True relationships: two individuals are genetically identical (for a trait) if they carry the same genotype at the causal QTL or genes
- Genomic relationships: due to shared (Identical By State) alleles at *causal genes*
  - If I share the blood group A with someone, we are like twins!
  - Most of the genes are unknown
  - We use proxies (SNP markers)

17

17

Early use of markers to infer **A**

- **A** = pedigree relationships: due to shared (Identical By Descent) alleles at *causal genes*
- In conservation genetics
- Gather markers, then reconstruct pedigrees, then construct **A**
  - Either estimates of A_{xy}, or estimates of « the most likely relation » (son-daughter, cousins, whatever)

  Li and Horvitz 1953, Cockerham 1969, Ritland 1996, Caballero & Toro 2002, and many others
- With abundant marker data we can do better than this

18

18

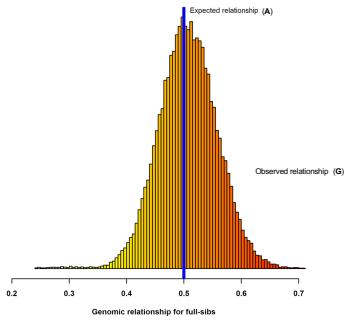
Pedigree vs. Genomic relationships

- Identical By Descent Relationships based on pedigree are average relationships which assume infinite loci
- « Real » IBD relationships are a bit different due to finite genome size (Hill and Weir, 2010)
- Therefore **A** is the expectation of 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
  - Estimators of IBD

19

19

Pedigree vs. Genomic relationships



Adapted from Lourenco et al. (2015)

20

20

Genomic relationships

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

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*

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

21

21

Some "interesting" properties of **G**

- If *p* are computed from the data
  This implies that E(Breeding Values)=0
- Positive and negative inbreeding
  Some individuals are more heterozygous than the average of the population
  (OK, no biological problem)
- Positive and negative genomic relationships
  Individuals *i* and *j* are more distinct than an average pair of individuals in the data
  Fixing negative estimates of relationships to 0 is a wrong praxis

22

22

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
- Stranden and Christensen (2011)
  - **G** is singular if *p's* are averages across the sample

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

- Blending ≈ Adding a residual polygenic effect

23

23

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"

24

24

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

$$\mathbf{A}=\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix} \qquad \mathbf{A}=\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{G}\end{bmatrix} \qquad \mathbf{H}=\mathbf{A}+\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}-\mathbf{A}_{22}\end{bmatrix}$$

Genotyped

Misztal et al., 2009

25

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 prediction of 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

26

Understanding **H**

- It is a projection of **G** matrix on the rest of individuals "so that" **G** matrix 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

27

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), this gives lots of numerical problems
- If no one is genotyped, Single-step is BLUP
- If everyone is genotyped, Single-step is GBLUP

28

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**)

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

29

29

Single-step Genomic BLUP (ssGBLUP)

- Because not all animals are genotyped
  - 5% to 10% in large populations

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

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

Aguilar et al., 2010
Christensen and Lund, 2010

30

30

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

31

31

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 on the same scale

32

32

Single-step

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

ssGBLUP

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

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

ssSNPBLUP or **ssBR**

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

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,*† M. Koivula,‡ J. Jensen,* I. Stranden,‡ 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
†Viking Genetics, Viking Genetic Evaluation, DK-6200 Aabenraa, Denmark
‡Natural Resources Institute Finland (Luke), FIN-71500 Jokioinen, Finland

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 $\mathbf{g}$ is partitioned as:

$$\mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{T}_{12}\mathbf{g}_1 \end{bmatrix},$$

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

33

33

Quality Control of SNP data and creating of genomic matrices with preGSf90

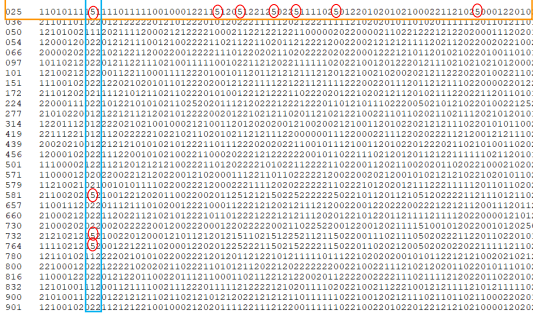
BLUPF90 TEAM, 11/2022

34

SNP data

SNP

ANIMAL



35

Quality control

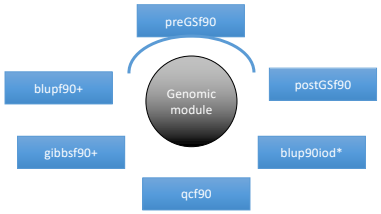
- Call rate
  - Animals
  - SNP
- Minor Allele Frequency (MAF)
- Hardy-Weinberg Equilibrium (HWE)
- Non-mapped SNP
- Mendelian Conflicts
- Duplicate genotypes
- Linkage disequilibrium (LD)

Which software in the BLUPF90 family?

36

preGSf90

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



37

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

38

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

39

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
```

40

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
```

41

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**

42

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   2110101100201201101101011011111
8014  2111010151110112022111011151111
 516  2110010120225202112021012110 21
 181  2111011111220112055020002010 10
```

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

43

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

44

SNP map file – new default

- OPTION chrinfo *<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       NUMG
31428 14 7926189 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
23666 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 79052023 ARS-BFGL-BAC-11000 15
24732 10 80410977 ARS-BFGL-BAC-11003 16
24741 10 80765719 ARS-BFGL-BAC-11007 17
24827 10 84516867 ARS-BFGL-BAC-11025 18
25065 11 21276136 ARS-BFGL-BAC-11039 21
```

45

Saving 'clean' files

- SNP excluded from QC are set to missing (i.e., Code=5)
  - 5 is replaced by 0 in calculations
- OPTION saveCleanSNPs
- Save clean genotype data without excluded SNP and individuals
  - For example, for a SNP_file named *marker.geno*
  - Clean fles 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

46

Only QC in preGSf90

- 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 createGInverse 0`
  - `OPTION createA22Inverse 0`
  - `OPTION createGimA22i 0`

47

No QC in the 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`

48

Use in application programs

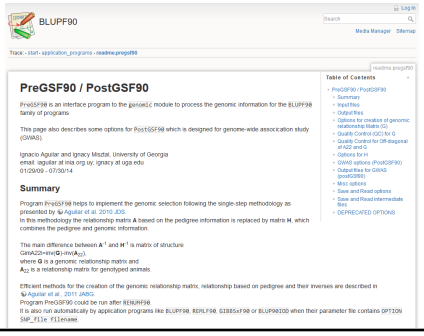
- Use `renumf90` for renumbering and creation of XrefID and files

`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



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

BLUP-based models

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$ BLUP — Henderson, 1963

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$ GBLUP — Nejati-Javaremi et al., 1997; Fernando, 1998; VanRaden, 2008

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{H}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$ ssGBLUP — Misztal et al. (2009); Legarra et al. (2009); Aguilar et al. (2010); Christensen & Lund (2010)

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} \qquad \mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

52

PreGSf90

- Created to construct the matrices using in ssGBLUP

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

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

$\mathbf{A_{22}}$ $\quad$ $\mathbf{A_{22}^{-1}}$

$\mathbf{G^{-1}-A_{22}^{-1}}$

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   21101011002012011011010110111111211111210100
8014 21110101511101120221110111511112101112210100
516  21100101202252021120210121102111202212111101
181  21110111112201120550200020101022212211111100
```


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

- $\mathbf{G}_0 = \frac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)}$ (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

57

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

58

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 : $\mathbf{G}^* = 11'\lambda + (1 - \lambda/_2)\mathbf{G}$
      - μ = (Genomic base) – (Pedigree base)
      - Vitezica et al. 2011

59

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($\mathbf{A}_{22}$)), mean(offdiag(G))=mean(offdiag($\mathbf{A}_{22}$)) (default)
  - 3: mean(G)=mean($\mathbf{A}_{22}$)
  - 4: Use Fst adjustment. Powell et al. (2010) & Vitezica et al. (2011)

$$\lambda = \frac{1}{n^2}\left(\sum_i\sum_j \mathbf{A}_{22_{ij}} - \sum_i\sum_j \mathbf{G}_{ij}\right) \qquad \mathbf{G}^* = 11'\lambda + (1 - \lambda/_2)\mathbf{G}$$

60

Storing and Reading Matrices

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

To save 'raw' genomic matrix:

- 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

61

Storing and Reading Matrices

- OPTION saveG [all] , OPTION saveGInverse, …
  - Saves in binary format
  - "Dumped" format to save space and time
  - To save as row, column, value:
    - OPTION no_full_binary
    - Still binary, but can be easily read and converted to text

62

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

63

Genomic Matrix - Population structure

```
OPTION plotpca
```

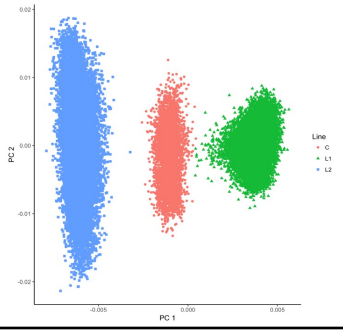
Plot first two principal components to look for stratification in the population.

```
OPTION extra_info_pca file col
```

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

64

Genomic Matrix - Population structure



65

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

66

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`

67

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

68

Use in application programs

- Use renumf90 for renumbering and creation of XrefID and files

```
SNP_FILE
marker.geno
```

- Run preGSf90 with quality control, saving clean files
- Option 1:
  run blupf90 with clean files
- Option 2:
  run preGSf90 with clean files (program saves **GimA22i**)
  run blupf90 with option to read **GimA22i** from the file

Reading external matrices

- BLUPF90 programs accept external matrices created outside
- http://nce.ads.uga.edu/wiki/doku.php?id=user_defined_files_for_covariances_of_random_effects
- File should be row, column, value in plain text format (lower OR upper triangular)

renf90.par

```
RANDOM_GROUP
 # genomic
 2
RANDOM_TYPE
user_file
FILE
 # matrix file
G1
```

Valid format

```
1 1 1
1 2 0.5
2 2 1
```

Non-valid format

```
1 1 1
1 2 0.5
2 1 0.5
2 2 1
```

- user_file: if providing the inverse of the covariance structure
- user_file_inv: if the program has to invert the covariance structure