A relationship matrix combining pedigree and markers when some individuals are not genotyped

Andres Legarra

A RELATIONSHIP MATRIX COMBINING PEDIGREE AND MARKERS (WHEN SOME INDIVIDUALS ARE NOT GENOTYPED)

Andrés Legarra

INRA, UMR1388, Toulouse, France,

Andres.Legarra@toulouse.INRA.fr

Thanks

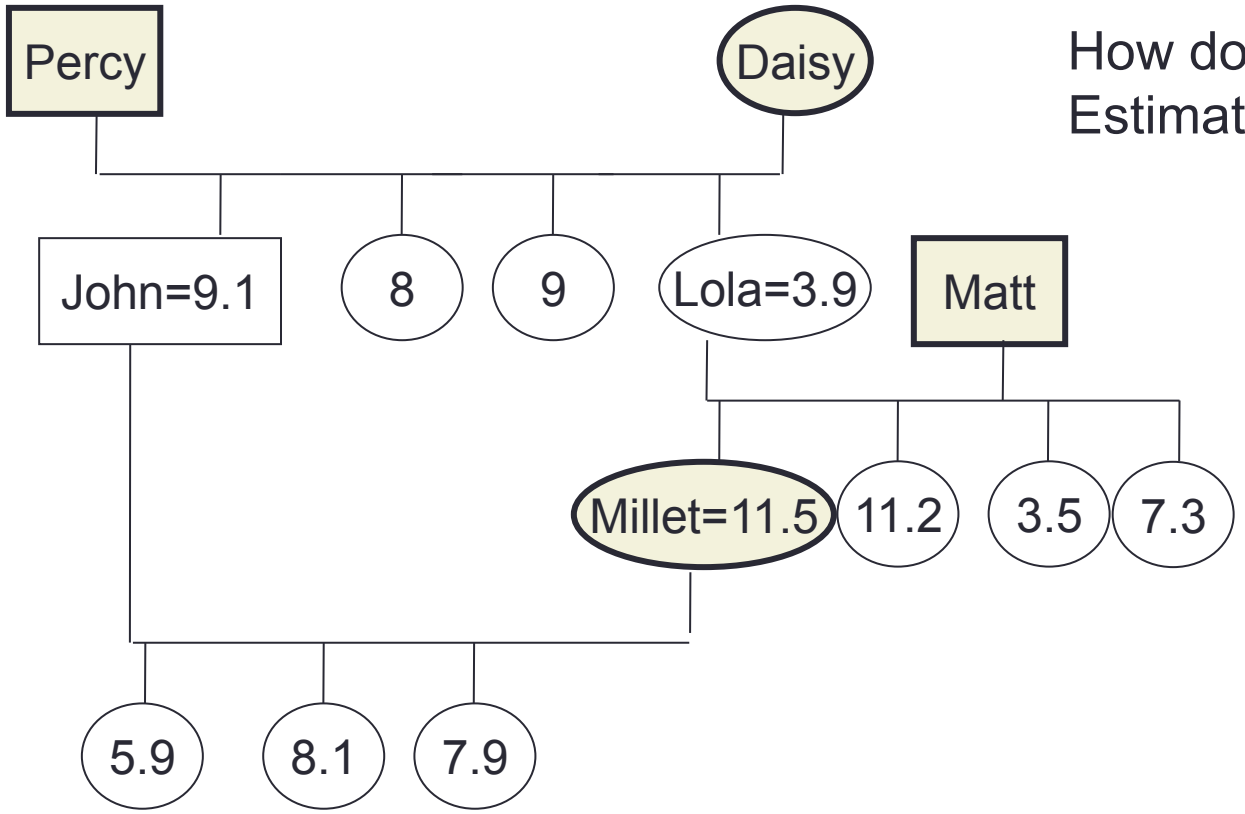
- Organisers for organizing everything & inviting me
- Projects GenSSeq and X-Gen (INRA)

- Work that I have been doing primarily with I Misztal (UGA, US), I Aguilar (INIA, Uruguay) and many other people

- Other group led by OF Christensen (University of Aarhus, DK) developed the theory in parallel
  - with fruitful cross-fecundation

Example

- Pedigree; grey is genotyped
- Numbers are records of a quantitative trait (e.g. weight)
- Can't easily assign a record to a genotyped individual



How do we predict?
Estimate heritablity? etc

Plan

- **Intro: pedigree & genomic relationship, why we need them**
- Derivation of a joint matrix **H**
- Compatibility of genomic and pedigree relationships

Pedigree relationships: **A**

- Additive relationships= 2*(kinships or coancestries)
  - $A_{ij} = 2\phi_{ij}$

- Pedigrees describe how genes are potentially transmitted
- Systematic "tabular" rules to compute any A_{ij} (Emik & Terrill 1947)
- The whole array of A_{ij} is disposed in a matrix $\boldsymbol{A}$.
- $\boldsymbol{A}^{-1}$ is very sparse and easy to create(Henderson 1976)
  - Extraordinary development of whole-pedigree methods in livestock genetics

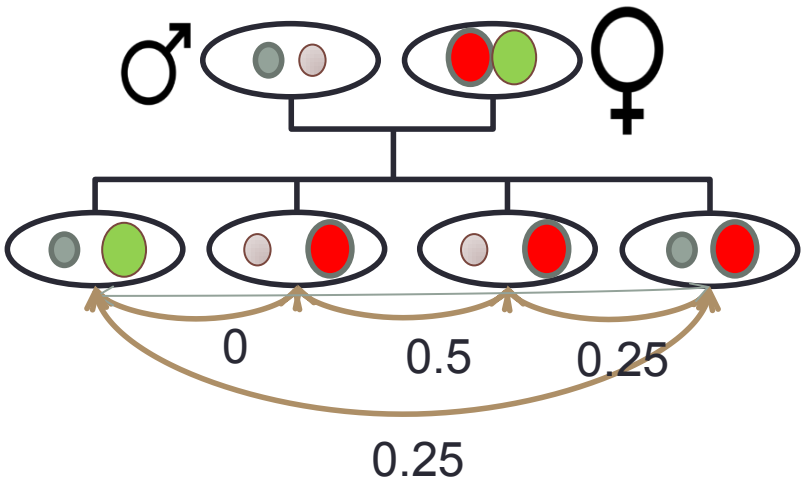
Genomic (or molecular) relationships: **G**

- The predecessors are poorly known
  - Li and Horvitz 1953, Cockerham 1969, Ritland 1996, Caballero & Toro 2002, VanRaden 2008 and many others

- Genomes are of finite size
  - Some sib pairs are more equal than others (Hill & Weir 2011, etc)
  - Pedigree relationships are not “fair”

Genomic (or molecular) relationships: **G**

- If we could see genes then we could just count
- Instead of genes, we see markers, which are *not* genes
  - Markers are stretches of DNA that can be accurately read across individuals
  - Biallelic SNP markers are used right now (e.g. A/a). Many of them: 50,000 to 800,000 / individual



VanRaden genomic relationships

- (VanRaden, 2008, more known as Yang et al., 2010)
- Crossproduct across numerically coded genotypes

- $G_{ij} = \frac{\mathbf{z}_i \mathbf{z}_j}{2\sum p_k q_k}$

  - $\boldsymbol{z}_i$: vector of n elements
    - with standardized genotypes as $\{0,1,2\} - 2p_k$ for genotypes $\{AA, Aa, aa\}$ at locus $k = 1, n$
  - p_k : across-population frequency of $\{a\}$ at locus k

- Whole-population $\boldsymbol{G} = \boldsymbol{ZDZ'}$
  - Semipositive definite, not easy to invert

Genomic and pedigree relationships

- Pedigree (**A**) are estimated IBD relationships, assuming « unrelated » founders
- Genomic (**G**) are Identical by state (IBS) relationships, corrected to be in IBD scale (see later)
- Genomic relationships are similar to pedigree relationships but more accurate

- If pedigree correct, typically crude $sd(\boldsymbol{G} - \boldsymbol{A}) \approx 0{,}04$

and $cor(\boldsymbol{A}, \boldsymbol{G}) \approx 0{,}80$

Applications

- Most applications come from the model
- $\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Wu} + \boldsymbol{e}$
  - Phenotype = environmental effects + genetic value + residual
- Assuming
  - $Var(\boldsymbol{u}) = \boldsymbol{A}\sigma_u^2$
  - $Var(\boldsymbol{u}) = \boldsymbol{G}\sigma_u^2$

  } Relationships
  - $Var(\boldsymbol{e}) = \boldsymbol{R}$ ⟶ Typically simple structure
- In (G)BLUP equations we use relationships:

$$\begin{pmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}W} \\ \boldsymbol{W'R^{-1}W} & \boldsymbol{W'R^{-1}W} + Var(\boldsymbol{u})^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{W'R^{-1}y} \end{pmatrix}$$

- $Var(\boldsymbol{u})^{-1} = \boldsymbol{G}^{-1}\sigma_u^{-2}$ or perhaps $Var(\boldsymbol{u})^{-1} = \boldsymbol{A}^{-1}\sigma_u^{-2}$

Genomic predictions and Pedigree predictions

- Relationships can be obtained from pedigree (pedigree relationships) or from markers (genomic relationships)
  - We expect markers to be better than pedigree because they are more “real” but they are expensive… (40-150 $ / individual)
  - We expect Artificial Selection based on markers (“Genomic Selection”) to be more efficient than based on pedigree

Genomic predictions and Pedigree predictions

- genomic predictions are 10-25% more accurate than pedigree predictions in terms of cross-validation R^2
  - e.g. VanRaden et al. 2009 (dairy cattle)

Table 2. Coefficients of determination ($R^2 \times 100$) for 2008 daughter d

Trait	pedigree	genomic
Net merit	11	28
Milk yield	28	47
Fat yield	15	42
Protein yield	27	47
Fat percentage	25	55
Protein percentage	28	51
Productive life	17	26
SCS	23	37

VanRaden et al **J. Dairy Sci. 92:16–24**

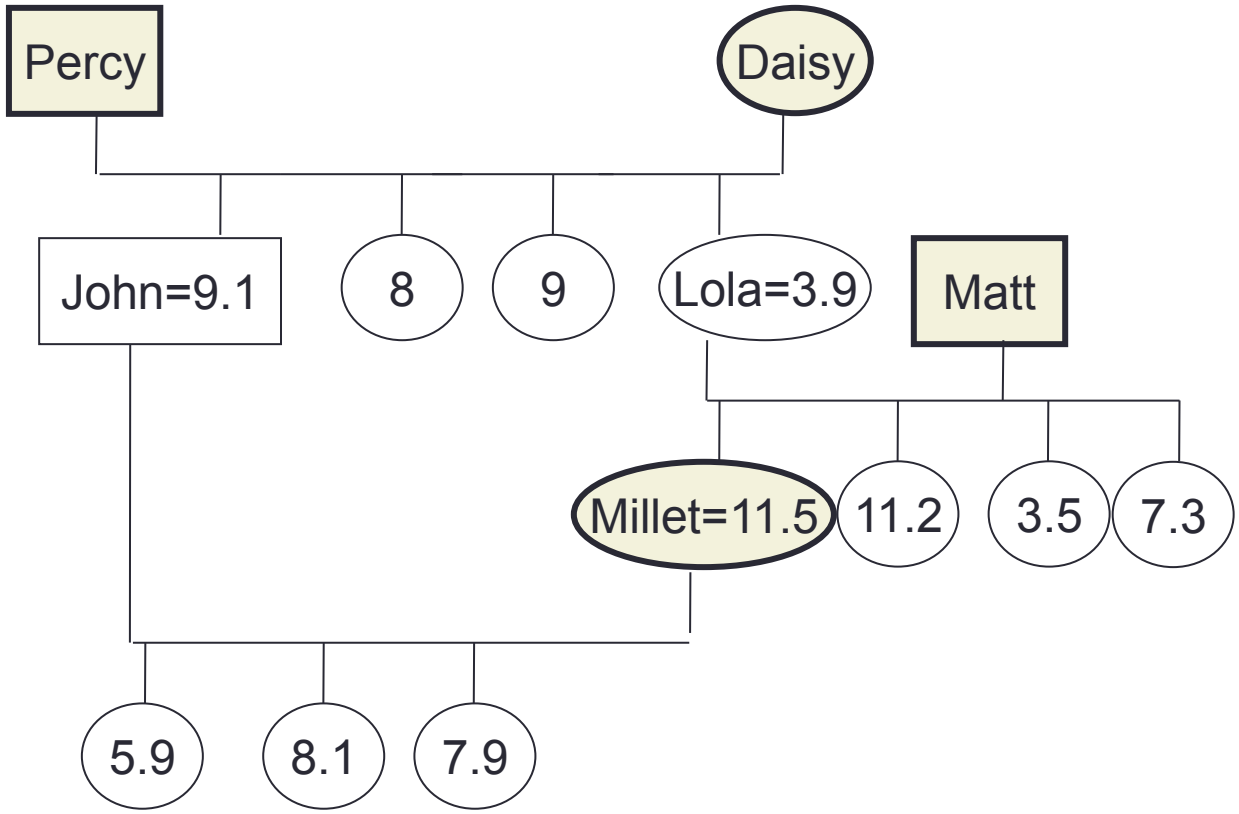
- Similar results in sheep, pigs, chicken and goats (and plants)

Pedigrees in livestock genetics

- They are deep and connect most animals
- From 100,000's to 1,000,000's
- However, only some animals are genotyped
  - Important animals such as bulls, also recent animals
  - MANY animals are ungenotyped (perhaps 99%)
- This makes us unhappy
  - **A** spans all animals but has no marker information and is less precise
  - **G** is more precise but does not include all animals
  - So far, we use horrible procedures for precorrection

Example

- Grey is genotyped
- Numbers are records (e.g. weight)
- Can't easily assign a record to a genotyped individual



Plan

- Intro: pedigree & genomic relationship, why we need them
- **Derivation of a joint matrix H**
- Compatibility of genomic and pedigree relationships

- Things would be simple if we had genomic relationships for everyone (Legarra et al., 2009)
- Things would be simple if we could add genotypes for all animals (Christensen et al., 2010)

- Things would be simple if we had genomic relationships for everyone (Legarra et al., 2009)
- **Things would be simple if we could add genotypes for all animals (Christensen et al., 2010)**

Single Step as a missing data problem

- We can see genotype as a missing data problem (Christensen & Lund, 2010)
- Use the prediction and the distribution of the prediction

Missing data

Fill-in missing data: data augmentation

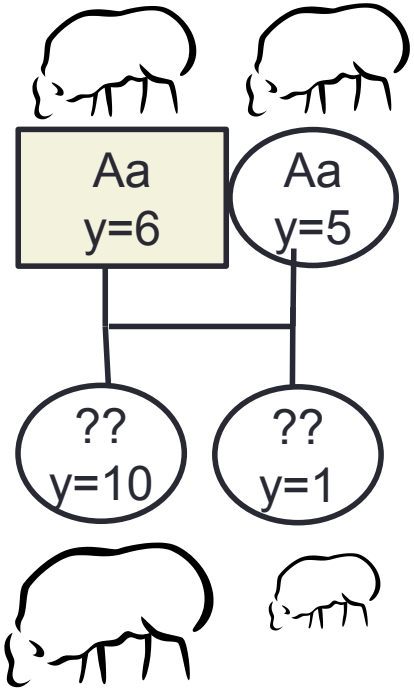
- *« data augmentation refers to a scheme of augmenting the observed data so as to make it more easy to analyze »* (Tanner & Wong, 1987)

- Augmenting = adding genotypes

- Imputing algorithms work from low to high density markers
- For animals nongenotyped (at all), they *may* give a point estimate based on most likely genotype
- Why is this bad?

Problem with point estimates of genotypes

- Imagine a major gene

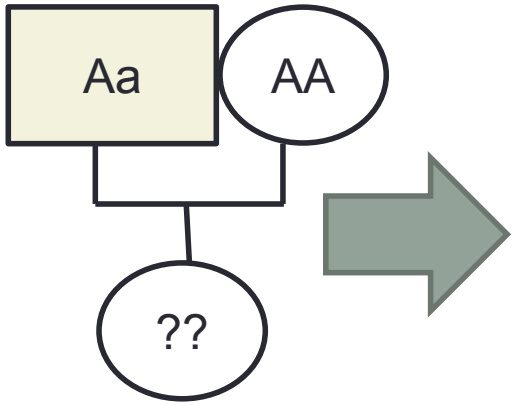


- Point estimate of genotype of the descendants: “Aa”
- Clearly, based on y there is Mendelian segregation where one descendant received “AA” and the other “aa”
- There is variation of true genotype around the point estimate of the genotype
- If we do not consider this variation we consider the offspring as identical twins

Augmenting genotypes

- Gengler et al. (2007) conceived an algebraic way to deal with these point estimates (== to McPeek et al. 2004)
- Christensen & Lund (2010) showed how to take the variation into account

- Genotype of descendants = half their parents + Mendelian sampling



AA with probability ½
Aa with probability ½

E(Genotype) = $\frac{3}{2}$"A" + $\frac{1}{2}$"a"

Variance(Genotype)=$\frac{1}{4}$"A" + $\frac{1}{4}$"a"

Augmenting genotypes

Genotype = $\frac{3}{2}"A" + \frac{1}{2}"a"$

Variance(Genotype)=$\frac{1}{4}"A" + \frac{1}{4}"a"$

- Yes this is weird but it allows linear and algebraic treatment of an almost impossible problem
- You can see it as a linear simplification of a superpolynomial problem

- This allows using the classical machinery of animal breeding (relationships and matrix algebra)

Inferring genotypes

- Gengler's gene content prediction (2007)
- Linear approximation to the imputation problem
- This method can be applied to any member of a pedigree and generalized to a set of individuals

Expected genotype

Observed genotype

$$\hat{\mathbf{z}}_{non\ genotyped} = E\left(\mathbf{z}_{non\ genotyped} \mid \mathbf{z}_{genotyped}\right) = \mathbf{A}_{1,2}\mathbf{A}_{2,2}^{-1}\left(\mathbf{z}_{genotyped} - \mathbf{2}p\right)$$

$$Var\left(\hat{\mathbf{z}}_{non\ genotyped}\right) = Var\left(\mathbf{z}_{non\ genotyped} \mid \mathbf{z}_{genotyped}\right) = \left(\mathbf{A}_{1,1} - \mathbf{A}_{1,2}\mathbf{A}_{2,2}^{-1}\mathbf{A}_{2,1}\right)2pq$$

non genotyped

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

genotyped

Christensen & Lund key idea:

$$\boldsymbol{u} = \begin{pmatrix} \boldsymbol{u}_{ng} \\ \boldsymbol{u}_{g} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}_{ng} \\ \boldsymbol{Z}_{g} \end{pmatrix} \boldsymbol{a}$$

Breeding values

SNP effects

ng: « non genotyped »
g: « genotyped »

Chistensen & Lund use $Var(A) = E\big(Var(A|B)\big) + Var\big(E(A|B)\big)$ to consider the prediction of the genotype and its variance

$$\mathrm{Var}(\boldsymbol{u}) = \begin{pmatrix} \widehat{\boldsymbol{Z}}_{ng} \\ \boldsymbol{Z}_{g} \end{pmatrix} Var(\boldsymbol{a}) \begin{pmatrix} \widehat{\boldsymbol{Z}}_{ng} & \boldsymbol{Z}'_{g} \end{pmatrix} + \begin{pmatrix} Var(\widehat{\boldsymbol{Z}}_{ng}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} Var(\boldsymbol{a})$$

$E(\boldsymbol{Z}_{ng}|\boldsymbol{Z}_{g})$

$1/2\Sigma p_i q_i$

$Var(\boldsymbol{Z}_{ng}|\boldsymbol{Z}_{g})$

Using Gengler's results

Resulting in:

Covariances of all animals

Legarra et al. 2009; Aguilar et al., 2010; Christensen & Lund, 2010

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}=\mathbf{H}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}=\begin{bmatrix}\overbrace{\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}+\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}}^{\text{non genotyped}} & \underbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}}_{\text{genotyped}}\\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}\end{bmatrix}$$

Let $\mathbf{A}=\begin{bmatrix}\overbrace{\mathbf{A}_{11}}^{\text{non genotyped}} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \underbrace{\mathbf{A}_{22}}_{}\end{bmatrix}$

Covariances of all animals

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}=\mathbf{H}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}=\left[\begin{array}{cc|c}\overbrace{\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}}+\overbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}} & & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\\ \hline \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & & \mathbf{G}\end{array}\right]$$

This is the variance of prediction of genotypes *from* genotyped *to* non-genotyped

This is the error in the prediction

The prediction « generates » a covariance

G comes from genotypes

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}=\mathbf{H}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}=$$

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

- Incredibly: $\mathbf{H}^{-1}$ is very simple:

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

…and avoiding « double counting »

Inverse of the regular pedigree relationship matrix

Correcting for genomic relationships…

- **Things would be simple if we had genomic relationships for everyone (Legarra et al., 2009)**
- Things would be simple if we could add genotypes for all animals (Christensen et al., 2010)

Overall modification

- Look at **A** as a « prior » (pedigree) relationship and to **G** as an « observed » (genomic) relationship
  - **G** is observed for some individuals only, whose « a priori » (pedigree) relationship matrix was $\mathbf{A}_{22}$
- Try to construct a « posterior » relationship matrix

Joint distributions

Unconditional distribution of genetic values of Genotyped individuals

$$p(\mathbf{u}_2) = N(\mathbf{0}, \mathbf{G}) \text{ and}$$

After seeing their genotypes !

Conditional distribution of Non-Genotyped individuals

$$p(\mathbf{u}_1 | \mathbf{u}_2) = N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)$$

Because they have no genotypes, this depends only on pedigree

$$p(\mathbf{u}_1, \mathbf{u}_2) = p(\mathbf{u}_2)\, p(\mathbf{u}_1 | \mathbf{u}_2)$$

Joint distribution

Joint distributions

$$p(\mathbf{u}_1, \mathbf{u}_2) = p(\mathbf{u}_1, | \mathbf{u}_2) p(\mathbf{u}_2)$$

$$= p(\mathbf{u}_1 | \mathbf{u}_2) p(\mathbf{u}_2)$$

prediction of non genotyped from genotyped

"Genomic" relationships

$$\propto \exp[-0.5(\mathbf{u}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2)'\mathbf{A}^{11}(\mathbf{u}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2)]\exp[-0.5\mathbf{u}_2'\mathbf{G}^{-1}\mathbf{u}_2]$$

$$= \exp\left(-0.5 \begin{bmatrix} \mathbf{u}_1' & \mathbf{u}_2' \end{bmatrix} \begin{bmatrix} \mathbf{A}^{11} & -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11} & \mathbf{G}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} \right)$$

$$= \exp\left(-0.5 \begin{bmatrix} \mathbf{u}_1' & \mathbf{u}_2' \end{bmatrix} \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{G}^{-1} + \mathbf{A}^{22} - \mathbf{A}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} \right).$$

…for those inclined to algebra

Covariances of all animals

Legarra et al. 2009; Aguilar et al., 2010; Christensen & Lund, 2010

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}=\mathbf{H}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}=\begin{bmatrix}\overbrace{\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}+\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}}^{\text{non genotyped}} & \underbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}}_{\text{genotyped}}\\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}\end{bmatrix}$$

Exactly same results…

Overall modification: example

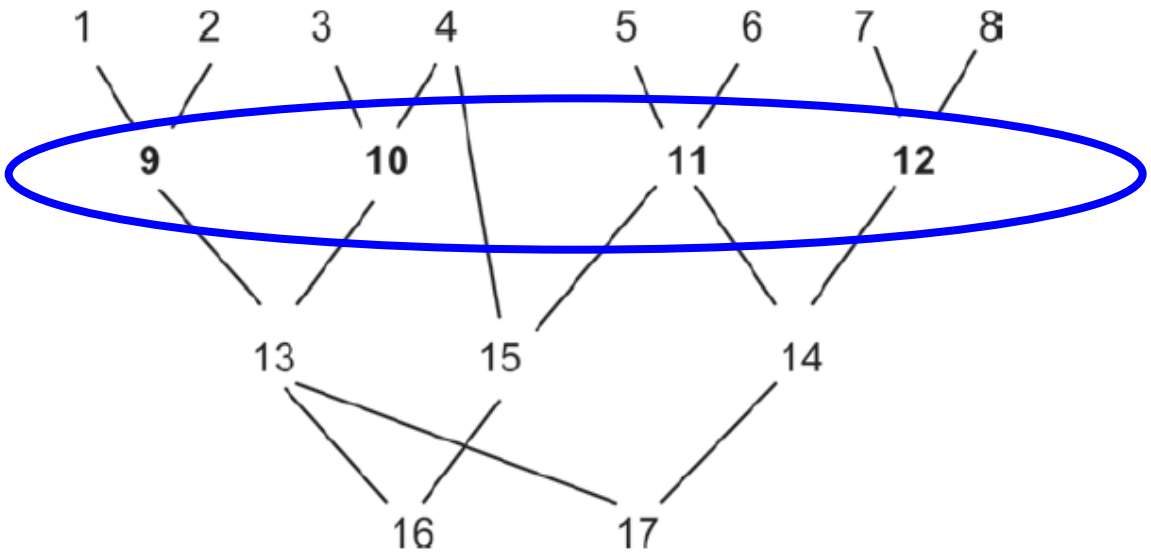


Figure 1. Example pedigree. Genotyped animals are in bold.

Overall modification: example

Table 1. Numerator relationship matrix **A** for the pedigree in Figure 1[1]

1.00								0.50				0.25			0.13	0.13
	1.00							0.50				0.25			0.13	0.13
		1.00							0.50			0.25			0.13	0.13
			1.00						0.50			0.25		0.50	0.38	0.13
				1.00						0.50			0.25	0.25	0.13	0.13
					1.00					0.50			0.25	0.25	0.13	0.13
						1.00					0.50		0.25			0.13
							1.00				0.50		0.25			0.13
0.50	0.50							**1.00**				0.50			0.25	0.25
		0.50	0.50						**1.00**			0.50		0.25	0.38	0.25
				0.50	0.50					**1.00**			0.50	0.50	0.25	0.25
						0.50	0.50				**1.00**		0.50			0.25
0.25	0.25	0.25	0.25					0.50	0.50			1.00		0.13	0.56	0.50
				0.25	0.25	0.25	0.25			0.50	0.50		1.00	0.25	0.13	0.50
			0.50	0.25	0.25				0.25	0.50		0.13	0.25	1.00	0.56	0.19
0.13	0.13	0.13	0.38	0.13	0.13			0.25	0.38	0.25		0.56	0.13	0.56	1.06	0.34
0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.25	0.25	0.25	0.25	0.50	0.50	0.19	0.34	1.00

[1]Cells with 0 are empty to show the pattern. Coefficients for genotyped animals are in bold. Matrix $\mathbf{A}_g$ is obtained by setting the out-of-diagonal coefficients of genotyped animals to 0.7.

This is the regular relationship matrix. Assume now that animals 9 to 12 have a genomic relationship of 0.7

Overall modification: example

Table 3. Modified relationship matrix **H** including genomic information for genotyped animals and all relatives for the pedigree in Figure 1[1]

1.00		0.18	0.18	0.18	0.18	0.18	0.18	0.50	0.35	0.35	0.35	0.43	0.35	0.26	0.34	0.39
	1.00	0.18	0.18	0.18	0.18	0.18	0.18	0.50	0.35	0.35	0.35	0.43	0.35	0.26	0.34	0.39
0.18	0.18	1.00		0.18	0.18	0.18	0.18	0.35	0.50	0.35	0.35	0.43	0.35	0.18	0.30	0.39
0.18	0.18		1.00	0.18	0.18	0.18	0.18	0.35	0.50	0.35	0.35	0.43	0.35	0.68	0.55	0.39
0.18	0.18	0.18	0.18	1.00		0.18	0.18	0.35	0.35	0.50	0.35	0.35	0.43	0.34	0.34	0.39
0.18	0.18	0.18	0.18		1.00	0.18	0.18	0.35	0.35	0.50	0.35	0.35	0.43	0.34	0.34	0.39
0.18	0.18	0.18	0.18	0.18	0.18	1.00		0.35	0.35	0.35	0.50	0.35	0.43	0.26	0.31	0.39
0.18	0.18	0.18	0.18	0.18	0.18		1.00	0.35	0.35	0.35	0.50	0.35	0.43	0.26	0.31	0.39
0.50	0.50	0.35	0.35	0.35	0.35	0.35	0.35	**1.00**	**0.70**	**0.70**	**0.70**	0.85	0.70	0.53	0.69	0.78
0.35	0.35	0.50	0.50	0.35	0.35	0.35	0.35	**0.70**	**1.00**	**0.70**	**0.70**	0.85	0.70	0.60	0.73	0.78
0.35	0.35	0.35	0.35	0.50	0.50	0.35	0.35	**0.70**	**0.70**	**1.00**	**0.70**	0.70	0.85	0.68	0.69	0.78
0.35	0.35	0.35	0.35	0.35	0.35	0.50	0.50	**0.70**	**0.70**	**0.70**	**1.00**	0.70	0.85	0.53	0.61	0.78
0.43	0.43	0.43	0.43	0.35	0.35	0.35	0.35	**0.85**	**0.85**	**0.70**	**0.70**	1.35	0.70	0.56	0.96	1.03
0.35	0.35	0.35	0.35	0.43	0.43	0.43	0.43	0.70	0.70	0.85	0.85	0.70	1.35	0.60	0.65	1.03
0.26	0.26	0.18	0.68	0.34	0.34	0.26	0.26	0.53	0.60	0.68	0.53	0.56	0.60	1.18	0.87	0.58
0.34	0.34	0.30	0.55	0.34	0.34	0.31	0.31	0.69	0.73	0.69	0.61	0.96	0.65	0.87	1.41	0.80
0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.78	0.78	0.78	0.78	1.03	1.03	0.58	0.80	1.53

[1]Cells with 0 are empty to show the pattern. Coefficients for genotyped animals are in bold.

This parents now are related

G

This guy now is inbred

Understanding H matrix

- 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 updating of the pedigree matrix based on new information from genotypes
- The approximation of multivariate normality is good because we have *many* markers

- Typically
  - $\mathbf{A}^{-1}$ in the millions but extremely sparse
  - **G** and $\mathbf{A}_{22}$ in the thousands
  - Leads to a very efficient method of genomic evaluation:
    - Single Step GBLUP

Single step GBLUP

Single Step = Your regular BLUP with small modifications

W: incidence matrix of animals on data

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}\mathbf{R}^{-1}\mathbf{W}+\mathbf{H}^{-1}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

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

G

A: pedigree relationship matrix

$\mathbf{A}_{22}$: pedigree matrix among genotyped individuals

This **G** could be *any* matrix describing « genomic » covariances of breeding values;

it does not restrict to VanRaden's (2008) GBLUP

Single Step GBLUP

- Easy modification to a general purpose BLUP software
  - Only changes: addition of $\boldsymbol{G}^{-1}$ and $\boldsymbol{A}_{22}^{-1}$
  - Matrices $\boldsymbol{G}^{-1}$ and $\boldsymbol{A}_{22}^{-1}$ can be computed with external tools

- Can fit any model (probit, GxE,…)

- Simple extraction of SNP effects for prediction or (multimarker) GWAS:

$$\hat{\boldsymbol{a}} = \boldsymbol{Z}'\boldsymbol{G}^{-1}\hat{\boldsymbol{u}}_2/k$$

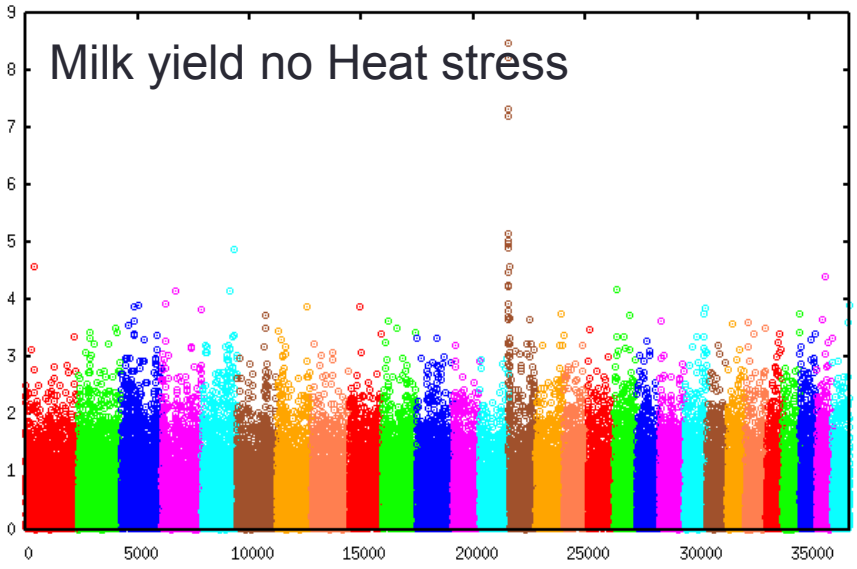
Some results in Pigs

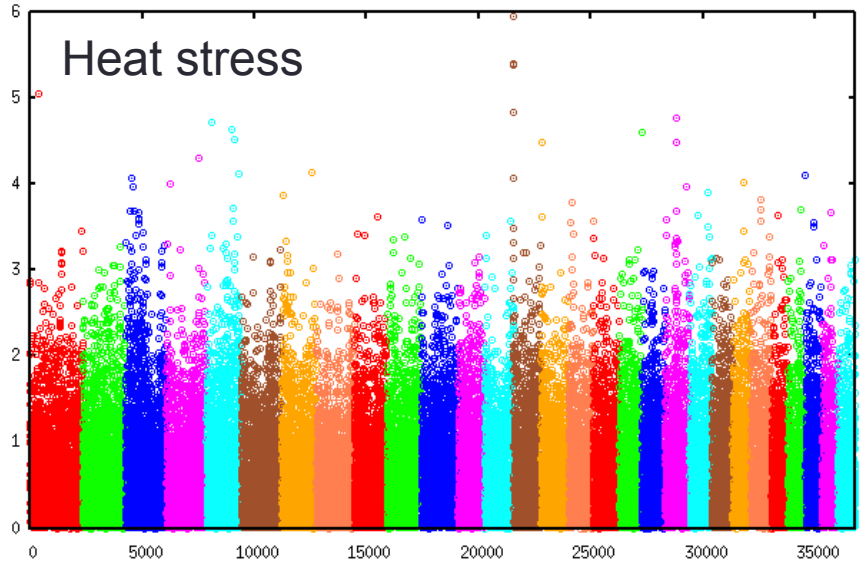
- Christensen et al., 2012
  - Joint two-trait analysis: daily gain (massively recorded) and feed efficiency (scarcely recorded)
  - 2600 genotyped, 300,000 records
  - Single Step increased accuracy by 0.10 in both traits compared to pedigree BLUP and reduced bias compared to simple GBLUP
- Lourenço et al., 2014, PIC data
  - Litter size, fertility
  - 2,000,000 animals in data, 5,000 animals genotyped
  - Single Step increased accuracy by 0.10-0.20 compared to pedigree BLUP

Single-Step Heat Stress GWAS

- Aguilar et al., unpublished
- Multiple-Trait Test-Day model heat tolerance
  - ~ 90 millions records, ~ 9 millions pedigrees
  - ~ 3,800 genotyped bulls
- Computing time
  - Complete evaluation ~ 16 h

Marker effects (after backsolving)





Plan

- Intro: pedigree & genomic relationship, why we need them
- Derivation of a joint matrix H
- **Compatibility of genomic and pedigree relationships**

Compatibility of marker and pedigree relationships

- Populations evolve with time, but genotypes came years after pedigree started
- Genomic Predictions are shifted from Pedigree Predictions
  - This makes them not directly comparable
- Underlying hypothesis of Christensen & Lund (allelic frequencies constant across time) or Legarra et al. (average genetic value does not change) false
- This can be modelled in a quantitative framework

Compatibility of marker and pedigree relationships

$\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{Z}'/k$ or $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}'$

Consider a model $\boldsymbol{y} = \mu + \boldsymbol{u} + \boldsymbol{e}$, $Var(u) = \boldsymbol{G}$

- Adding or substracting constants from **G** shifts $\hat{\boldsymbol{u}}$ by a constant absorbed by $\hat{\mu}$
- Multiplying **G** by a constant changes the genetic variance

Compatibility of marker and pedigree relationships

- The population for which $\text{average}(\boldsymbol{u}) = 0$ and for which the genetic variance is defined is called the *genetic base*
  - *Founders of the pedigree in classical* ***A***
  - *Whole set of genotyped animals in most typical* ***G***

- Typically, genotyped animals come *after* pedigree starts
  - e.g. Lacaune sheep pedigree go back to 1960 but genotypes start in 1995
- Drift (and selection) causes :
  - Average genetic values “drift” (in particular in small populations)
  - Genetic variance reduces

Compatibility of marker and pedigree relationships

- Vitezica et al. (2011) and Christensen et al. (2012) provided an unbiased method that forces the same genetic base across **G** and **A** :
  - $\boldsymbol{G}^* = a + b\boldsymbol{G}$
  - a accounts for old relationships among non genotyped ancestors
  - b accounts for reduction in the genetic variance
  - a and b can be obtained equating average inbreeding and average relationships:

$$a + b\,\overline{\boldsymbol{G}} = \overline{\boldsymbol{A}}_{22}$$

$$a + b\,\overline{diag(\boldsymbol{G})} = \overline{\left(diag(\boldsymbol{A_{22}})\right)}$$

In H-W $b = 1 - a/2$ and this is Wright's fixation index (Powell et al., 2011):

$$\left(1 - \frac{G^*_{ij}}{2}\right) = \left(1 - \frac{G_{ij}}{2}\right)\left(1 - \frac{(\overline{\boldsymbol{A}}_{\boldsymbol{22}} - \overline{\boldsymbol{G}})}{2}\right)$$

Christensen, 2012

- Christensen (2012) suggests fitting **A** to **G** instead of the opposite
  - Ancestral relationships that can be seen in **G** go undetected in **A**
- Christensen analitically integrates out p_i (=allele frequencies) in a model that
  - uses $p = 0.5$ as reference in ALL loci
  - uses a relationship matrix $\mathbf{A}^{\gamma}$ with related founders

Relationship across founders

Classically we assume

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- Christensen changes this into:

$$\mathbf{A}^{\gamma} = \begin{pmatrix} 1+\frac{\gamma}{2} & \gamma & \gamma & \gamma \\ \gamma & 1+\frac{\gamma}{2} & \gamma & \gamma \\ \gamma & \gamma & 1+\frac{\gamma}{2} & \gamma \\ \gamma & \gamma & \gamma & 1+\frac{\gamma}{2} \end{pmatrix}$$

He was unaware of Jacquard (1974) who posited this structure

Conclusions

- We have a rather good theory on mixing pedigree and genomic relationships for a single population
- This theory is useful for genomic predictions and for GWAS in complex scenarios such as livestock
- The associated computational methods are quite efficient

- BUT
- It is sensible to pedigree or genotyping mistakes (label switching)
- Compatibility needs a reasonable data set (representative samples)

TODO list

- Extend to multiple origins (=crosses of lines or breeds)
- Include linkage among markers (useful ?)
- Improve computational algorithms

- Understand those differences between *realized* (**G**) and *expected* (**A**) relationships, in order to come up with a comprehensive theory

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta *et al.*, 2010 Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci 93**:** 743-752.

Aguilar, I., I. Misztal, A. Legarra and S. Tsuruta, 2011 Efficient computations of genomic relationship matrix and other matrices used in the single-step evaluation. Journal of Animal Breeding and Genetics 128**:** 422-428.

Emik, L. O., and C. E. Terrill, 1949 Systematic procedures for calculating inbreeding coefficients. J Hered 40**:** 51-55.

LI, C. C., and D. G. HORVITZ, 1953 Some methods of estimating the inbreeding coefficient. Am J Hum Genet 5**:** 107-117.

Cockerham, C. C., 1969 Variance of gene frequencies. Evolution 23**:** 72-84.

Ritland, K., 1996 Estimators for pairwise relatedness and individual inbreeding coefficients. Genetical research 67**:** 175-185.

Caballero, A., and M. A. Toro, 2002 Analysis of genetic diversity for the management of conserved subdivided populations. Conservation genetics 3**:** 289.

VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91**:** 4414-4423.

Hill, W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res (Camb)**:** 1-18.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42**:** 565-569.

VanRaden, P. M., C. P. V. Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel *et al.*, 2009 Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92**:** 16-24.

Legarra, A., I. Aguilar and I. Misztal, 2009 A relationship matrix including full pedigree and genomic information. J Dairy Sci 92**:** 4656-4663.

Christensen, O. F., and M. S. Lund, 2010 Genomic prediction when some animals are not genotyped. Genet Sel Evol 42**:** 2.

Tanner, M. A., and W. H. Wong, 1987 The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82**:** 528-540.

Gengler, N., P. Mayeres and M. Szydlowski, 2007 A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. animal 1**:** 21-28.

McPeek, M. S., X. Wu and C. Ober, 2004 Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60**:** 359-367.

Christensen, O. F., 2012 Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. GENETICS SELECTION EVOLUTION 44**:** 37.

Lourenco, D., I. Misztal, S. Tsuruta, I. Aguilar, T. Lawlor *et al.*, 2014 Are evaluations on young genotyped animals benefiting from the past generations? Journal of Dairy Science 97**:** 3930-3942.

Chen, C., I. Misztal, I. Aguilar, S. Tsuruta, S. Aggrey *et al.*, 2011 Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. Journal of Animal Science 89**:** 23-28.

Vitezica, Z., I. Aguilar, I. Misztal and A. Legarra, 2011 Bias in genomic predictions for populations under selection. Genetics Research**:** In press.

Christensen, O. F., 2012 Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. GENETICS SELECTION EVOLUTION 44**:** 37.

Jacquard, A., 1970 Genetic structures of populations. Structures genetiques des populations.

General review:

Legarra, A., O. F. Christensen, I. Aguilar and I. Misztal, 2014 Single Step, A General Approach For Genomic Selection. Livestock Science.