Statistical Indicators
E-38
Pseudo-records

▪ Introduction

In December 2014 GES introduced a new method of breeding value estimation: the pseudo-record system (the PSR system). In this system all information (parents, offspring, own performance and genomic) is incorporated into a single evaluation to produce *genomically enhanced breeding values* (GEBV).

The pseudo-record system is a system of breeding values estimation in which the genomic breeding values (*direct genomic values* or DGV) are used as a fourth source of information, after data on own performance, parents and offspring. The DGV of genotyped animals are treated as observations on a virtual trait (the pseudo-trait) that is correlated to the actual trait. A DGV thus becomes an observation on the DNA of an animal for this pseudo-trait. This observation is called the pseudo-record of an animal. Since we use DGV as observations on the pseudo-trait, the pseudo-record of an animal is its DGV (Stoop et al., 2014).

In the breeding value estimation pseudo-traits can be treated the same as actual traits, resulting in a single evaluation in which all traits are evaluated simultaneously. The correlation between pseudo-trait and actual trait results in a natural integration of genomic data with conventional data (by which we mean data from parents, offspring and own performance). The resulting breeding values for actual traits are therefore GEBV: genomically enhanced breeding values.

For example, genotyped bulls have DGV for the trait ‘milk-production’ that are used as pseudo-records: observations of own performances (its own DNA) for the trait ‘pseudo-milk-production’. A young bull, genotyped but without daughters, will have more information available than just the average of its parents. The DGV, through the correlation between the trait ‘milk production’ and the ‘pseudo-milk-production’ trait, will produce a breeding value for ‘milk production’ that is made up of its parent average and information we have about its DNA, in the form of its DGV.

This method has a number of advantages over the old method of integration, *blending*, the three most important of which are:

1. Simplification of the integration process
2. More efficient use of genomic data
3. Reduction of bias as a result of genomic pre-selection

Ad 1) When *blending* is used, the breeding value estimation encompasses two separate processes, the conventional breeding value estimation and the genomic evaluation, that are combined in a separate post-processing step, the actual blending of breeding values. With pseudo-records a separate integration step is no longer required, as the genomic data (DGV) is directly integrated with conventional data to produce GEBV directly. In Figure 1 the differences between *blending* and *pseudo-records* are presented in a schematic way.

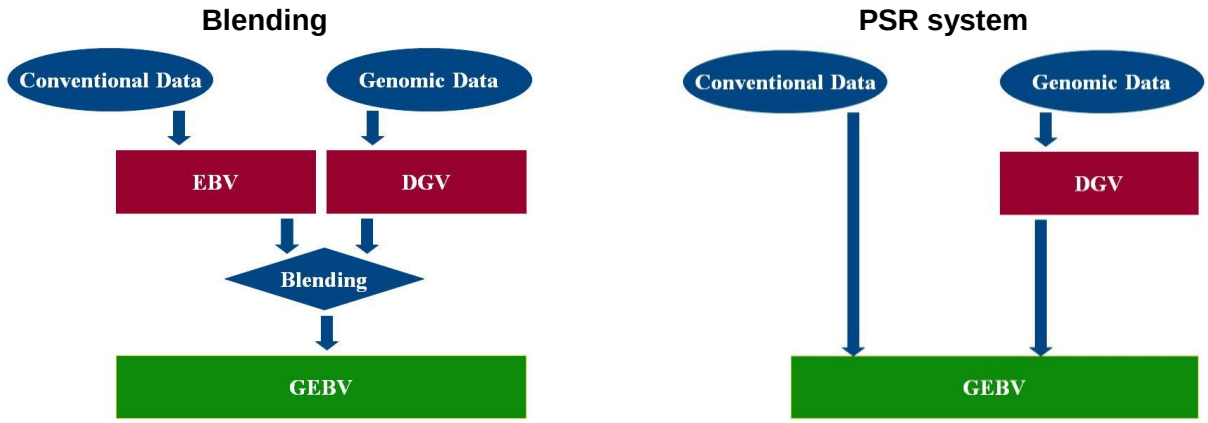

Figure 1. The integration of genomic and conventional data when using blending (left) and in the method using pseudo-records (right) in schematic form. When **blending** is used, the **GEBV** estimation encompasses two separate processes, the conventional **EBV** (based on **conventional data**) and the genomic evaluation (**DGV** based on **genomic data**), that are combined in a separate post-processing step, the actual **blending** of breeding values. With pseudo-records a separate integration step is no longer required, as the **genomic data** (**DGV**) is directly integrated with **conventional data** to produce **GEBV** directly.

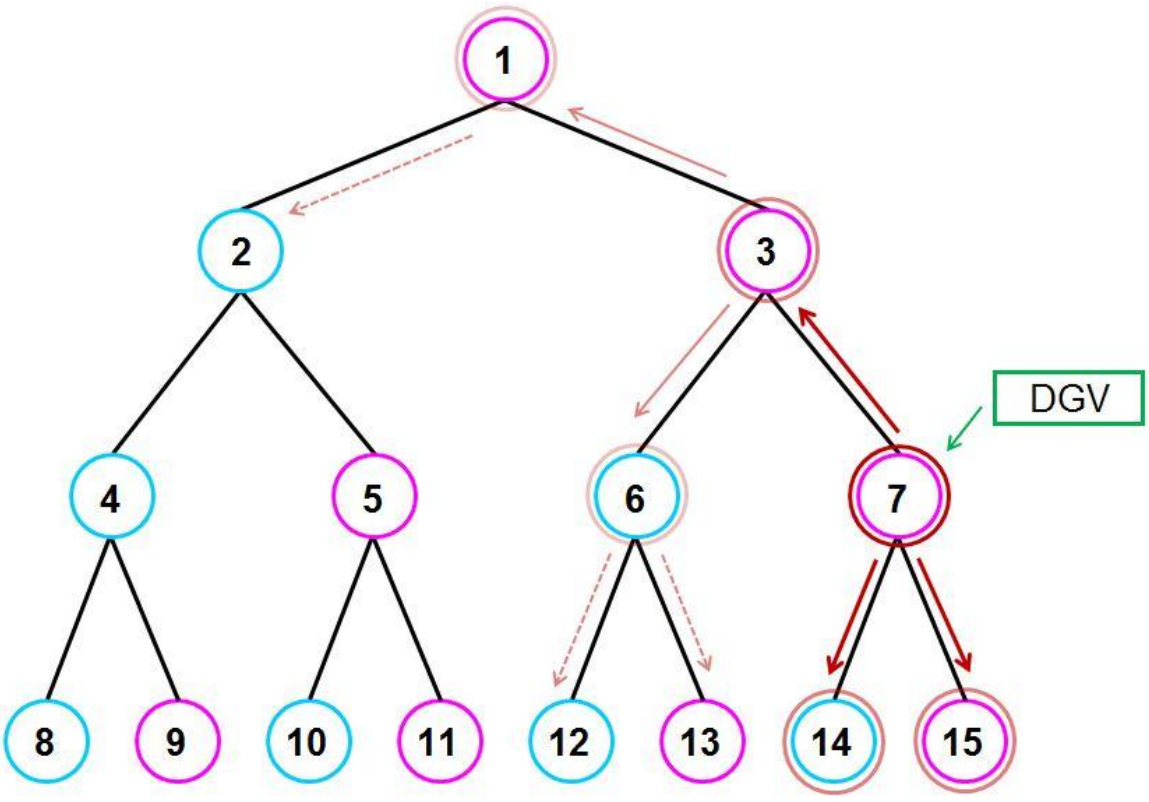

Figure 2. Illustration of the effect of one DGV in the PSR system on the pedigree of a genotyped animal. Sire 7 has been genotyped and has a DGV used as input in the PSR system. This has increased the reliability of its breeding value, indicated with the red circle, because it is now GEBV. However, some of that increase (roughly ¼ of it) is passed on to its offspring (animals 14 and 15) and its parents (sire 3, in this case), indicated with the lighter shade circles. Hence, these breeding values also are GEBV. A quarter of the increase in reliability for sire 3 passes on further up and down the pedigree, diminishing with every step. This effect vanishes quite quickly: the increase in reliability in animals 12 and 13 is only about 1/64 of the increase in sire 7.

Ad 2) If an animal has a valid DGV its GEBV will be more reliable than its conventional breeding value. In the PSR system, this increase in reliability is no longer restricted to the animal itself. Its parents and siblings (if not genotyped) will also benefit from through the structure of the pedigree. This is illustrated in Figure 2.

Ad 3) Genomic selection can cause a bias in genetic evaluations if mean genetic potential of the group of genotyped animal (those with a DGV) differs from the mean genetic potential of group of animals without a DGV. This may happen because the *Mendelian sampling* term is more accurately estimated in genotyped animals then it is in conventionally evaluated animals, or because of non-random mating of animals, based on genomic evaluations. In general we can say that a bias occurs as soon as selection takes place based on information that is not accounted for in the genetic evaluation. By incorporating the DGV of genotyped animals in the genetic evaluation of the entire population, the occurrence of bias is avoided.

▪ Principles of pseudo-records

The basic idea in the PSR system is that the DGV of a genotyped animal is used as an observation of a pseudo-trait, the pseudo-record. When pseudo-records are available for a particular trait, a pseudo-trait is used in a multiple-trait analysis, where it is correlated to the actual trait (Mäntysaari en Strandèn; 2010). When this is done correctly, the pseudo-record can be treated as an actual observation on the pseudo-trait. The pseudo-trait is then analyzed like a normal trait in the breeding value estimation.

In formula form the actual trait – pseudo-trait system looks like:

$$\begin{bmatrix} Y \\ D \end{bmatrix}_i = \begin{bmatrix} a_Y \\ a_D \end{bmatrix}_i + \begin{bmatrix} e_Y \\ e_D \end{bmatrix}_i$$

and

$$\text{var}\begin{bmatrix} a_Y \\ a_D \end{bmatrix} = \begin{bmatrix} \sigma_g^2 & r_g\sigma_g^2 \\ & \sigma_g^2 \end{bmatrix}$$

Where Y and D are observations on the actual and the pseudo-trait. The variables a_Y and a_D are the breeding values for Y and D. In the current context a_Y corresponds to the GEBV and a_P to the DGV of an animal. Finally, e_Y and e_D are the residuals (estimation-errors) for Y and D.

The second formula describes the genetic relation between Y and D, where σ_g is the genetic variance. Because the DGV is a breeding value the genetic variance of the pseudo-trait D is equal to the genetic variance of actual trait Y.

The integration of genomic data with conventional data takes place in the second formula above. This formula takes the genomic breeding value (a_D) and uses the r_g (in the upper right corner of the right-hand side) to integrate the information of a_D into a_Y (see Figure 3). The genetic correlation of the pseudo-trait with the actual trait (r_g) is proportional to the reliability of the genomic data in the DGV. Hence, the amount of information from the DGV (which is a_D) into de breeding value a_Y (the GEBV) is limited to the accuracy with which the DGV was estimated.

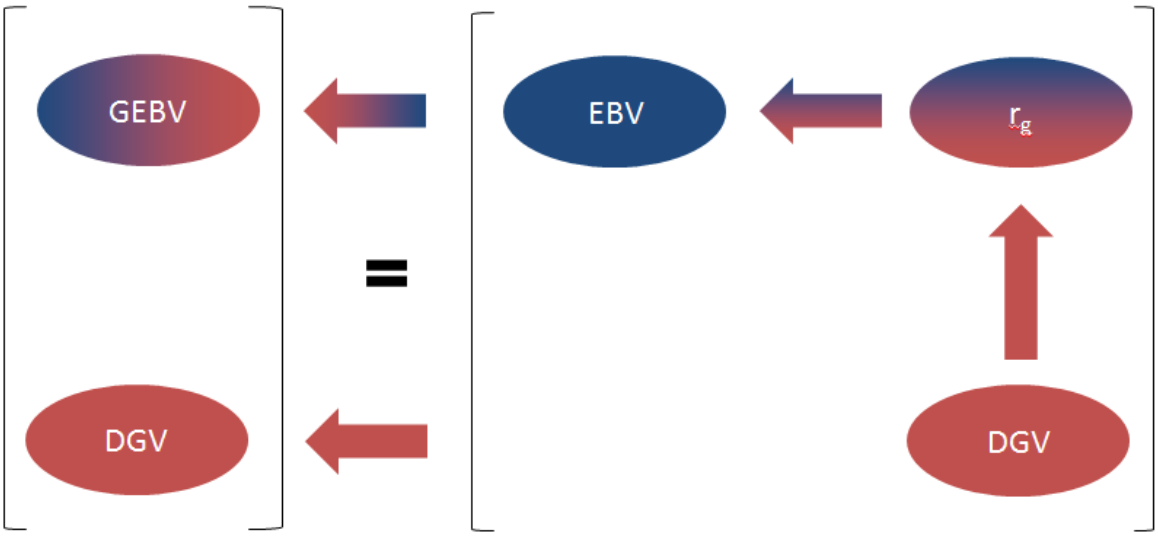

Figure 3. A schematic representation of the way genomic data (DGV) influence the breeding value of an animal to become a GEBV in the system of formulas described above. On the right side of the figure, the genetic correlation r_g between a trait and a correlated genomic trait passes information from the DGV into the conventional breeding value EBV, such that the latter becomes a GEBV in the end-result on the left. Note that because a heritability of 1 is assumed for correlated genomic traits, the DGV is the same on the right and on the left.

The pseudo-trait is assumed to have a heritability of (nearly) one. Therefore, breeding value a_D of a genotyped animal will be equal to the DGV of an animal. Because the heritability ≈ 1, related animals (parents or offspring) with a DGV will not influence the breeding value a_D, avoiding double-counting of genomic information and overestimation of GEBV reliability.

▪ Selection of data

The observations on pseudo-traits in the PSR system are the DGV from the genomic evaluation. Animals with DGV will have a corresponding pseudo-record in the PSR system, when:

- The animal is a bull and
  - 10 months or older and owned by an AI organization participating in the genomic evaluation
  - is a Eurogenomics animal
  - is not an AI bull and has been culled
- Or when the animal is a female

Currently, AI organizations participating in the Dutch national genomic evaluations are CRV and KI Kampen.

Eurogenomics is a European network of AI companies performing genomic evaluations. To more accurately estimate DGV and widen the scope of genomic evaluations, the participants in Eurogenomics have agreed to exchange genotype information of their sires for use in genomic evaluations in the participants respective countries. In practice this means that a Eurogenomics bull with a genotype will have a Dutch/Flemish DGV used in the PSR system.

▪ Determination of parameters

For each pseudo-trait the conventional breeding value estimation must be extended to include genetic and residual variance components modelling the link between pseudo-trait and actual trait. To do this the (genetic) correlation between a pseudo-trait and the associated actual trait is needed. This genetic correlation is equal to the square root of the reliability of the marker effects r^2_{mark} (Stoop et al., 2014). Hence the basic PSR system described above changes to:

$$\begin{bmatrix} Y \\ D \end{bmatrix}_i = \begin{bmatrix} a_Y \\ a_D \end{bmatrix}_i + \begin{bmatrix} e_Y \\ e_D \end{bmatrix}_i$$

$$\text{var}\begin{bmatrix} a_Y \\ a_D \end{bmatrix} = \begin{bmatrix} \sigma^2_g & \sqrt{r^2_{mark}} \cdot \sigma^2_g \\ & \sigma^2_g \end{bmatrix}$$

Note that r_g is replaced by $\sqrt{(r^2_{\text{mark}})}$ in the above. In the PSR system a pseudo-trait is allowed to only directly influence the actual trait it corresponds to. Hence, in a multiple-trait breeding values estimation, the covariance between a pseudo-trait and all other, non-corresponding traits are proportional to the correlation between pseudo-trait and corresponding actual trait. For a more in-depth, technical account on how to derive correct parameter matrices for use in a breeding value estimation when multiple (pseudo-) traits are involved, see Appendix A.

▪ Method of GEBV estimation

Due to the nature of the PSR system, it can only be applied to breeding value estimations where breeding values are estimated using a system of linear equations (generalized linear models). Currently the breeding value estimations that include the PSR system are:

- Production
- Conformation
- Fertility
- Udder health
- Claw health
- Milking speed & Temperament
- Calving ease
- Liveability
- Beef production

Pseudo-traits are analysed the same in all breeding value estimations where the PSR system is applied. The statistical model generally used to analyse pseudo-traits is:

$$y_{pi} = \mu_p + animal_{pi} + error_{pi}$$

Where

μ : the mean effect
y : the DGV for psr-trait p and the i-th
animal : the genetic effect
error : the residual effect (estimation error)

Because the $h^2 \approx 1$, the effect *animal* will be equal to the pseudo-record, which is the DGV.

Because the estimation of DGV is dependent on accurate and timely conventional breeding values, the genetic evaluation with the PSR system takes place at the end of the genetic evaluation process. The genetic evaluation process is done in the following order:

1) Conventional breeding value estimation NL/FL
2) DGV estimation based on breeding values from 1) and previous MACE run
3) Breeding value estimation NL/FL using the PSR system

This ensures that the PSR evaluation starts with the most up-to-date DGV as pseudo-observations.

Table 1 gives an overview of all traits for which DGV (and hence pseudo-records) are available. The table shows per trait the reliability of the marker data and the genetic correlation between pseudo-trait and actual trait. Note that for a number of breeding value estimation an overall index is missing as a pseudo-trait (Overall conformation, Udder health index). This is because they are indices of underlying traits for which pseudo-records exist. A pseudo-trait corresponding to the index does not add any new information and is hence omitted. The actual overall indices are formed from underlying traits that are GEBV and hence are GEBV themselves, including increased reliability.

▪ Reliability of GEBV

Relative to conventional breeding values the reliability of the GEBV is increased, especially for young bulls with no or few daughters. See Figure 3, for instance, where the reliability of the GEBV for Udder health (in blue) is given, relative to the reliability of the conventional breeding value (red line) for a number of bulls. For bulls that already have a reliable conventional breeding value (around 70%) the distance between the red line and the blue ‘x’-signs is modest. For young bulls without daughters the increase in reliability is most dramatic.

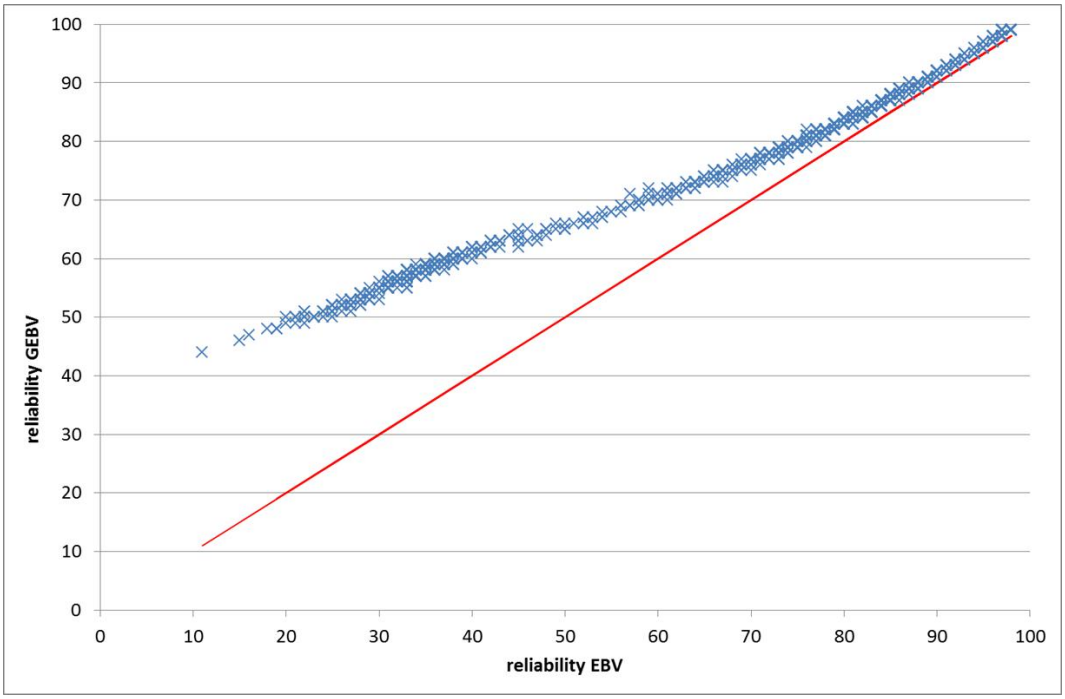

Figure 3. Reliability of GEBV for Udder health versus the reliability of the conventional breeding value (EBV) for genotyped bulls. GEBV reliabilities in blue “x”, the solid red line indicates EBV reliabilities.

Table 1. Overview of all traits for which pseudo-records are available, their heritability (h^2) the reliability of the marker data (r^2_{mark}) and the genetic correlation between pseudo-trait and actual trait (r_g). Values given were used in the December 2014 evaluation

BVE	Trait	h^2	r^2_{mark}	r_g
Production	milk yield	0,57	0,688	0,829
	fat yield	0,58	0,639	0,799
	prot yield	0,50	0,605	0,778
Conformation	stature	0,52	0,686	0,828
	chest width	0,24	0,641	0,801
	body depth	0,31	0,588	0,767
	angularity	0,11	0,620	0,787
	body condition	0,30	0,653	0,808
	rump angle	0,34	0,642	0,801
	rump width	0,40	0,669	0,818
	rear legs rear	0,15	0,465	0,682
	rear legs side	0,23	0,639	0,799
	foot angle	0,14	0,610	0,781
	locomotion	0,14	0,482	0,694
	fore udder attachment	0,27	0,600	0,775
	front teat placement	0,38	0,653	0,808
	front teat length	0,38	0,659	0,812
	udder depth	0,38	0,713	0,844
	rear udder height	0,23	0,602	0,776
	udder support	0,23	0,661	0,813
	rear teat placement	0,32	0,662	0,814
	frame	0,28	0,621	0,788
	dairy strength	0,14	0,266	0,516
	overall udder score	0,29	0,623	0,790
	overall feet leg	0,16	0,503	0,709
Udder health	somatic cell count	0,37	0,632	0,795
	subcl. mastitis	0,06	0,644	0,803
	clin. mastitis	0,06	0,569	0,754
Calving ease	direct calving ease	0,07	0,717	0,847
	maternal calving ease	0,05	0,384	0,620
	direct stillbirth	0,04	0,326	0,571
	maternal stillbirth	0,09	0,643	0,802
Fertility	non return 56	0,04	0,485	0,696
	interval calving – 1st insemination	0,17	0,625	0,791
	calving interval	0,15	0,624	0,790
	int first last ins	0,08	0,635	0,797
MS&T	milking speed	0,23	0,587	0,766
	temperament	0,12	0,482	0,694
Claw health	claw health index	0,18	0,317	0,563
Beef	beef index	0,25	0,564	0,751

▪ References

Mäntysaari, E.A. & Strandén, I. (2010) Proc. 9th WCGALP, Leipzig, Germany.

Stoop W.M., H. Eding, M.L. van Pelt, L.C.M. de Haer and G. de Jong, (2014) Proc. 10th WCGALP, Vancouver, Canada

▪ Appendix A

This appendix is a more detailed, technical account on how to derived parameter matrices for use in a breeding value estimation including the PSR system. The first part explains the derivation of matrices when multiple pseudo-traits are involved. In the second part pseudo-traits that are composite traits (indices) correlating to a number of underlying actual traits are dealt with.

Principle formula of correlated genomic traits

The basic idea for including pseudo-records into a national genetic evaluation as formulated by Mantysaari and Strandèn (2010) is of the following form:

$$\begin{bmatrix} Y \\ P \end{bmatrix}_i = \begin{bmatrix} Z_Y & 0 \\ 0 & Z_P \end{bmatrix} \begin{bmatrix} a_Y \\ a_P \end{bmatrix}_i + \begin{bmatrix} e_Y \\ e_P \end{bmatrix}_i$$

$$\text{var}\begin{bmatrix} a_Y \\ a_P \end{bmatrix} = \mathbf{G} = \begin{bmatrix} \sigma^2_{g,Y} & r_g\sigma^2_{g,Y} \\ & \sigma^2_{g,Y} \end{bmatrix}$$

$$\text{var}\begin{bmatrix} e_Y \\ e_P \end{bmatrix} = \mathbf{E} = \begin{bmatrix} \sigma^2_{e,Y} & 0 \\ & (1-h^2)\sigma^2_{g,Y} \end{bmatrix}$$

Where Y and P are the actual and pseudo-trait, a_Y and a_P are the breeding values for Y and P and e_Y and e_P are residuals of Y and P. Z_Y and Z_P are incidence matrices linking observations to animals.

Conventional and genomic breeding value information is linked through a genetic correlation r_g which is the square root of the reliability of the DGV (minus pedigree information. Usually this is taken to be the increase in in reliability in terms of *expected daughter contributions* of DGV relative to the reliability of pedigree information, i.e. parent averages or sire indices. The heritability h^2 is assumed to be (nearly) one, hence the residual variance of correlated genomic traits is a small fraction ($1 - h^2$) of the genetic variance.

Basic structure of genetic covariance matrix

Let $\mathbf{G_c}$ be a genetic correlation (or covariance) matrix of N conventional traits. Let $\mathbf{G_{22}}$ be the partition of $\mathbf{G_c}$ containing traits for which pseudo-records exist. Hence, the (conventional) genetic correlation (covariance) matrix is:

$$\mathbf{G}_c = \begin{bmatrix} \mathbf{G_{11}} & \mathbf{G_{12}} \\ \mathbf{G_{21}} & \mathbf{G_{22}} \end{bmatrix}$$

Finally, let $\mathbf{r^2}$ be a vector reliability of the marker effects r^2_{mark} of *P* pseudo-traits obtained from a validation study. To obtain correlations between pseudo-trait and corresponding existing trait we transform $\mathbf{r^2}$ into a diagonal matrix **R**, such that $R_{i,i} = \sqrt{(r^2_i)}$ and $R_{i,j\neq i} = 0$.

$$\mathbf{R} = \begin{bmatrix} \sqrt{r_1^2} & 0 & 0 & 0 \\ & \sqrt{r_2^2} & 0 & 0 \\ & & \ddots & 0 \\ & & & \sqrt{r_P^2} \end{bmatrix}$$

We assume $h^2 = 1.0$ for the DGV, hence the genetic covariance structure among pseudo-traits is equal to the genetic covariance structure among the associated actual traits $\mathbf{G_{22}}$. Then the correct genetic correlation matrix is obtained through:

$$\mathbf{G} = \begin{bmatrix} \mathbf{G_c} & \mathbf{G_{c,psr}} \\ & \mathbf{G_{psr}} \end{bmatrix} = \begin{bmatrix} \mathbf{G_{11}} & \mathbf{G_{12}} & \mathbf{G_{12}R} \\ & \mathbf{G_{22}} & \mathbf{G_{22}R} \\ & & \mathbf{G_{22}} \end{bmatrix}$$

This structure ensures that traits in $\mathbf{G_{psr}}$ only affect their counterparts in $\mathbf{G_{22}}$ with no direct influence on any other trait.

*Deriving **G** for multiple pseudo-traits*

The structure of the **G** matrix described above is easily obtained using a matrix **Φ** such that

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{I_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I_2} \\ \mathbf{0} & \mathbf{W} \end{bmatrix}$$

Where $\mathbf{I_1}$ and $\mathbf{I_2}$ are identity matrices corresponding to $\mathbf{G_{11}}$ and $\mathbf{G_{22}}$ and **W** is a matrix describing the relation between pseudo-trait and corresponding existing trait. If each trait in $\mathbf{G_{22}}$ has a single counterpart pseudo-trait, matrix $\mathbf{W} = \mathbf{I_2}$.

Using this matrix we produce:

$$\mathbf{G^*} = \mathbf{\Phi G \Phi'}$$

The resulting matrix **G*** is of size *N+P,* i.e. covariances of pseudo-traits are now included. However, in **G*** the correlation between pseudo-trait and existing trait is not yet accounted for. To obtain the correct **G**, we partition **G***, separating conventional and pseudo-traits and multiply off-diagonal partitions by **R** such that:

$$\mathbf{G_{psr}} = \begin{bmatrix} \mathbf{G^*_{cc}} & \mathbf{G^*_{cp}R} \\ & \mathbf{G^*_{pp}} \end{bmatrix}$$

Where the subscript **c** indicates existing conventional traits, while the subscript **p** indicates pseudo-traits.

*Deriving **G** for composite traits and indices*

The method described above is easily extended to include pseudo-traits that are linear combinations of existing traits (for instance, when the pseudo-trait is an overall trait, while the existing conventional traits are lactation specific). To do this the matrix **W** describing the relation between pseudo-trait and corresponding existing trait should contain weights in the linear combination.

Suppose we have two overall traits, for which we have pseudo-records, which are indices with two underlying traits each. The matrix **W** then takes the form:

$$\mathbf{W} = \begin{bmatrix} w_{11} & 0 & w_{12} & 0 \\ 0 & w_{21} & 0 & w_{22} \end{bmatrix}$$

Where w_{ij} is the *j*-th trait in index *i*. Using this in matrix **Φ** = [**I** ; **W**] in the previously described method to calculate **G*** and multiplying $\mathbf{G_{pc}}$ (and $\mathbf{G_{cp}}$!) with an appropriate **R** produces correct covariances between pseudo-traits, underlying traits and all other traits in **G**.

The residual covariance matrix

No residual covariances between pseudo-traits is assumed. And since pseudo-traits have a heritability $h^2 \approx 1$ the residual covariance matrix **E** is:

$$\mathbf{E} = \begin{bmatrix} \mathbf{E_c} & \mathbf{0} \\ & \mathbf{E_{psr}} \end{bmatrix} = \begin{bmatrix} \mathbf{E_c} & \mathbf{0} \\ & (1-h^2)\cdot diag(\mathbf{G_{pp}}) \end{bmatrix}$$

Where *diag*($\mathbf{G_{pp}}$) indicates a matrix with only the diagonal elements of $\mathbf{G_{pp}}$.