



UNIVERSITY OF
GEORGIA

College of Agricultural &
Environmental Sciences

*Animal Breeding and
Genetics Group*

Introduction to Genomics

Daniela Lourenco

BLUPF90 TEAM – 08/2024

INIAV Course 2024

*Daniela Lourenco
Fernando Bussiman*

Genomic Information

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century¹⁻³ sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless effort to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, two animals and one plant.

Here we report the results of a collaboration involving 20 countries from the United States, the United Kingdom, Japan, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a map covering more than 96% of the euchromatic part of the genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. Sequence data have been made available without restriction, updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the bringing of the vast majority of the sequence to this standard is straightforward and should proceed rapidly.

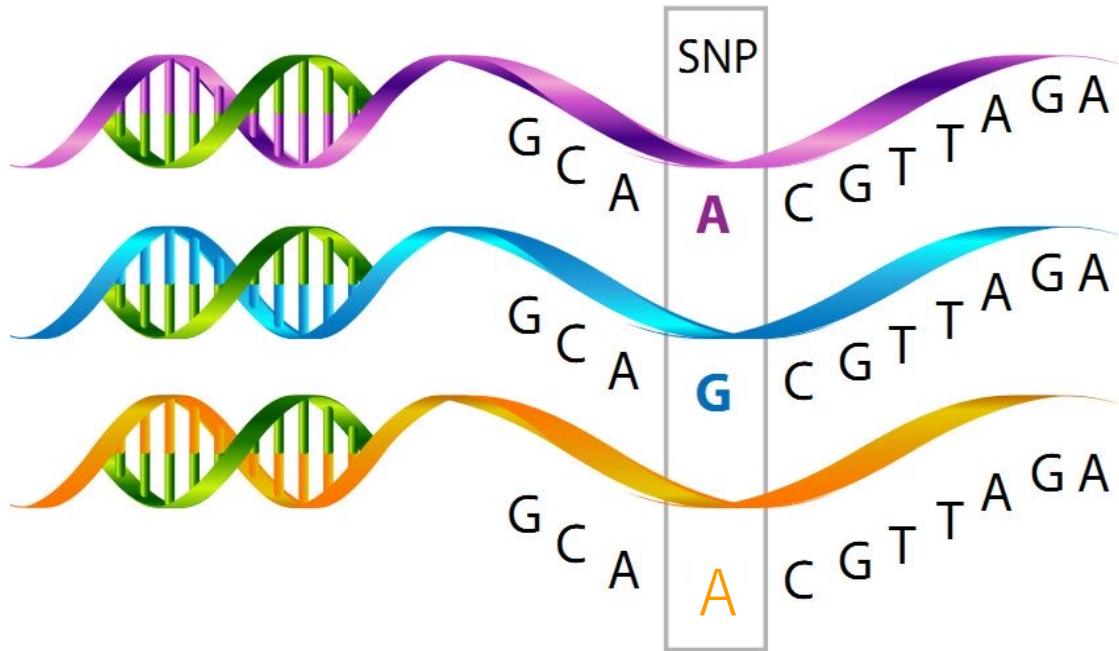
The sequence of the human genome is of interest in many respects. It is the largest genome to be extensively sequenced, being 25 times as large as any previously sequenced genome and eight times as large as the sum of all such genomes. It is the first vertebrate genome to be extensively sequenced. And, uniquely, it is the genome of our own species.

Much work remains to be done to produce a complete sequence, but the vast trove of information that has become available through this collaborative effort allows a global perspective on the human genome. Although the details will change as the sequence is finished, many points are already clear.

The genomic landscape shows marked variation in the distribution of a number of features, including genes, transposable elements, GC content, CpG islands and recombination rates. These features give us important clues about function. For example, the developmentally important HOX gene clusters are the most repetitive regions of the human genome, probably reflecting the very c

coordinate regulation of the genes in the clusters.

- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from



<http://neuroendocrine.files.wordpress.com/2014/03/snp.png>

Mutation < 1% < SNP

What are SNP used for?

Theor Appl Genet (1983) 67:25–33



Genetic polymorphism in varietal identification and genetic improvement *

M. Soller¹ and J. S. Beckmann²

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



Excitement about genomics

Copyright © 2001 by the Genetics Society of America

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

Genotyping became cheaper in 2008

- First genomic evaluation for dairy and beef cattle in 2009
 - \$300 in 2009 vs. \$25 in 2024
 - 50,000 SNP

What about statistical methods able to fit genomic information?

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

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

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

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

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



Henderson's MME for dairy in 1989

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



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

National genetic improvement programs for dairy cattle in the United States

G. R. Wiggans

J Anim Sci 1991. 69:3853-3860.

Challenges

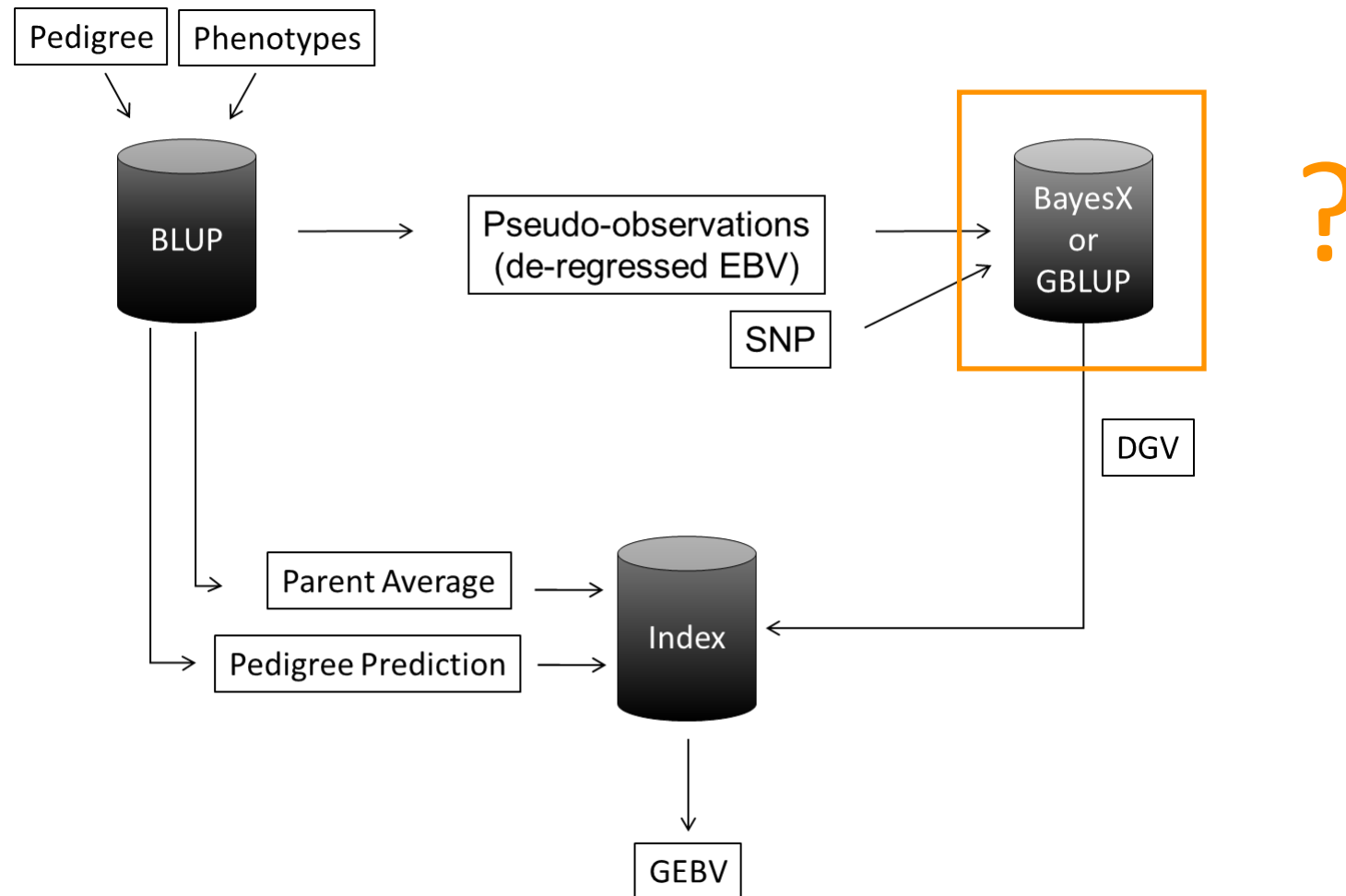
Genetic improvement programs are in a period of rapid change. Advances in computer capability enable adoption of sophisticated computational procedures. Advances in repro-

ACKNOWLEDGMENTS

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

From 1989 to 2009

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



Multistep

Bayesian Alphabet

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

Bayesian Alphabet

- BayesC (Habier et al., 2011)

- $p(a_i | \sigma_a^2) = \begin{cases} N(0, \sigma_a^2) \text{ with probability } (1 - \pi) \\ 0 \text{ with probability } \pi \end{cases}$

- BayesR (Erbe et al., 2012)

- $p(a_i | \pi, \sigma_a^2) = \pi_1 \times N(0, 0 \times \sigma_u^2) + \pi_2 \times N(0, 10^{-4} \times \sigma_u^2) + \pi_3 \times N(0, 10^{-3} \times \sigma_u^2) + \pi_4 \times N(0, 10^{-2} \times \sigma_u^2)$

- BayesRC (MacLeod et al., 2016)

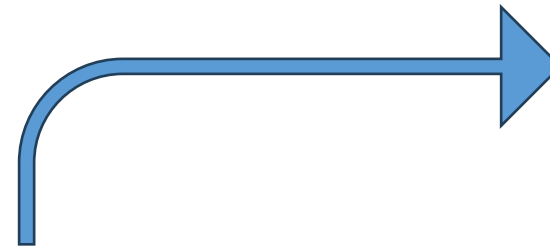
- BayesR using biological information to assign SNP to classes

- High computing cost and simple models

- After > 10 years, assumption of normality is good enough!

SNP-BLUP (ridge regression)

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



$$y = X\beta + Za + e$$

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

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

```

1101011115111101111001000122115120512212502251110250122010201021000221121025000122010:
2110110102201212222012101222010120222111112021222111112102020101101020111112011012110:
121010021112021112000212122210002112212212211000002022000021102212221212202000112020:
12000120022012121110012100222211021122110201121222120022002121212111202112022002022100:
2000020202210212211200220012222111012202021102022202022000122212101201021022010011010:
1011021202201211221110210011110010221121202211111020221001201222012111021021021012000:
121002120220012211000111122201001012011212111212012211020202021211222022010022110:
111001020221220210201011012220200121221112212211211122200220111201121211022000022012:
211012020211112101211021102220100102212122211022202012210202121120102111220221120110:
220001110221012210102110252020111212022212221222011012101110222005021012022010022125:
210102200121221211212021012222002012210212110201121021221002211011020211021112021012010:
1220111201222021021001000212100112012020200121002002121001120102202121211022010101100:
22111221012112022221022102110201021121211122000001112200022112202022211212001212110:
2002010012212101010210122211011122202020221100101112100112010220122202110210011020:
1200010202211122001010210022110002022212122222001011022111021201201121221111102112010:
1110002122112120121212100222110120222101022112222110220011202110020201102022100021020:
1100001202022002212120220012102000111221101102222120022002012001010212121022102010110:
112100210210010101111022002221200022211120202222211022210120201211122211112011011020:
21100202152100122120201100220020112512121502252222250221011201121051202222112111012110:
110011120220111211101020012221000112221212021211121200220012202220022212212112001112011:
210002120221120221121021012221011012221222121211202012210122011211121112022000012101:
21000202022002022222001200222000122022220021102252200122001202111151001012022001012025:
212102121521002201200012101121201215110215122521121150220011102111050202221122011022010:
11110212152001221221102000122020122522211502152222115022011020212005020202202211112110:
1211010211222021010102200222120120112122101211110111221020202001010112212121002021021:
2210001202212222102020211022211010121102212022222002210022112121021202011022010111010:
1100012202201212201100220111211000110211221212122002011222200222211021111212022011022010:
12101001112001211110021112220111112122221210201110202210021122210012121112101211110:
21010011022012212110211021210121202212121211011111022100120212110211011021100022020:
12100102022112121221001000212020111122111212200111110221002201022012212121021000012020:

```

- 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

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

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

$$\text{Var}(\mathbf{u}) = \text{Var}(\mathbf{Z}\mathbf{a})$$

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

$$\text{Var}(\mathbf{u}) = \mathbf{Z}\mathbf{Z}'\sigma_a^2$$

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

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

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

Genomic
relationship matrix
VanRaden (2008)

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

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



GBLUP assumption!!!

GBLUP: equivalent to SNP-BLUP

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

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

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

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

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

VanRaden (2008)

Genomic relationship matrix

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

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

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 additive 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)

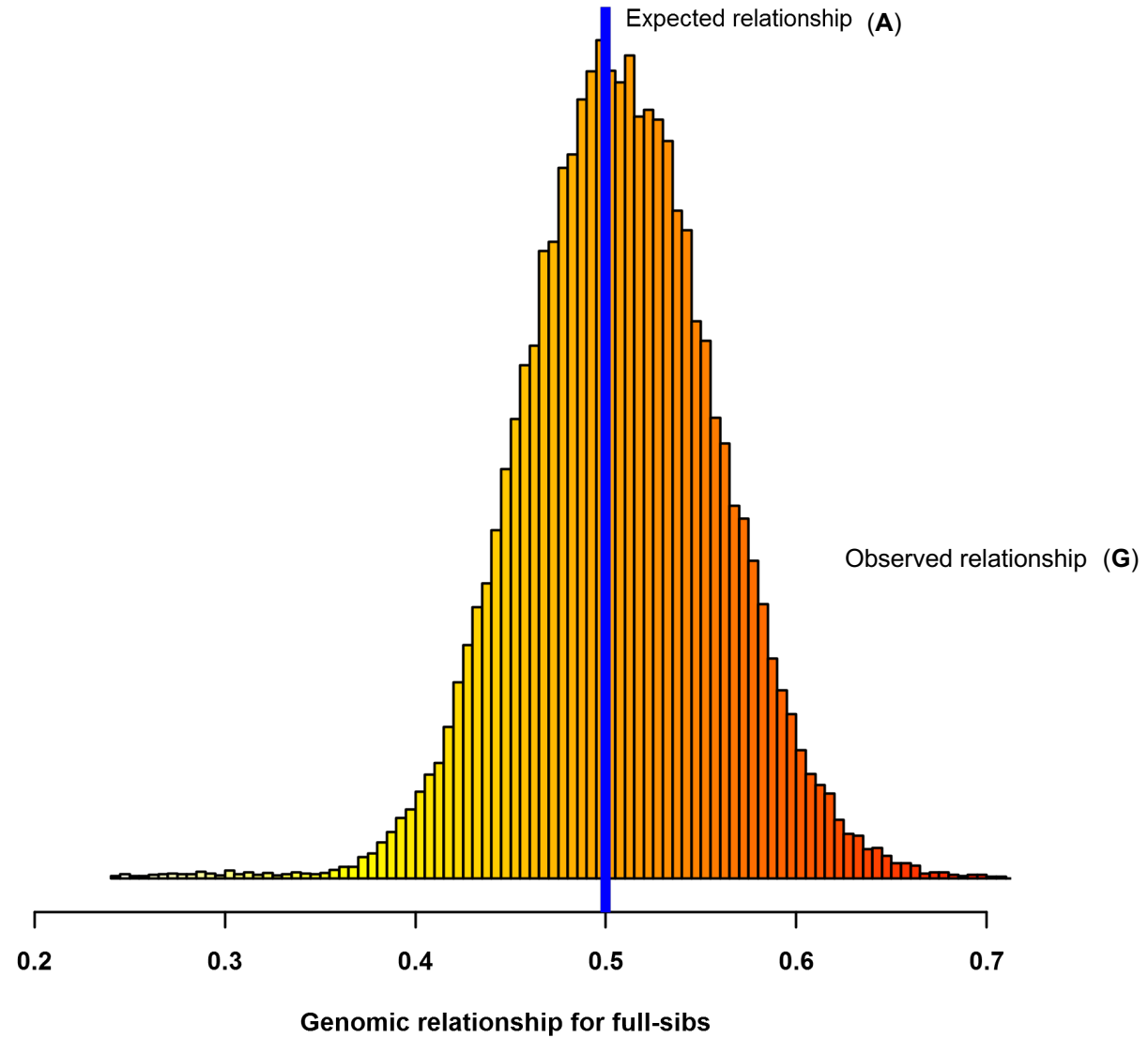
Early use of markers to infer **A**

- **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

Pedigree vs. Genomic relationships

- Identical By Descent Relationships (IBD) based on pedigree
 - 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

Pedigree vs. Genomic relationships



Genomic relationships

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

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, \mathbf{G} is an unbiased and efficient estimator of IBD realized relationships

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 \approx Adding a residual polygenic effect

Some “interesting” properties of **G**

- For all matrices of the kind

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2 \sum p_i(1 - p_i)} = \frac{(\mathbf{M} - 2\mathbf{P})(\mathbf{M} - 2\mathbf{P})'}{2 \sum p_i(1 - p_i)}$$
 - We don't need to put the same p 's in the upper and and in the lower part
- Changing allele frequencies in \mathbf{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”

GBLUP

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

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

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

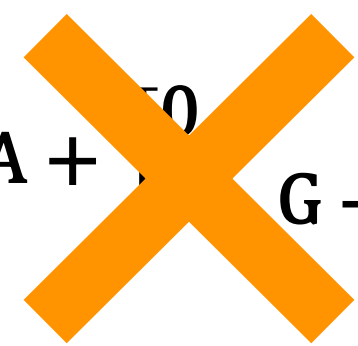
Only for
genotyped individuals!!!

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

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!

$$\begin{array}{c}
 \text{Non-genotyped} \rightarrow \\
 \mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \\
 \leftarrow \text{Genotyped}
 \end{array}
 \qquad
 \mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G} \end{bmatrix}
 \qquad
 \mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix}$$


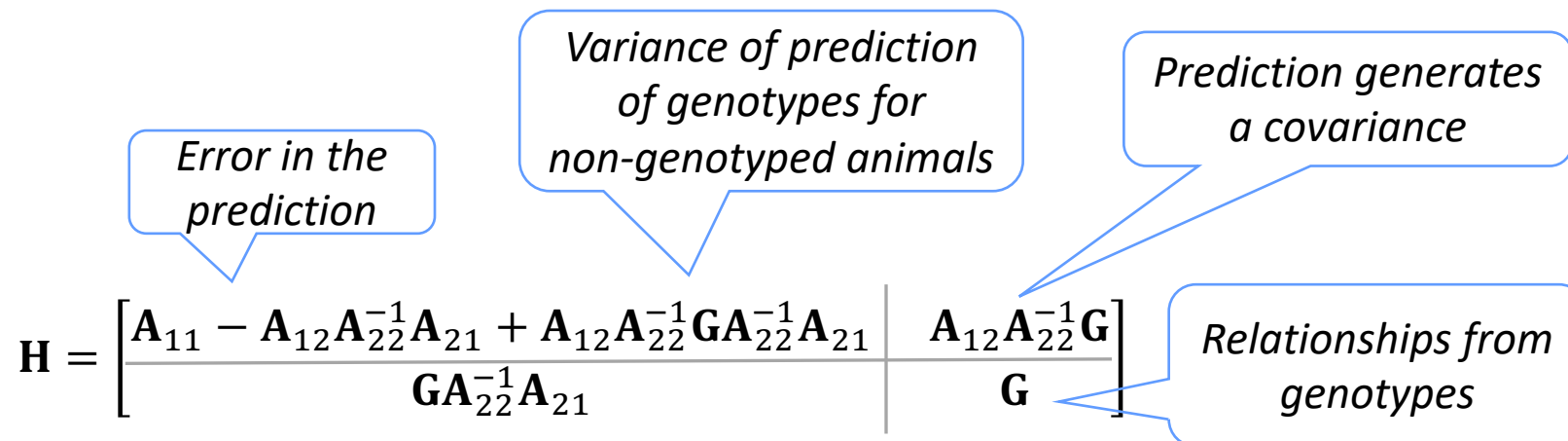
Not all animals are genotyped

- 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

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



The diagram shows the H matrix partitioned into two columns. The first column contains the variance-covariance matrix for non-genotyped animals, and the second column contains the relationships from genotypes. Callouts explain the components:

- Error in the prediction:** Points to the term $\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ in the top-left element of the first column.
- Variance of prediction of genotypes for non-genotyped animals:** Points to the term $\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ in the top-right element of the first column.
- Prediction generates a covariance:** Points to the term $\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}$ in the top-right element of the second column.
- Relationships from genotypes:** Points to the term \mathbf{G} in the bottom-right element of the second column.

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

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 **A**₂₂ in the thousands
 - Leads to a very efficient method of genomic evaluation:
 - Single Step GBLUP

Some properties of \mathbf{H}

- Always semi-positive definite
 - eigenvalues are always positive or zero
- Positive definite & invertible if \mathbf{G} is invertible
- In practice, if \mathbf{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**)

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

Single-step Genomic BLUP (ssGBLUP)

- Because not all animals are genotyped
 - 5% to 15% 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

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

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

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

$$\mu = (\text{Pedigree base}) - (\text{Genomic base})$$

Tries to put \mathbf{G} and \mathbf{A} in the same scale

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)

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

ssSNPBLUP or ssBR

Fernando et al. (2014)
Liu et al. (2014)
Mantysaari & Strandén (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 \mathbf{g} is partitioned as:

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

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



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

© 2018, The Authors. Published by FASS Inc. and Elsevier Inc. on behalf of the American Dairy Science Association®.
This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

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

H. Gao,*† 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

Bases for Genomic Predictions

Bases for Genomic Prediction

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

2022-05-11





UNIVERSITY OF
GEORGIA
College of Agricultural &
Environmental Sciences

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

SNP data

SNP

ANIMAL

025	110101110511110111110010001221151205122125022511110250122010201021000221121025000122010
036	21101101022012122222012101222010120222111112021222111112102020101101020111112011012110:
050	121010021112021111200021212222100021122122122110000020220000211022122212122020001112020:
054	120001200220121211100121002222110211221102011212221200220021212121111202112022002022100:
066	200002020221021221120022001222211101220202110202222020220001222121011201021022010011010:
097	101102120220121122111021001111100102211212022111111020221001201222012111021021021012000:
101	121002120220011221100011112220100101120112121211121201221002102002021211222022010022110:
151	111001020221220210201011012220200121221111221221121111222002201112011212111022000022012:
172	211012020211112101211021102220101001221212221102220201221020212112010211122022112011010:
224	220001110221012210101021102520201112120222122212220110121011102220050210121022010022125:
277	21010220012122121121202101222002012210212110201121021221002211011020211021112021012010:
314	122011120122220210210010002121001120120202001210020021210011201022021212111022010101100:
419	22111221012112022222102210211020102112121112200000011220002211122020222112120012121110:
439	200202100122121210101021012221101112220202022110010111210011201022012220211021010011020:
456	1200010202211122001010210022110002022212122222001011022111021201201121221111102112010:
501	111000021221121201212121002221101202222101022112222110220011202110020201102022100021020:
571	110000120202200221212022001210200011122110110222221200220020212001010212121022102010110:
579	1121002102100101011110220022212000222111120202222211022210120201211122211112011011020:
581	211002025100122120201100220020112512121502252222250221011201121051202222112111012110:
657	110011120220111211101020012221000112221212021211121200220012202220022212212112001112011:
660	210002120221120221121021012221011012221222121211120201221012201121111211112022000012101:
730	210002020220020222220012002220001220222220021102252200122001202111151001012022001012025:
732	2121021251002201200012101121201215110215122521211150220011102111050202221122011022010:
764	11110212520012212211020001220201225222115021522221150220110202120050202022022111112110:
780	12110102112220210101022002221201201121221012111110111221020202001010112212121002021021:
800	22100012022122221020202110222110101211202212022222200221002211121021202011022010111010:
816	11000122022012122011002201112110001102112212122002011222200222111021111212022011022010:
832	121010011120011211110021112220111112122221210201111020221002112221001212111121012111110:
900	210100110220122121211021102121012120221212121101111110221001202121110211011021100022020:
901	121001020221121212210010002120201111221112122001111110221002201022012212121021000012020:

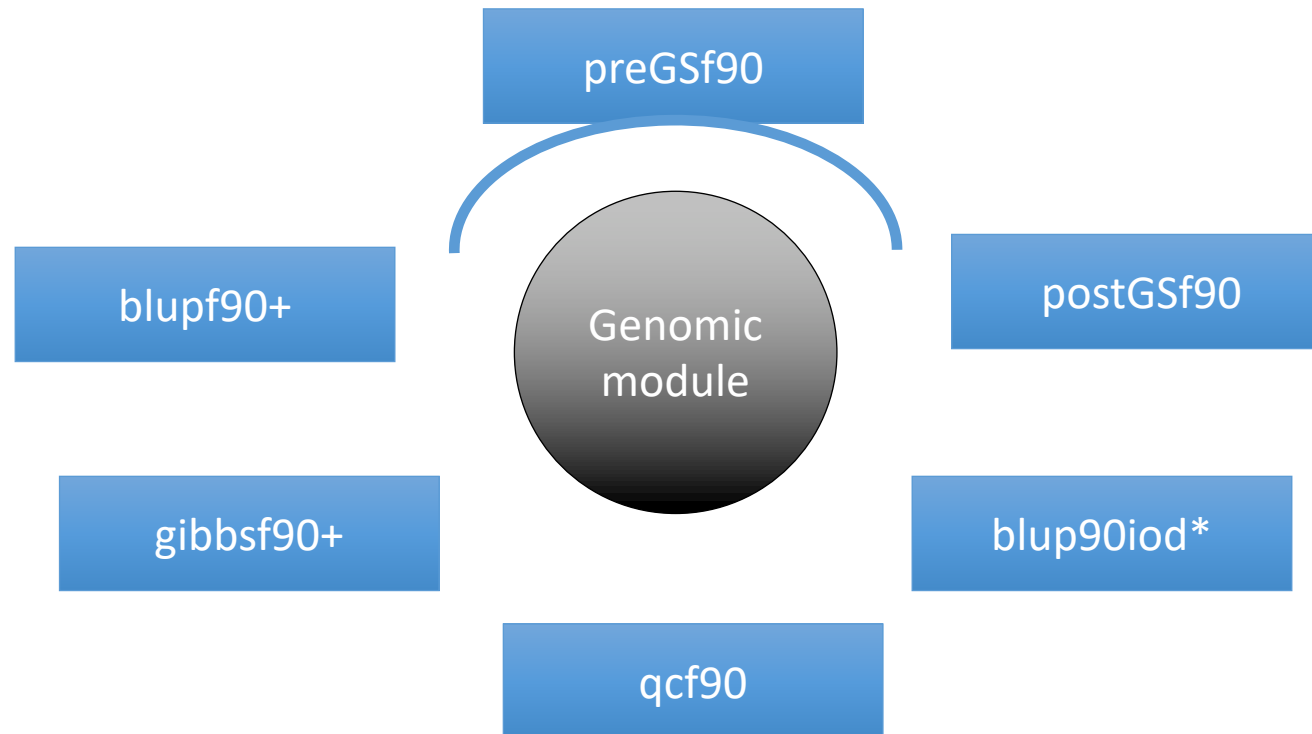
Quality control

- 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?**

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 (**A₂₂**)
 - 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 211010110020120110110101101111
8014 211101015111011202211101115111
516 211001012022520211202101211021
181 211101111122011205502000201010
```

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

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

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      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

- 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 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

- 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`

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

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

 BLUPF90

[Log In](#)

[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@inia.org.uy; ignacy@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
$$\text{GimA22} = \text{inv}(\mathbf{G}) - \text{inv}(\mathbf{A}_{22}),$$
where **G** is a genomic relationship matrix and **A₂₂** 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 / PostGSF90](#)
- ◊ [Summary](#)
- ◊ [Input files](#)
- ◊ [Output files](#)
- ◊ [Options for creation of genomic relationship Matrix \(G\)](#)
- ◊ [Quality Control \(QC\) for G](#)
- ◊ [Quality Control for Off-diagonal of A22 and G](#)
- ◊ [Options for H](#)
- ◊ [GWAS options \(PostGSF90\)](#)
- ◊ [Output files for GWAS \(postGSF90\)](#)
- ◊ [Misc options](#)
- ◊ [Save and Read options](#)
- ◊ [Save and Read intermediate files](#)
- ◊ [DEPRECATED OPTIONS](#)

preGSf90

- Performs Quality Control of SNP information
- Creates the genomic relationship matrix (**G**)
 - and relationships based on pedigree (**A₂₂**)
 - Inverse of relationship matrices



BLUP-based models

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

BLUP

Henderson, 1963

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

GBLUP

Nejati-Javaremi et al., 1997
Fernando, 1998
VanRaden, 2008

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

ssGBLUP

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

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

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

PreGSf90

- Created to construct the matrices using in ssGBLUP

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

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

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

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

Genomic Relationship Matrix - G


- $\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2 \sum p_i(1-p_i)}$ (VanRaden, 2008)

- Z = matrix for SNP marker
- Dimension of Z = $n \times i$
- 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

- $\mathbf{G}_0 = \frac{\mathbf{Z}\mathbf{Z}'}{2 \sum p_j(1-p_j)}$ (VanRaden, 2008)
- With:
 - \mathbf{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 \mathbf{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
 - $p(\mathbf{u}_2) = N(\mathbf{1}\mu, \mathbf{G})$
 - Integrate $\mu : \mathbf{G}^* = \mathbf{1}\mathbf{1}'\lambda + (1 - \lambda/2)\mathbf{G}$
 - $\mu = (\text{Genomic base}) - (\text{Pedigree base})$
 - Vitezica et al. 2011

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

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

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

Storing and Reading Matrices

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

To save 'raw' genomic matrix:

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

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

- OPTION saveGInverse
 - Only the final \mathbf{G} , after blending, scaling, etc. is inverted !!!

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

- OPTION saveA22 and OPTION saveA22Inverse

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

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

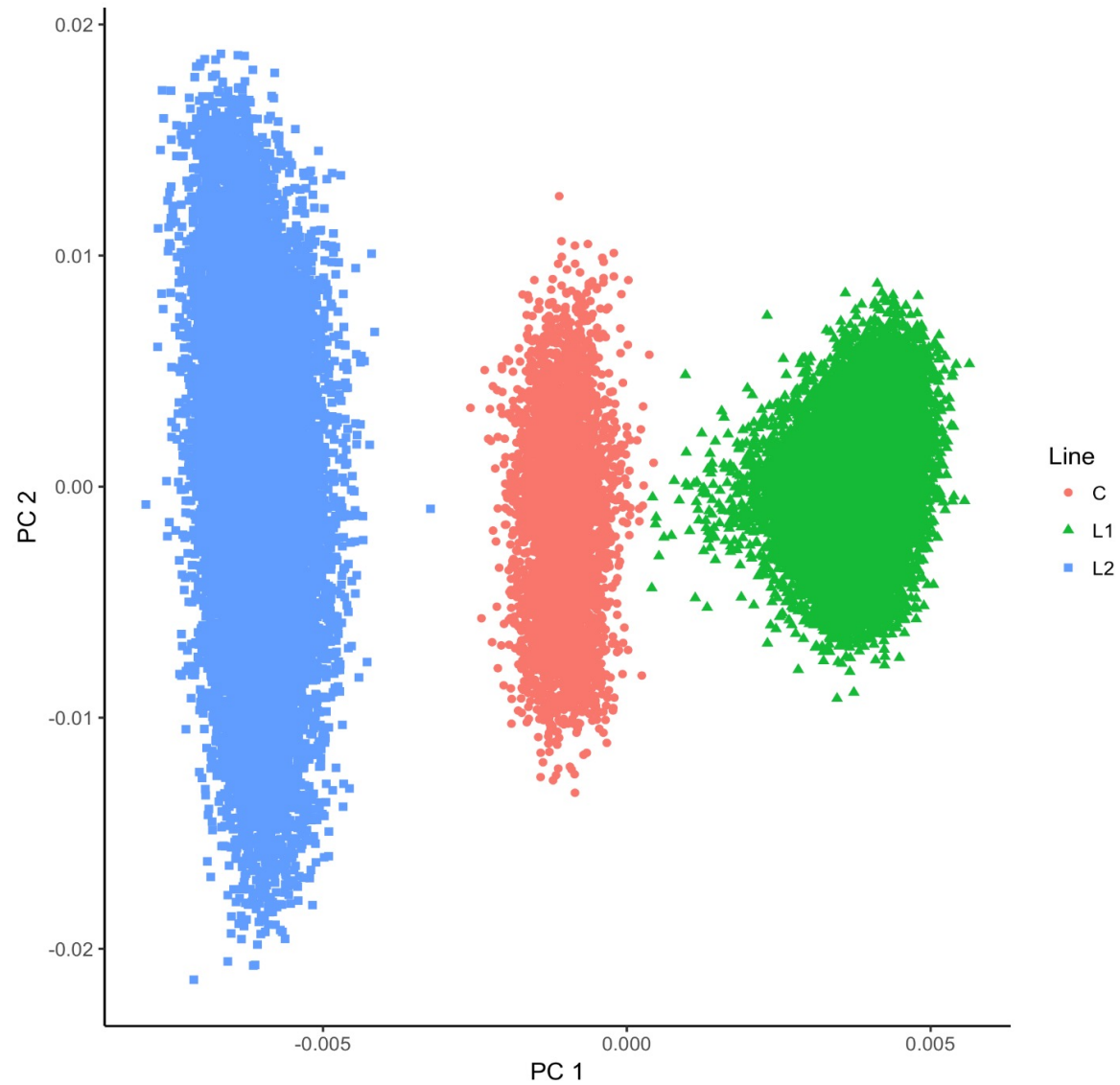
```
OPTION plotpca
```

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 $\rightarrow G = 1.00*\mathbf{G} + 0.00*\mathbf{I}$
- OPTION AlphaBeta 0.95 0.05 $\rightarrow G = 0.95*\mathbf{G} + 0.05*\mathbf{I}$

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 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 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
Gi
```

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