# Across-Breeds Ancestral Relationships and Metafounders for Genomic Evaluation

**A. Legarra[1], O.F. Christensen[2], Z.G. Vitezica[1], I. Aguilar[3], I. Misztal[4]**
[1]INRA, UMR1388, Toulouse, France, [2]University of Aarhus, Foulum, Denmark, [3]INIA, Las Brujas, Uruguay,
[4]University of Georgia, Athens, Georgia, USA.

**ABSTRACT:** A theory to account for across-founder ancestral relationships is presented, which can consider relationships within and across populations. The theory assumes finite size of the ancestral population. Ancestral relationships can be represented as metafounders, pseudo-individuals that can be seen as pool of gametes and as extensions of random genetic groups. Simple rules exist for computation of relationships, inbreeding, and the inverse of the relationship matrix. Values of the Ancestral relationships can be inferred from marker relationships and pedigrees. Once they are computed, use in Single Step GBLUP is straightforward and compatibility of marker and pedigree information is automatic.
**Keywords:** Relationships; BLUP; Genomic; Pedigree; Effective size

## Introduction

Markers reveal relationships across populations with no common pedigree. For instance, VanRaden et al. (2011) and Legarra et al. (2014) quantified relationships between "unrelated" breeds of cattle and sheep, respectively. Markers also reveal relationships across founders of a population. These relationships exist due to the finite size of the population and of the genome. The more we use markers for genomic evaluation, the worse the hypothesis of "unrelatedness" becomes. This makes comparison across pedigree and genomic relationships awkward and has relevance for some applications, in particular for Single Step methods. In addition, genomic relationships do not depend on pedigree length, whereas pedigree relationships do.

Jacquard (1974) presented relationship matrices allowing for across-founder relationships due to finite size of the population. VanRaden (1992) presented algorithms to compute inbreeding where animals with missing parents had non-zero inbreeding. VanRaden et al. (2011) used across-breeds relationships based on markers. Christensen (2012) presented a method in which across-founder relationships were constructed and estimated from marker data in a subset of individuals and a pedigree, in a Single Step context. Sullivan and Schaeffer (1994) suggested the consideration of unknown parent groups as random. The purpose of this work is to give a general and coherent framework for the consideration of relationships across founders *within* population and across founders *across* population through the use of *metafounders*, which are pseudo-individuals that can be seen as pools of ancestral gametes. Use of metafounders simplifies the consideration of ancestral relationships, extends naturally the notion of unknown parent groups, simplifies the construction of relationship matrices for single or crossed populations, and leads naturally to easy algorithms for inversion of the relationship matrix. Compatibility of genomic and pedigree relationships is warranted if ancestral relationships are inferred from markers.

## Theory

**Ancestral relationships.** Assume that the founders of a single population are drawn with replacement from a larger, but limited, population with $2N_e$ gametes (Figure 1). This generates self- and across- relationships of, respectively, $1+\gamma/2$ and $\gamma$, where the ancestral relationship $\gamma = 1/N_e$. Several, possibly overlapping, populations can be equally considered (Figure 2), with an extended set of ancestral relationships $\boldsymbol{\Gamma}$.

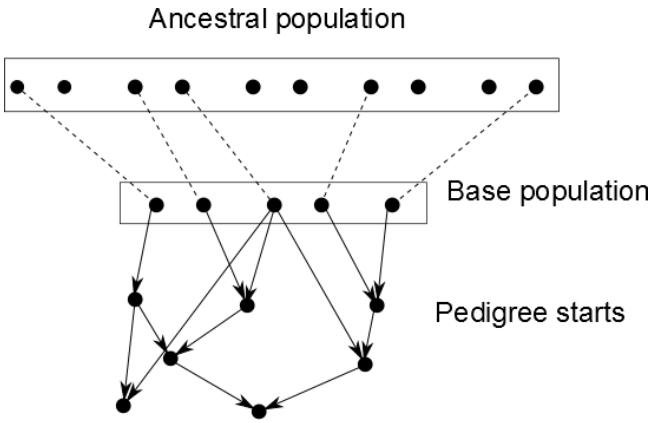


**Figure 1: Ancestral and base population and pedigree**

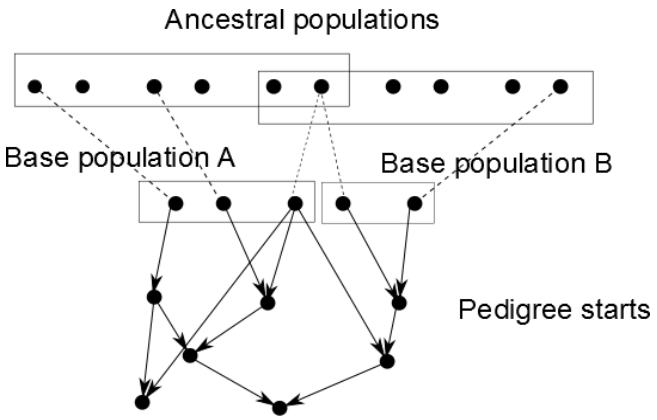


**Figure 2: Two populations**

**Metafounders.** Instead of setting up across-individuals relationships using $\gamma$ coefficients, one can define metafounders. These are pseudo-individuals that represent pools of gametes (Figure 3). They have a self-relationship of $\gamma$ and an inbreeding coefficient of $\gamma - 1$. The interpretation of $\gamma - 1$ is the heterozygosity of the pool of gametes. Metafounders can also be seen as an extension of unknown parent groups, where these unknown parent groups now would be random and contain information on inbreeding of their "offspring" – something that is absent from the original formulation but in agreement with VanRaden (1992) idea for inbreeding with missing parentships. Also, if the ancestral population is considered to be infinite, $\gamma = 0$ and the usual structure of the population is found.

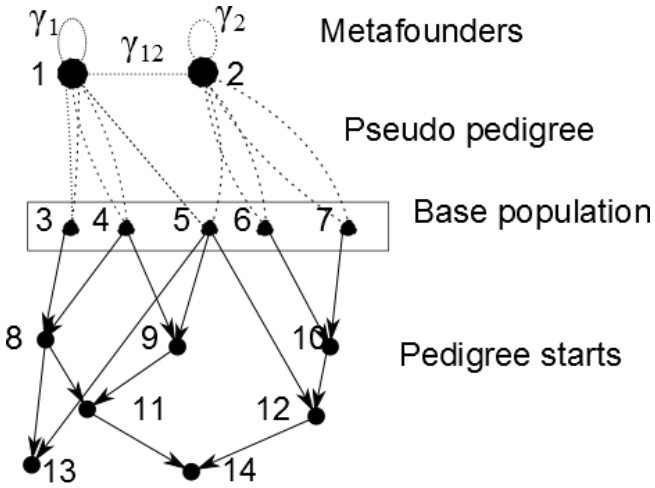


**Figure 3: Two populations and metafounders**

Using metafounders one can create pseudo-pedigrees (Figure 4). Using pseudo-pedigrees, it is possible to use the usual recursion rules, and more importantly, Henderson's rules (with very minor modifications) for decomposition and sparse inversion of A. In short, Henderson's rules do not change, with the proviso that Mendelian sampling variances need to be computed previously using the typical rule $D_{ii} = 0.5 - 0.25(F_s + F_d)$ where if for instance the sire is a metafounder then $F_s = \gamma - 1$. Note that the algorithm works even for $= 0$ , which is the regular case.

1 0 0
2 0 0
3 1 1
4 1 1
5 1 2
…
14 11 12

**Figure 4: Pedigree file for Figure 3**

**Inferring ancestral relationships.** Ancestral relationships can be inferred by setting as the reference point allelic frequencies of 0.5 for all markers. This amounts to marginalize the likelihood over the unknown distribution of allele frequencies (Christensen (2012)). The likelihood of observed genotypes is:

$$\log p(\boldsymbol{m}|\boldsymbol{\Gamma}, s) = const - \frac{pn_2}{2}\log(s) - \frac{p}{2}\log\left(\left|\boldsymbol{A}_{22}^{\boldsymbol{\Gamma}}\right|\right) - \frac{p}{2s}tr\left(\boldsymbol{A}_{22}^{\boldsymbol{\Gamma}-\mathbf{1}}\boldsymbol{Z}\boldsymbol{Z}'\right)$$

Where **m** refers to the marker genotypes, $p$ contains the allelic frequencies, $n_2$ denotes the number of genoyped animals, $\boldsymbol{Z}$ is a matrix with genotypes coded as {-1,0,1}, and $\boldsymbol{A}_{22}^{\boldsymbol{\Gamma}}$ is the relationship matrix across genotyped animals including ancestral relationships. Inferring **Γ**and $s$ from this likelihood is not straightforward. An alternative is to use a method of moments, equating average pedigree and genomic relationships and average pedigree and genomic inbreeding (Vitezica et al. (2011); Christensen et al. (2012)). Consider for instance two populations A and B. For large enough populations **Γ**and $s$ can be expressed as:

$$s = \frac{trace\begin{pmatrix}\boldsymbol{Z}_A\boldsymbol{Z}_A' & \\ & \boldsymbol{Z}_B\boldsymbol{Z}_B'\end{pmatrix}}{\frac{n_A\gamma^{A,A}}{2} + \frac{n_B\gamma^{B,B}}{2} + trace\begin{pmatrix}\boldsymbol{A}_{A,A} & \\ & \boldsymbol{A}_{B,B}\end{pmatrix}}$$

$$\gamma^A = \left(\overline{\boldsymbol{G}}_{A,A} - \overline{\boldsymbol{A}}_{A,A}\right);\ \gamma^B = \left(\overline{\boldsymbol{G}}_{B,B} - \overline{\boldsymbol{A}}_{B,B}\right);$$

$$\gamma^{A,B} = \overline{\boldsymbol{G}}_{A,B}$$

where

$$\boldsymbol{G} = \begin{pmatrix}\boldsymbol{G}_{A,A} & \boldsymbol{G}_{B,A} \\ \boldsymbol{G}_{B,A} & \boldsymbol{G}_{B,B}\end{pmatrix} = \begin{pmatrix}\boldsymbol{Z}_A\boldsymbol{Z}_A' & \boldsymbol{Z}_A\boldsymbol{Z}_B' \\ \boldsymbol{Z}_B\boldsymbol{Z}_A' & \boldsymbol{Z}_B\boldsymbol{Z}_B'\end{pmatrix}/s$$

The procedure is very similar to VanRaden et al. (2011). Note that this proceeds by fitting **A** (pedigree relationships) to the genetic base of **G** (genomic relationships) and not the opposite. This makes sense because genomic relationships are free from missing pedigree, pedigree errors or pedigree depth.

**Use in the Single Step**

The final aim of this theory is to refine the framework for the Single Step GBLUP (Aguilar et al. (2010); Christensen and Lund (2010)). In practice, the theory in this paper can be used as follows. First, infer ancestral relationships **Γ** and scaling factor $s$ based on existing pedigree and markers. Construct $\boldsymbol{A}^{\boldsymbol{\Gamma}-\mathbf{1}}$ , $\boldsymbol{A}_{22}^{\boldsymbol{\Gamma}-\mathbf{1}}$ and $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{Z}'/s$, and combine them in

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

Compatibility of relationships is automatic by use of **Γ**and marginalization over unknown allelic frequencies

(Christensen (2012)). Note that tuning for crossbreds is also automatic, whereas this tuning is otherwise difficult (Harris and Johnson (2010)). It can also be shown that inclusion of ancestral relationships increases the range of values of $\boldsymbol{A}_{22}^{\boldsymbol{\Gamma}}$, and therefore decreases the values of $\boldsymbol{A}_{22}^{\boldsymbol{\Gamma}-1}$.

This explains why, in practical applications, Aguilar et al. (2010) and Tsuruta et al. (2011, 2013) found more accurate and less biased valuations using a weight $\omega < 1$ on $\boldsymbol{A}_{22}^{-1}$.

## Example

Consider the pedigree in Figure 4. Assuming unrelated founders, relationships between individuals 8 (pure breed 1), 10 (pure breed 2) and 14 (crossbred, 56% breed 1 and 44% breed 2, grandson of 8 and of 10) are $A_{subset} = \begin{pmatrix} 1 & 0 & 0.313 \\ 0 & 1 & 0.25 \\ 0.313 & 0.25 & 1.063 \end{pmatrix}$. If we consider within and across-base population relationships in $\boldsymbol{\Gamma} = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.2 \end{pmatrix}$ we obtain:

$$A_{subset} = \begin{pmatrix} 1.05 & 0.05 & 0.375 \\ 0.05 & 1.10 & 0.341 \\ 0.375 & 0.341 & 1.095 \end{pmatrix}$$

where the relationship between 8 and 10 appears, which in turn slightly increases the inbreeding coefficient of 14.

## Conclusion

We have sketched a coherent theory to consider ancestral relationships in pedigreed populations, how to use them efficiently, and how these can be estimated from genotypic data. The notion of metafounder condensates the information of ancestral population and allows for simple algorithms. This work lays the foundations for the genomic analysis of complex populations, possibly with crosses and missing parentships. Testing in real data sets is needed.


## Acknowledgements

This project has been financed by X-Gen and GenSSeq actions from SelGen metaprogram (INRA). We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing resources.


## Literature Cited


Aguilar, I., Misztal, I., Johnson, D.L. et al. (2010). J Dairy Sci 93:743-752

Christensen, O.F. (2012). Gen Sel Evol 44:37

Christensen, O.F., Lund, M.S. (2010). Gen Sel Evol 42:2

Christensen, O., Madsen, P., Nielsen, B. et al. (2012) Animal 6:1565-1571

Jacquard, A. (1974).The Genetic Structure of Populations. Springer-Verlag.

Legarra, A., Baloche, G., Barillet, et al. (2014). J. Dairy Sci. (Accepted)

Sullivan, P. G., Schaeffer, L. R. (1994). In Proc. 6th World Congr. Genet. Appl. to Lives. Prod 18: 483-486.

Tsuruta, S., Misztal, I., Aguilar, I. et al. (2011). J. Dairy Sci. 94 :4198–4204.

Tsuruta, S., Misztal, I., Lawlor, I. (2013). J. Dairy Sci. 96:1–4.

VanRaden, P.M., Olson, K.M., Wiggans, G.R.. et al. (2011). J. Dairy Sci., 94:5673-5682.

VanRaden, P.M. (1992). J. Dairy Sci. 75:3136-3144.

Vitezica, Z., Aguilar, I., Misztal, I. et al (2011). Gen. Res. 93:357-366.