UNIVERSITY OF
GEORGIA
College of Agricultural &
Environmental Sciences

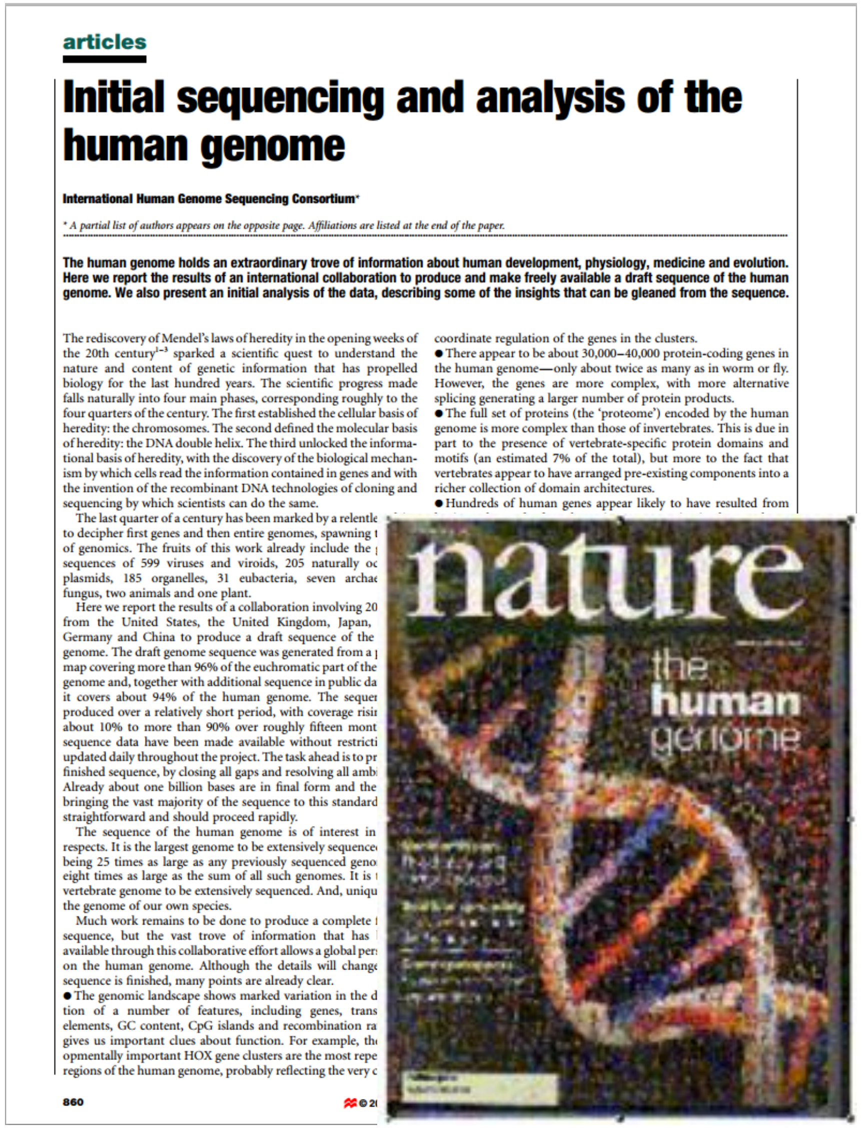
Introduction to genomics

Daniela Lourenco
UGA USA

Andres Legarra
INRA France

BLUPF90 TEAM, 02/2022

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 century[1–3] 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 relentle
to decipher first genes and then entire genomes, spawning t
of genomics. The fruits of this work already include the g
sequences of 599 viruses and viroids, 205 naturally oc
plasmids, 185 organelles, 31 eubacteria, seven archae
fungus, two animals and one plant.

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

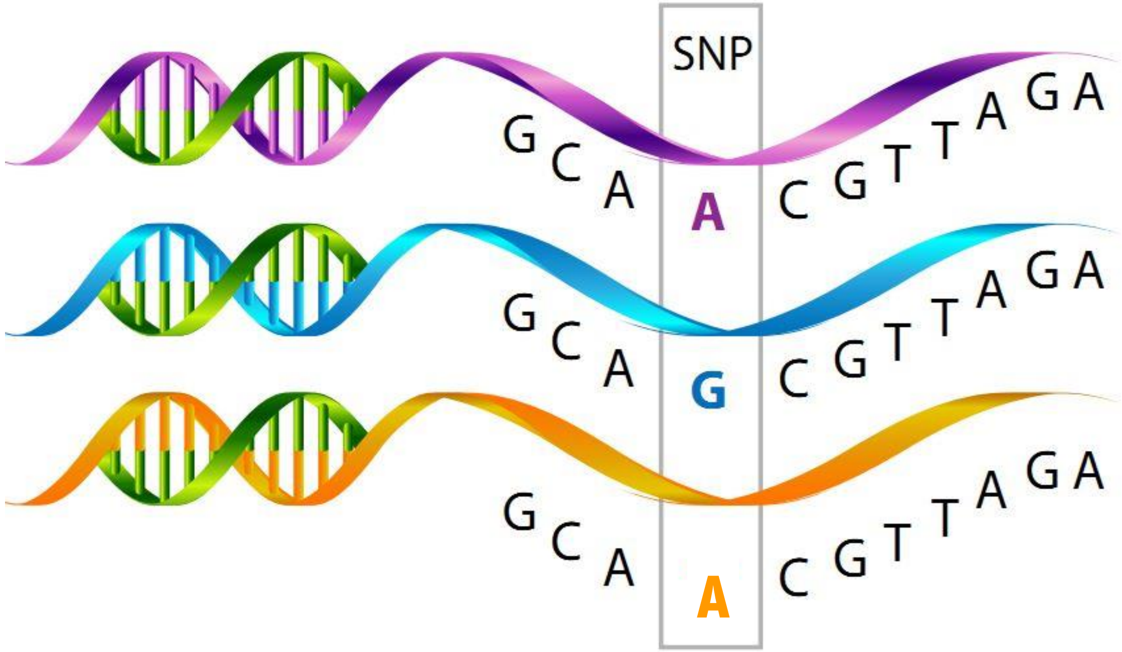
The sequence of the human genome is of interest in
respects. It is the largest genome to be extensively sequence
being 25 times as large as any previously sequenced geno
eight times as large as the sum of all such genomes. It is t
vertebrate genome to be extensively sequenced. And, uniqu
the genome of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

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

- The genomic landscape shows marked variation in the d
tion of a number of features, including genes, trans
elements, GC content, CpG islands and recombination ra
gives us important clues about function. For example, th
opmentally important HOX gene clusters are the most repe
regions of the human genome, probably reflecting the very c

coordinate regulation of the genes in the clusters.

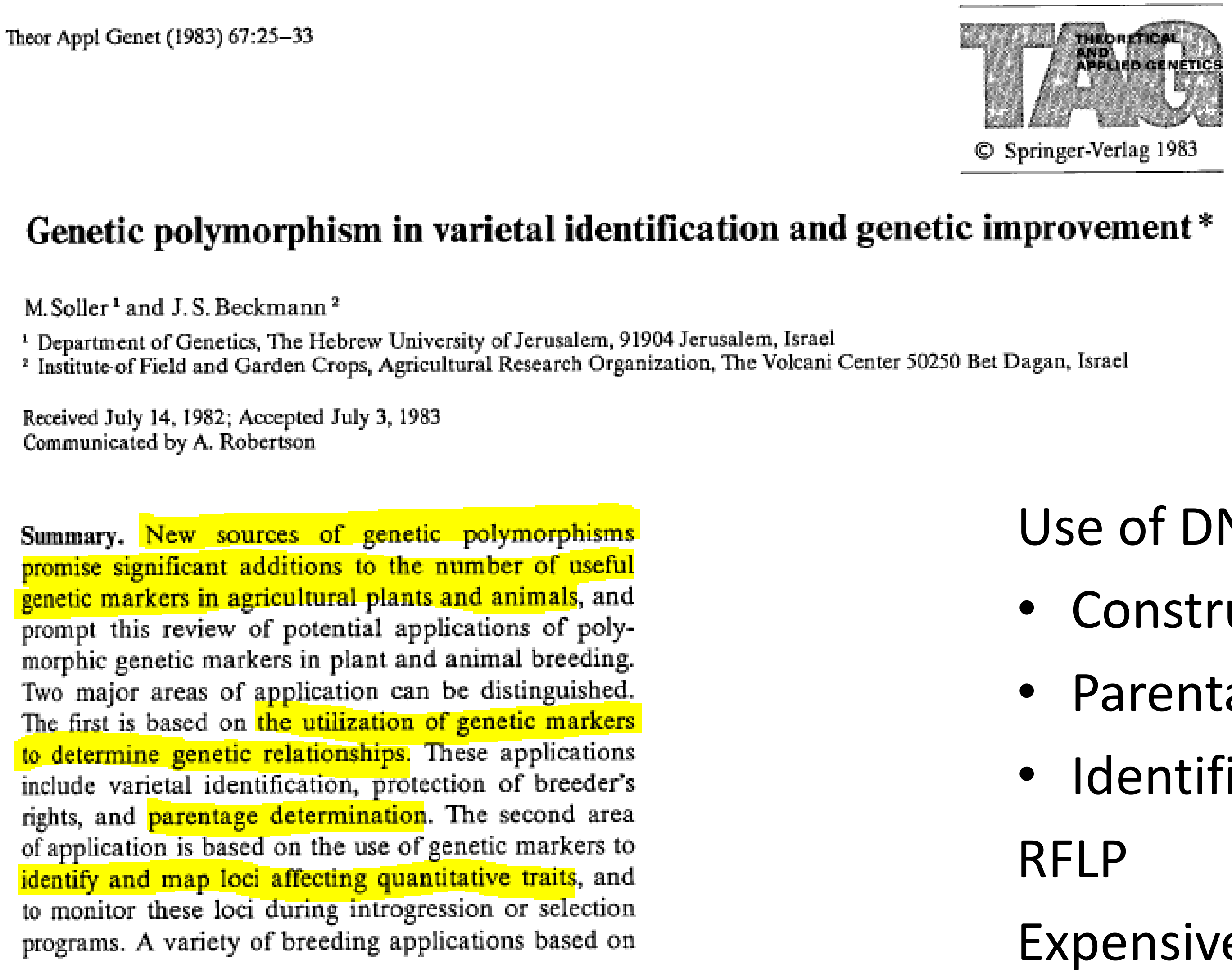
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the 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://neuroendoimmune.files.wordpress.com/2014/03/snp.png

Mutation < 1% < SNP

What are SNP used for?



Theor Appl Genet (1983) 67:25–33

THEORETICAL AND APPLIED GENETICS

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

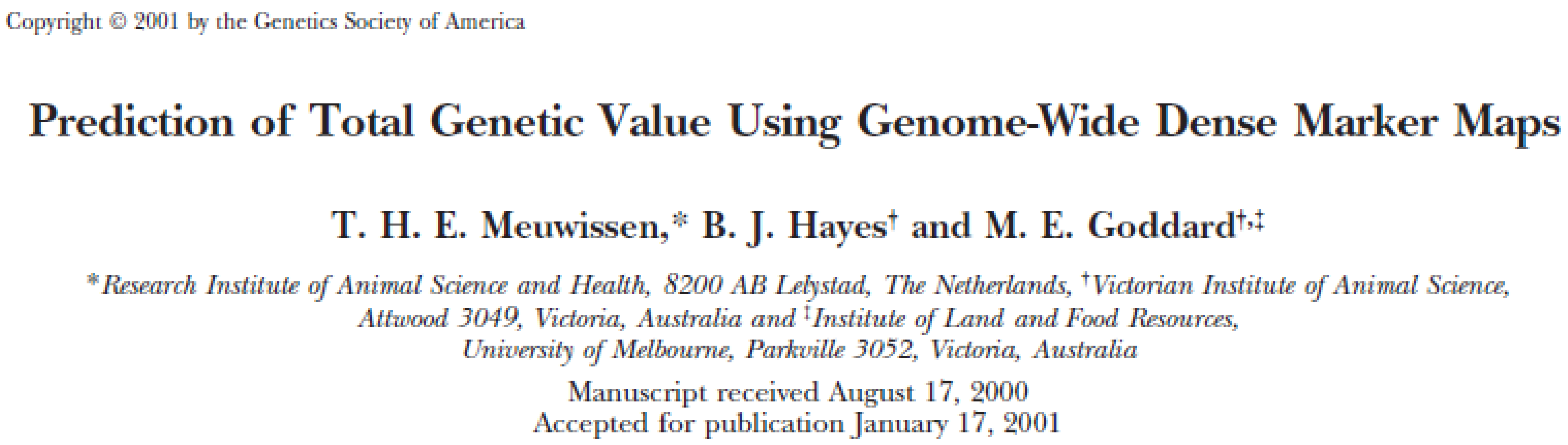
Use of DNA polymorphisms as genetic markers

- Construct genetic relationships
- Parentage determination
- Identification of QTL

RFLP

Expensive

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

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?

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

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

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

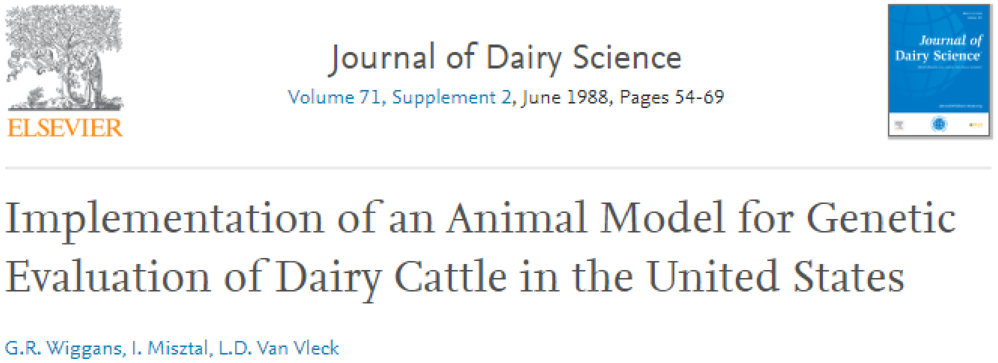
$$p(\mathbf{y},\mathbf{u}) = p(\mathbf{y}|\mathbf{u})\, p(\mathbf{u}) = \frac{1}{\sqrt{2\pi|\mathbf{R}|}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Wu})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Wu})} \frac{1}{\sqrt{2\pi|\mathbf{G}|}} e^{-\frac{1}{2}(\mathbf{u}-\mathbf{0})'\mathbf{G}^{-1}(\mathbf{u}-\mathbf{0})}$$

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

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

Henderson’s MME for dairy in 1989

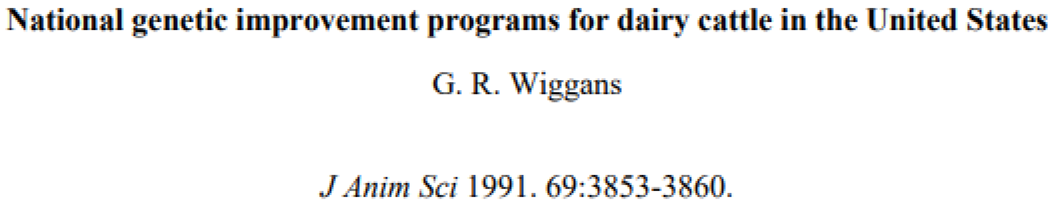
- BLUP (Henderson, 1949 - 1976)
- Implementation for dairy in 1989



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

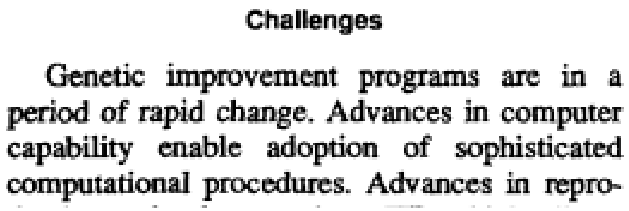


National genetic improvement programs for dairy cattle in the United States

G. R. Wiggans

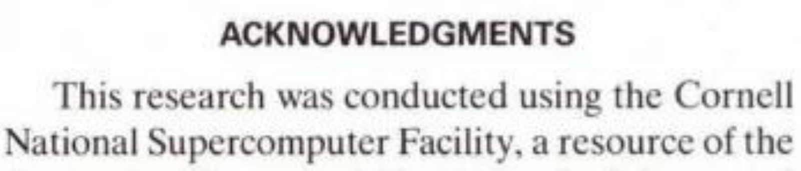
J Anim Sci 1991. 69:3853-3860.

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



Challenges

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

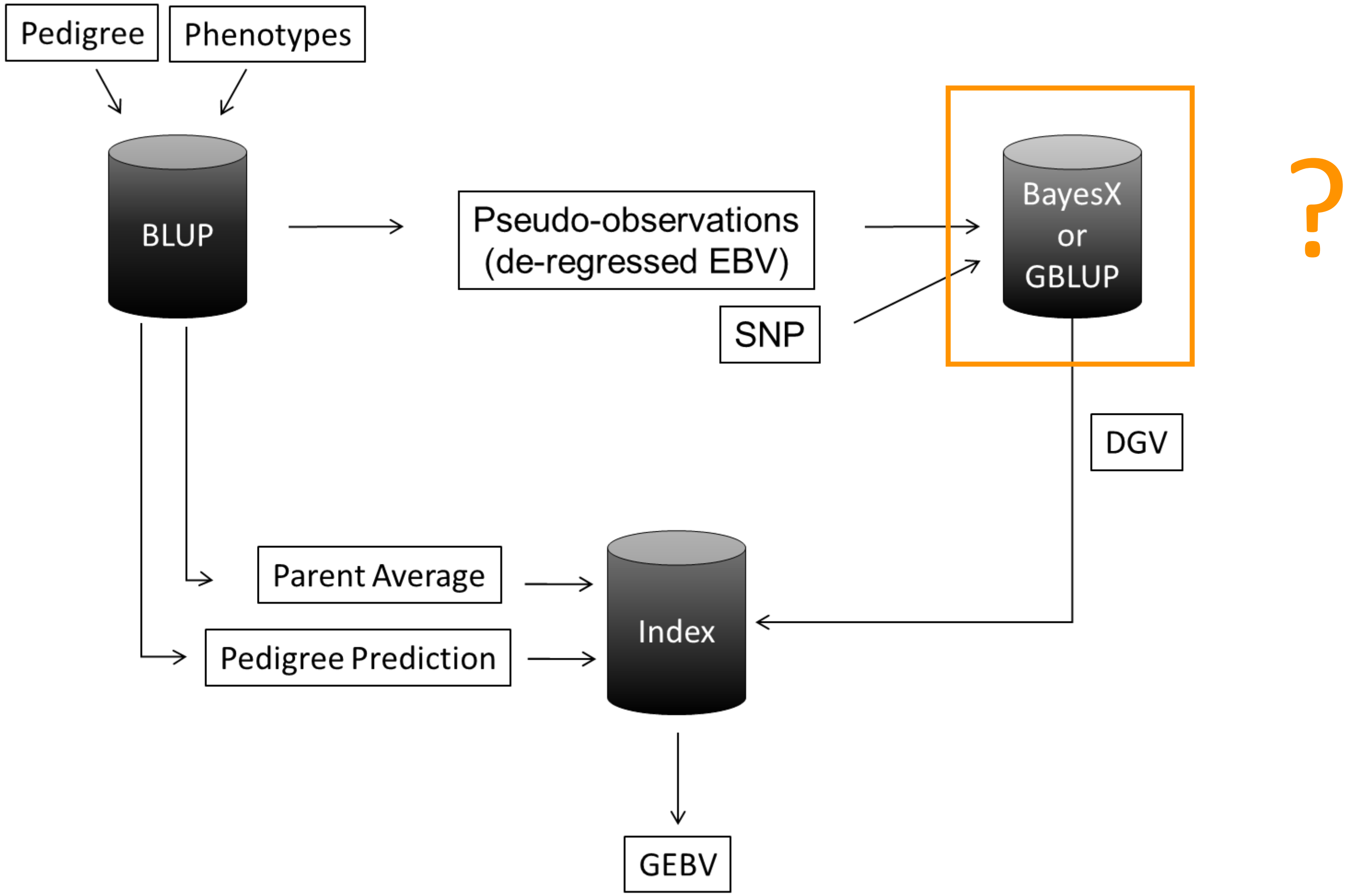


ACKNOWLEDGMENTS

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) $a_i \sim N\left(\mu, \sigma_{a_i}^2\right)$
  - Different variances for each SNP

- BayesB (Meuwissen et al., 2001)
  - $p\left(a_i \middle| \sigma_{a_i}^2, \pi\right) = \begin{cases} t\left(0, v, \sigma_{a_i}^2\right) or\ N\left(0, \sigma_{a_i}^2\right)\ with\ probability\ (1-\pi) \\ 0\ with\ probability\ \pi \end{cases}$

- When π = 0, BayesB becomes BayesA

Bayesian Alphabet

- BayesC (Habier et al., 2011)
  - $p(a_i|\sigma_a^2) = \begin{cases} N(0, \sigma_a^2)\ with\ probability\ (1-\pi) \\ 0\ with\ probability\ \pi \end{cases}$

- 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

- $a \sim N(0, \sigma_a^2)$

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

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

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

- 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

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

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

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

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

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

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

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

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

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!!!

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

…

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

Reversing the expression gives

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

GBLUP: equivalent to SNP-BLUP

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

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

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

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

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

VanRaden (2008)

Genomic relationship matrix

Genotypes {0,1,2}

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

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

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 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)

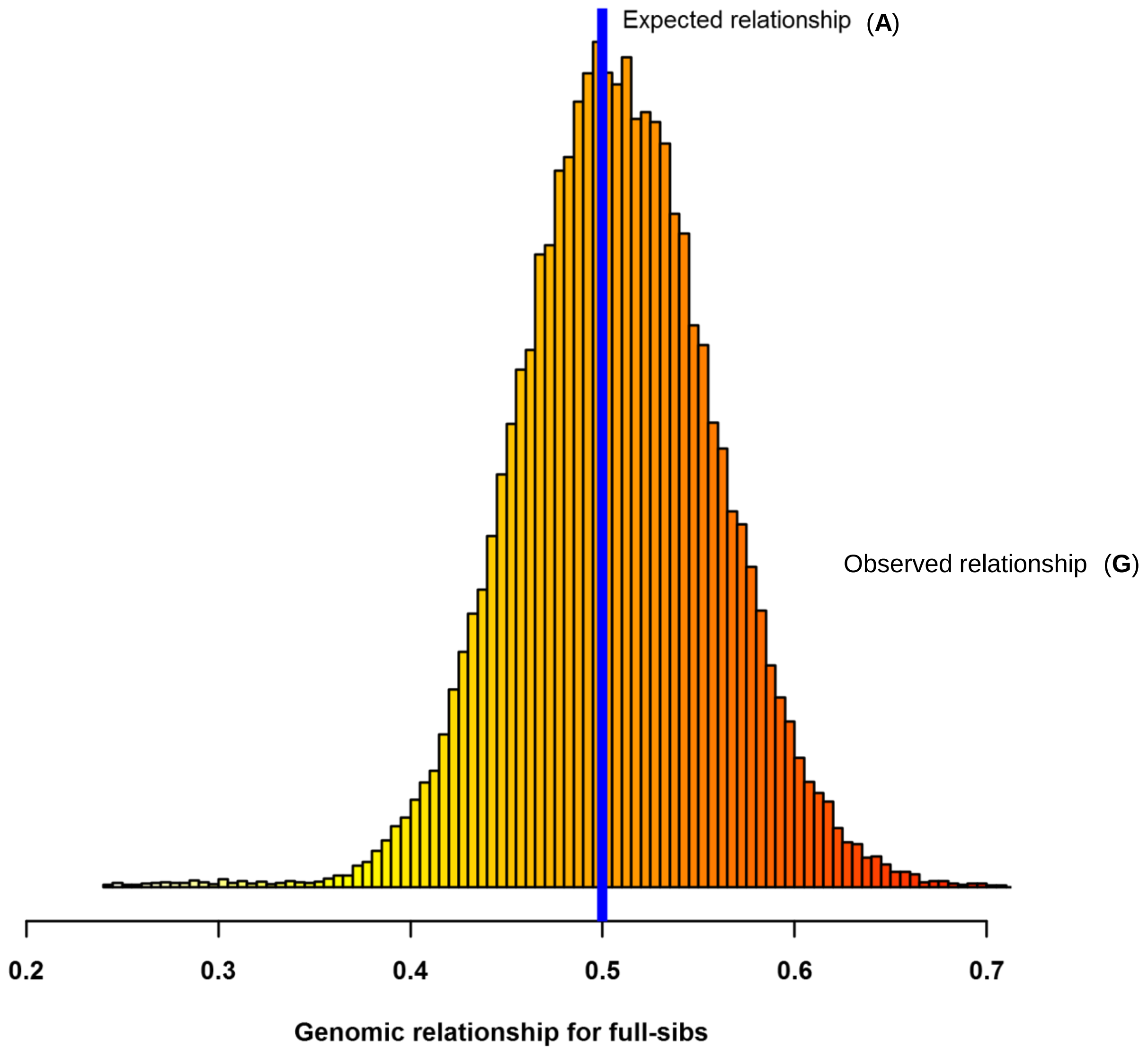
  Li and Horvitz 1953, Cockerham 1969, Ritland 1996, Caballero & Toro 2002, and many others

- With abundant marker data we can do better than this

Pedigree vs. Genomic relationships

- Identical By Descent Relationships 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 <u>expectation</u> 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



Adapted from Lourenco et al. (2015)

Genomic relationships

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

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

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

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

- 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”

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

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

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

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 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

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

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

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 10% in large populations

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

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

Aguilar et al., 2010
Christensen and Lund, 2010

Combining two sources of relationships

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

- **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 on the same scale

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

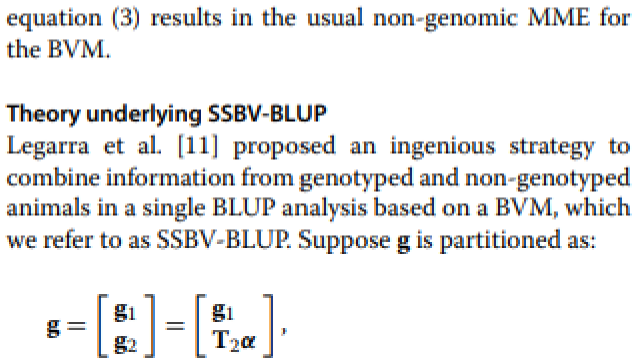
Misztal et al. (2009)
Legarra et al. (2009)
Aguilar et al. (2010)
Christensen & Lund (2010)

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

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

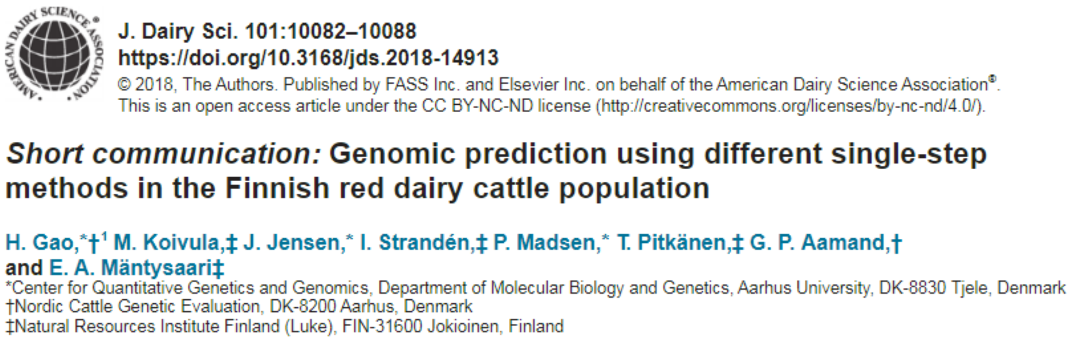


equation (3) results in the usual non-genomic MME for the BVM.

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 $\mathbf{g}$ is partitioned as:

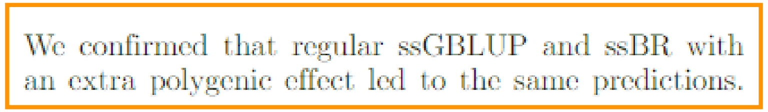
$$\mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{T}_2\alpha \end{bmatrix},$$



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,*†[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.

QC of SNP data in BLUPF90

ssGBLUP and GBLUP in BLUPF90