Properties of genomic relationships for estimating current genetic variances within and genetic correlations between populations

Yvonne C.J. Wientjes*, Piter Bijma*, Jérémie Vandenplas*, Mario P.L. Calus*

* Wageningen University & Research, Animal Breeding and Genomics, 6700 AH Wageningen, The Netherlands

15 Running title: Genetic correlation between populations

16

17 Key words: genetic correlation between populations, genomic relationships, genetic variance,

18 multi-trait model

19

20 Author information:

21 Yvonne Wientjes

22 Wageningen University & Research

23 Animal Breeding and Genomics

24 P.O. box 338, 6700 AH Wageningen, the Netherlands

25 E-mail: yvonne.wientjes@wur.nl

26 Phone: +31 317 481 904

27

28

29

ABSTRACT

Different methods are available to calculate multi-population genomic relationship matrices. Since those matrices differ in base population, it is anticipated that the method used to calculate the genomic relationship matrix affect the estimate of genetic variances, covariances and correlations. The aim of this paper is to define a multi-population genomic relationship matrix to estimate current genetic variances within and genetic correlations between populations. The genomic relationship matrix containing two populations consists of four blocks, one block for population 1, one block for population 2, and two blocks for relationships between the populations. It is known, based on literature, that current genetic variances are estimated when the current population is used as base population of the relationship matrix. In this paper, we theoretically derived the properties of the genomic relationship matrix to estimate genetic correlations and validated it using simulations. When the scaling factors of the genomic relationship matrix fulfill the property $k_{12} = \sqrt{k_1}\sqrt{k_2}$, the genetic correlation is estimated even though estimated variance components are not necessarily related to the current population. When this property is not met, the correlation based on estimated variance components should be multiplied by $\frac{\sqrt{k_1}\sqrt{k_2}}{k_{12}}$ to rescale the genetic correlation. In this study we present a genomic relationship matrix which directly results in current genetic variances as well as genetic correlations between populations.

INTRODUCTION

When estimating additive genetic values of individuals, the relationships between individuals are used to describe the covariance between additive genetic values for a specific trait. Those covariances between individuals are best represented by the relationships at causal loci. Since causal loci are generally unknown, different approaches have been developed to estimate relationships from genomic marker data (e.g., VanRaden 2008; Powell *et al.* 2010; Yang *et al.* 2010). As long as causal loci and genomic markers have the same properties, such as allele frequency distribution, relationships at the markers are observed to be an unbiased estimate of relationships at the causal loci (Yang *et al.* 2010; Yang *et al.* 2015).

Relationships are expressed relative to a base population, consisting of unrelated individuals that have average self-relationships of one, for which the additive genetic variance is estimated. The base population of a genomic relationship matrix depends on the method used to calculate the relationship matrix, therefore estimated variances differ across methods (Speed and Balding 2015; Legarra 2016). By using the current allele frequencies to calculate the genomic relationship matrix, the current population is the base population for which additive genetic variances are estimated (Hayes *et al.* 2009).

Genomic relationships can also be calculated between distantly related individuals, for example between individuals from different populations. Those relationships can be used to estimate genetic correlations between populations using a multi-trait model (Karoui *et al.* 2012), where the same trait in each population is modelled as a different trait. Due to differences in environments and allele frequencies, in combination with non-additive effects, the allele substitution effects of causal loci can differ between populations (e.g., Fisher 1918; Fisher 1930; Falconer 1952). Moreover, some causal loci might only segregate in one of the populations. Therefore, the genetic correlation between populations can differ from 1.

The genetic correlation between populations is an important parameter, since it is used to understand the genetic architecture and evolution of complex traits, such as disease traits in humans (De Candia *et al.* 2013; Brown *et al.* 2016). Moreover, the genetic correlation determines whether information can be shared across populations as done in multi-population genomic prediction (Wientjes *et al.* 2015; Wientjes *et al.* 2016), which is of importance for animals (e.g., Karoui *et al.* 2012; Olson *et al.* 2012), plants (e.g., Lehermeier *et al.* 2015) and humans (e.g., De Candia *et al.* 2013).

Different methods are available to calculate multi-population genomic relationship matrices (Harris and Johnson 2010; Erbe *et al.* 2012; Chen *et al.* 2013; Makgahlela *et al.* 2013). The two most important differences between the methods are: 1) the assumed relation between effect size and allele frequency of markers; namely assuming effect size and allele frequency are independent (e.g., method 1 of VanRaden (2008)) or assuming that markers with a lower allele frequency have a larger effect (e.g., method 2 of VanRaden (2008) and Yang (2010)), and 2) the allele frequency that is used; namely allele frequencies specific to each population, the average allele frequency across the populations, or the estimated allele frequency when the populations separated. Since relationships between individuals differ across those methods, it is anticipated that the method used to calculate the genomic relationship matrix affects the estimate of the genetic correlation.

Therefore, the aim of this paper is to define a multi-population genomic relationship matrix to estimate current genetic variances within and genetic correlations between populations. We theoretically derive a relationship matrix with this property and validate it with simulations. To rule out the effect of differences in linkage disequilibrium between markers and causal loci, we will focus in the entire paper on a situation where causal loci are used to calculate the relationships.

98 # MATERIALS AND METHODS

99 ## Theory

100 The additive genetic correlation, r_g, is the correlation between additive genetic values (A)
101 for two traits of the same individual (Bohren *et al.* 1966; Falconer and Mackay 1996). In an
102 additive model and under the assumptions that the correlation originates from pleiotropy,
103 genetic values are independent between loci, and allele substitution effects are independent
104 from allele frequency, r_g is equal to the average correlation between allele substitution effects
105 of the two traits, denoted as trait 1 and 2, at causal loci. This equality can be shown for
106 individual i by considering both genotypes (z) and allele substitution effects (α) at all n_c
107 causal loci as random:

108
$$Var(A_{i1}) = Var\left(\sum_j z_{ij}\alpha_{1j}\right) = E\left(\left(\sum_j z_{ij}\alpha_{1j}\right)\left(\sum_l z_{il}\alpha_{1l}\right)\right) = \sum_j E(z_{ij}z_{ij})E(\alpha_{1j}\alpha_{1j})$$
$$= n_c E(z_{ij}z_{ij})E(\alpha_{1j}\alpha_{1j})$$

109
$$Var(A_{i2}) = n_c E(z_{ij}z_{ij})E(\alpha_{2j}\alpha_{2j})$$

110
$$Cov(A_{i1}, A_{i2}) = Cov\left(\sum_j z_{ij}\alpha_{1j}, \sum_l z_{il}\alpha_{2l}\right) = E\left(\left(\sum_j z_{ij}\alpha_{1j}\right)\left(\sum_l z_{ij}\alpha_{2j}\right)\right) = \sum_j E(z_{ij}z_{ij})E(\alpha_{1j}\alpha_{2j})$$
$$= n_c E(z_{ij}z_{ij})E(\alpha_{1j}\alpha_{2j})$$

111
$$r_g = \frac{Cov(A_{i1}, A_{i2})}{\sqrt{Var(A_{i1})}\sqrt{Var(A_{i2})}} = \frac{n_c E(z_{ij}z_{ij})E(\alpha_{1j}\alpha_{2j})}{\sqrt{n_c E(z_{ij}z_{ij})\ E(\alpha_{1j}\alpha_{1j})}\sqrt{n_c E(z_{ij}z_{ij})\ E(\alpha_{2j}\alpha_{2j})}}$$
$$= \frac{E(\alpha_{1j}\alpha_{2j})}{\sqrt{E(\alpha_{1j}\alpha_{1j})}\sqrt{E(\alpha_{2j}\alpha_{2j})}} = \frac{\sigma_{\alpha_{12}}}{\sqrt{\sigma^2_{\alpha_1}}\sqrt{\sigma^2_{\alpha_2}}} = r_\alpha \qquad (1)$$

112 where j and l denote the different causal loci. Genotypes are represented by allele counts
113 coded as 0, 1 and 2 that are centered by subtracting $2p$, where p is the allele frequency for the
114 counted allele.

115 Similar to genetic correlations between traits in one population, the genetic correlation (r_g)
116 between populations can be estimated in a multi-trait model using a relationship matrix and

REML by modelling the phenotypes of two populations as different traits (Karoui *et al.* 2012). This approach is also known as multi-trait GREML. In the following, we will refer to trait 1 as the trait expressed in population 1 and to trait 2 as the trait expressed in population 2. When considering performance in different populations as different traits, individuals have a phenotype for only one trait. Therefore, the (co)variance structure of the additive genetic values can be written as (Visscher *et al.* 2014):

$$\begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}Var(\mathbf{a}_1) & Cov(\mathbf{a}_1,\mathbf{a}_2)\\ Cov(\mathbf{a}_2,\mathbf{a}_1) & Var(\mathbf{a}_2)\end{bmatrix}\right) = N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\mathbf{G}_{11}\sigma_1^2 & \mathbf{G}_{12}\sigma_{12}\\ \mathbf{G}_{21}\sigma_{12} & \mathbf{G}_{22}\sigma_2^2\end{bmatrix}\right). \quad (2)$$

where $\mathbf{a}_1$ is the vector with additive genetic values for individuals from population 1 for trait 1, $\mathbf{a}_2$ is the analogous vector for individuals from population 2 for trait 2, σ_1^2 and σ_2^2 are genetic variances for the two traits, σ_{12} is the genetic covariance between the traits, $\mathbf{G}_{11}$ is a matrix with genomic relationships within population 1, $\mathbf{G}_{22}$ is a matrix with genomic relationships within population 2, and $\mathbf{G}_{12}$ and $\mathbf{G}_{21}(=\mathbf{G}_{12}^{'})$ are matrices with genomic relationships between population 1 and 2.

To derive the definition of the genomic relationships in Equation 2, we derive the variances and covariance of the additive genetic values for the two traits. Naturally, this will result in an equation to calculate the genomic relationship matrix (**G**) across populations to estimate (co)variances in the current populations.

When both populations are in Hardy-Weinberg equilibrium, allele substitution effects are independent from allele frequency, and effects of causal loci are independent from each other, the genetic variance for trait 1 can be written as $\sigma_1^2 = \sum 2p_{1j}(1-p_{1j})\sigma_{\alpha_1}^2$, where p_{1j} is the allele frequency at locus j in population 1 (Falconer and Mackay 1996). Hence, the variance of $\mathbf{a}_1$ is:

$$Var(\mathbf{a}_1) = Var(\mathbf{Z}_1\boldsymbol{\alpha}_1) = \mathbf{Z}_1\mathbf{Z}_1^{'}\sigma_{\alpha_1}^2 = \frac{\mathbf{Z}_1\mathbf{Z}_1^{'}}{\sum 2p_{1j}(1-p_{1j})}\sigma_1^2, \quad (3)$$

where $\mathbf{Z}_1$ is a n_1 x n_c matrix of centered genotypes for all individuals from population 1 (n_1) for all causal loci, and $\boldsymbol{\alpha}_1$ is a vector of length n_c with allele substitution effects at causal loci for trait 1.

Similarly,

$$Var(\mathbf{a}_2) = \frac{\mathbf{Z}_2\mathbf{Z}_2'}{\sum 2p_{2j}(1-p_{2j})}\sigma_2^2 . \quad (4)$$

The genetic covariance between the two traits is:

$$\sigma_{12} = r_g\sqrt{\sigma_1^2}\sqrt{\sigma_2^2} = r_g\sqrt{\sum 2p_{1j}(1-p_{1j})\sigma_{\alpha_1}^2}\sqrt{\sum 2p_{2j}(1-p_{2j})\sigma_{\alpha_2}^2} =$$

$$= \sigma_{\alpha_{12}}\sqrt{\sum 2p_{1j}(1-p_{1j})}\sqrt{\sum 2p_{2j}(1-p_{2j})} . \quad (5)$$

Therefore, the covariance between genetic values of population 1 and 2 is:

$$Cov(\mathbf{a}_1,\mathbf{a}_2) = Cov(\mathbf{Z}_1\boldsymbol{\alpha}_1,\mathbf{Z}_2\boldsymbol{\alpha}_2) = \mathbf{Z}_1\mathbf{Z}_2'\sigma_{\alpha_{12}} = \frac{\mathbf{Z}_1\mathbf{Z}_2'}{\sqrt{\sum 2p_{1j}(1-p_{1j})}\sqrt{\sum 2p_{2j}(1-p_{2j})}}\sigma_{12} . \quad (6)$$

From Equation 3, 4 and 6, it follows that the genomic relationship matrix (**G**) is:

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix} = \begin{bmatrix} \frac{\mathbf{Z}_1\mathbf{Z}_1'}{\sum 2p_{1j}(1-p_{1j})} & \frac{\mathbf{Z}_1\mathbf{Z}_2'}{\sqrt{\sum 2p_{1j}(1-p_{1j})}\sqrt{\sum 2p_{2j}(1-p_{2j})}} \\ \frac{\mathbf{Z}_2\mathbf{Z}_1'}{\sqrt{\sum 2p_{1j}(1-p_{1j})}\sqrt{\sum 2p_{2j}(1-p_{2j})}} & \frac{\mathbf{Z}_2\mathbf{Z}_2'}{\sum 2p_{2j}(1-p_{2j})} \end{bmatrix} . \quad (7)$$

When allele frequencies from the current population are used, **G** from Equation 7 estimates current genetic (co)variances. Lourenco *et al.* (2016) presented a comparable **G** matrix for combining purebred and crossbred animals. Note that the covariance of the genotypes between the populations, $\mathbf{Z}_2\mathbf{Z}_1'$, is divided by the standard deviations of the genotypes in each population, $\sqrt{\sum 2p_{1j}(1-p_{1j})}$ and $\sqrt{\sum 2p_{2j}(1-p_{2j})}$. Therefore, the relationships in this **G** matrix are defined as correlations between the individuals.

By interpreting $\sum 2p_{1j}(1-p_{1j})$, $\sum 2p_{2j}(1-p_{2j})$ and $\sqrt{\sum 2p_{1j}(1-p_{1j})}\sqrt{\sum 2p_{2j}(1-p_{2j})}$ as scaling factors (i.e. k_1, k_2, and k_{12}) of **G**, the variance-covariance matrix in Equation 2 becomes:

$$\begin{bmatrix} \mathbf{G}_{11}\sigma_1^2 & \mathbf{G}_{12}\sigma_{12} \\ \mathbf{G}_{21}\sigma_{12} & \mathbf{G}_{22}\sigma_2^2 \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1\mathbf{Z}_1' \dfrac{\sigma_1^2}{k_1} & \mathbf{Z}_1\mathbf{Z}_2' \dfrac{\sigma_{12}}{k_{12}} \\ \mathbf{Z}_2\mathbf{Z}_1' \dfrac{\sigma_{12}}{k_{12}} & \mathbf{Z}_2\mathbf{Z}_2' \dfrac{\sigma_2^2}{k_2} \end{bmatrix}. \quad (8)$$

Equation 8 shows that the scaling factors of **G** and the variance components are completely confounded. Therefore, other scaling factors of **G** can be used to estimate the genetic correlation as:

$$\hat{r}_g = \frac{\dfrac{\hat{\sigma}_{12}}{k_{12}}}{\sqrt{\dfrac{\hat{\sigma}_1^2}{k_1}}\sqrt{\dfrac{\hat{\sigma}_2^2}{k_2}}} = \frac{\sqrt{k_1}\sqrt{k_2}}{k_{12}} \frac{\hat{\sigma}_{12}}{\sqrt{\hat{\sigma}_1^2}\sqrt{\hat{\sigma}_2^2}}. \quad (9)$$

Equation 9 shows that the genetic correlation is directly estimated from the variance components when the scaling factors of **G** fulfil the property $k_{12} = \sqrt{k_1}\sqrt{k_2}$. When $k_{12} \neq \sqrt{k_1}\sqrt{k_2}$, the correlation based on variance components should be multiplied by $\left(\dfrac{\sqrt{k_1}\sqrt{k_2}}{k_{12}}\right)$ to correct the estimated genetic correlation. By changing the scaling factors, the genetic variances change as well. When genetic variances of the current population are of interest, the within-population blocks in **G** should be scaled as in Equation 7 (Legarra 2016).

Equation 8 and 9 show that the genetic correlation is estimated when the scaling factors in **G** are the same for all blocks. When all scaling factors are equal to 1, so effectively no scaling factor is used, the (co)variances represent the (co)variances of the causal effects i.e., $\sigma_{\alpha_1}^2$, $\sigma_{\alpha_2}^2$, and $\sigma_{\alpha_{12}}$. A disadvantage of this scaling is that elements of **G** can become very large, which can result in very small variance components that may be flagged as too small in statistical

178 software. This might be prevented by either scaling up the phenotypic variance by multiplying
179 all phenotypes by a constant, or by scaling down the elements in **G** by dividing all elements
180 by the same constant. Both scaling approaches have no influence on the genetic correlation,
181 but do affect the genetic (co)variances.
182
183 **Simulations**
184 Simulations were used to validate the results above. Two populations of 2500 individuals
185 each with phenotypes for a trait influenced by the same 15 000 loci were simulated. Allele
186 frequencies of the loci were sampled from a U-shape distribution, independently in both
187 populations. Genotypes were allocated to individuals according to Hardy-Weinberg
188 equilibrium, assuming that loci were segregating independently. Therefore, genetic
189 correlations between populations were only affected by pleiotropy and not by linkage
190 disequilibrium.
191 Allele substitution effects were sampled from a bi-variate normal distribution with means
192 zero and variances 1, and a correlation of 0.5 between allele substitution effects in both
193 populations. The allele substitution effects were multiplied with the corresponding genotypes
194 to calculate additive genetic values for individuals, assuming additive gene action.
195 Environmental effects were sampled from a normal distribution with variance ($\frac{1}{h^2}$-1) times
196 the genetic variance, where the genetic variance was calculated across all individuals in both
197 populations. The heritability was set to 0.9, to ensure that there was sufficient power in the
198 data to estimate the (co)variances. Phenotypes were the sum of additive genetic and
199 environmental effects, and were standardized to an average of 0 and a standard deviation of
200 100. Simulations were replicated 100 times.
201 Phenotypes were analyzed in a two-trait model, using four different **G** matrices; two **G**
202 matrices derived above, and two commonly used **G** matrices for multiple populations (Chen

et al. 2013; Makgahlela *et al.* 2013). In all four methods, genotypes at causal loci were used to calculate **G**. The methods differed in scaling factors as well as in centering of genotypes, being performed either within or across populations.

In the first three methods, the genotypes in **Z** were centered within population as $g_{ijm} - 2p_{jm}$, where g_{ijm} is the allele count of individual i from population m at locus j and p_{jm} is the allele frequency at locus j in population m. The first method, **G**_New, scaled **G** following Equation 9:

$$\mathbf{G}\_\text{New} = \begin{bmatrix} \frac{\mathbf{Z}_1\mathbf{Z}_1'}{\sum 2p_{1j}(1-p_{1j})} & \frac{\mathbf{Z}_1\mathbf{Z}_2'}{\sqrt{\sum 2p_{1j}(1-p_{1j})}\sqrt{\sum 2p_{2j}(1-p_{2j})}} \\ \frac{\mathbf{Z}_2\mathbf{Z}_1'}{\sqrt{\sum 2p_{1j}(1-p_{1j})}\sqrt{\sum 2p_{2j}(1-p_{2j})}} & \frac{\mathbf{Z}_2\mathbf{Z}_2'}{\sum 2p_{2j}(1-p_{2j})} \end{bmatrix}.$$

In the second method, **G**_1, scaling factors were equal to 1:

$$\mathbf{G}\_1 = \begin{bmatrix} \mathbf{Z}_1\mathbf{Z}_1' & \mathbf{Z}_1\mathbf{Z}_2' \\ \mathbf{Z}_2\mathbf{Z}_1' & \mathbf{Z}_2\mathbf{Z}_2' \end{bmatrix}.$$

The third method, **G**_Chen, calculated **G** according to Chen *et al.* (2013):

$$\mathbf{G}\_\text{Chen} = \begin{bmatrix} \frac{\mathbf{Z}_1\mathbf{Z}_1'}{\sum 2p_{1j}(1-p_{1j})} & \frac{\mathbf{Z}_1\mathbf{Z}_2'}{\sum 2\sqrt{p_{1j}(1-p_{1j})p_{2j}(1-p_{2j})}} \\ \frac{\mathbf{Z}_2\mathbf{Z}_1'}{\sum 2\sqrt{p_{1j}(1-p_{1j})p_{2j}(1-p_{2j})}} & \frac{\mathbf{Z}_2\mathbf{Z}_2'}{\sum 2p_{2j}(1-p_{2j})} \end{bmatrix}.$$

The fourth method, **G**_Across, used the average allele frequency across both populations instead of population-specific allele frequencies to center the genotypes (e.g., Makgahlela *et al.* 2013). Thus, the matrix of genotypes, denoted **Z***, had elements $g_{ijm} - 2\bar{p}_j$, where $\bar{p}_j$ is the average allele frequency across both populations at locus j. The scaling factor was the same for all blocks:

220 $$\mathbf{G}\_\text{Across} = \begin{bmatrix} \dfrac{\mathbf{Z}_1^{*}\mathbf{Z}_1^{*'}}{\sum 2\bar{p}_j(1-\bar{p}_j)} & \dfrac{\mathbf{Z}_1^{*}\mathbf{Z}_2^{*'}}{\sum 2\bar{p}_j(1-\bar{p}_j)} \\ \dfrac{\mathbf{Z}_2^{*}\mathbf{Z}_1^{*'}}{\sum 2\bar{p}_j(1-\bar{p}_j)} & \dfrac{\mathbf{Z}_2^{*}\mathbf{Z}_2^{*'}}{\sum 2\bar{p}_j(1-\bar{p}_j)} \end{bmatrix}.$$

221 **G**_New, **G**_1 and **G**_Across fulfilled the property $k_{12} = \sqrt{k_1}\sqrt{k_2}$ to directly estimate the
222 genetic correlation. In **G**_Chen, $k_{12} \neq \sqrt{k_1}\sqrt{k_2}$ when allele frequencies in the two populations
223 were different. Therefore, the genetic correlation estimated with **G**_Chen was multiplied by
224 $\dfrac{\sqrt{k_1}\sqrt{k_2}}{k_{12}} = \dfrac{\sqrt{\sum 2p_{1j}(1-p_{1j})}\sqrt{\sum 2p_{2j}(1-p_{2j})}}{\sum 2\sqrt{p_{1j}(1-p_{1j})p_{2j}(1-p_{2j})}}$ to correct the estimate. Moreover, the current
225 populations were the base population for the within-population blocks of **G**_New and
226 **G**_Chen, so those **G** matrices estimated the genetic variances within the current populations
227 (Speed and Balding 2015; Legarra 2016). As explained before, the variances of **G**_1
228 represented the variances of the causal effects. For **G**_Across, the base population was not
229 clearly defined, so the interpretation of the estimated genetic variances is unclear. See
230 supporting information for the R-script and seeds used to simulate genotypes and phenotypes
231 and to calculate the different **G** matrices.

RESULTS

Variance components

In Figure 1, the estimated genetic variance using **G**_New is plotted against the simulated genetic variance. This figure shows that the estimates varied only slightly around the simulated values. This shows that **G**_New unbiasedly estimated the genetic variance in the current populations.

As expected, **G**_New and **G**_Chen estimated the same genetic variances (Figure 2 and 3). The variances of **G**_1 represented the variances of the causal effects. By multiplying those variances by $\sum 2p_{jm}(1-p_{jm})$ for population *m*, genetic variances identical to **G**_New and **G**_Chen were obtained. The genetic variance estimated with **G**_Across was approximately a factor 1.5 higher than the genetic variance estimated with **G**_New and **G**_Chen. Also the scaling factors k_1 and k_2 were approximately a factor 1.5 higher. Hence, when multiplying the variances estimated with **G**_Across by the ratio in scaling factors, estimates became identical to those with **G**_New and **G**_Chen. So, the difference in estimated variances between methods was completely explained by the difference in scaling factors, while centering genotypes within or across populations had no effect on estimated variances. Estimated residual variances were exactly the same for the four different **G** matrices.

Genetic correlation

Despite differences in (co)variance estimates, **G**_New, **G**_1, and **G**_Across yielded the same estimated genetic correlation (Figure 4) which was an unbiased estimate of the simulated genetic correlation (Figure 5). This is because differences in genetic covariances among models were compensated by corresponding differences in genetic variances. The

genetic correlation estimated using **G**_Chen was ~20% lower. When multiplying this estimate by $\frac{\sqrt{k_1}\sqrt{k_2}}{k_{12}}$ =1.23, the genetic correlation became identical to the other three methods.

DISCUSSION

The aim of this paper was to define a multi-population genomic relationship matrix to estimate current genetic variances within and genetic correlations between populations. We derived a genomic relationship matrix, **G**_New, that yields unbiased estimates of current genetic variances, covariances and correlations. Moreover, we showed the required property for other genomic relationship matrices to estimate the genetic correlation between populations, even though estimated variance components are not necessarily related to the current populations.

Methods to calculate the genomic relationship matrix

From the four methods used in this paper to calculate **G**, **G**_New was the only matrix correctly estimating both current genetic variances as well as genetic correlations. **G**_Chen also estimated current genetic variances, but the estimated genetic correlation had to be multiplied by $\frac{\sqrt{k_1}\sqrt{k_2}}{k_{12}}$. **G**_1 estimated the correct genetic correlation, but estimated the variance of causal effects instead of the genetic variance. Although the base population in **G**_Across was not well defined, genetic correlations were correctly estimated but there was no clear interpretation of the estimated genetic variances. Results also showed that genetic variances were not affected by centering the allele count, as shown before by Strandén and Christensen (2011).

Table 1 gives an overview of the most frequently used methods to calculate **G** across multiple populations, with scaling factors and correction factors for the estimated genetic correlation. **G**_New, **G**_1, **G**_Across, and the method described by Erbe *et al.* (2012) directly estimate the correct genetic correlation. The **G**_Chen method does not directly estimate the genetic correlation, but the estimate can be corrected using the scaling factors. Those five

methods all assume that allele substitution effects are independent of allele frequency, similar to method 1 of VanRaden (2008). This is in contrast to another regularly used method, namely method 2 of VanRaden (2008), also described by Yang (2010). This method yields a valid relationship matrix only when the average effect at a locus is proportional to the reciprocal of the square root of expected heterozygosity at that locus (Appendix, Equation A8). So, this method assumes that marker effects are determined by their allele frequency, with larger effects for rarer alleles. For a trait determined by relatively few genes and undergoing directional selection, this assumption may be plausible, since selection acts stronger on causal loci with a larger effect (Haldane 1924; Wright 1931, 1937). It is, however, a very strong assumption in general. Many traits may experience only weak selection, and/or are determined by many genes. In those cases, allele frequency distribution is determined mainly by the interplay of mutation and drift, and a direct relationship between effect size and allele frequency is not expected. Therefore, the assumption of independence between allele frequency and allele substitution effects seems more realistic for most traits. Moreover, when allele substitution effects would depend on allele frequency, effects for exactly the same trait would differ between populations when allele frequencies differ. This makes the interpretation of a genetic correlation estimated using method 2 of VanRaden (2008) rather difficult. Therefore, we advise to use **G** matrices based on method 1 instead of method 2 of VanRaden (2008), especially when multiple populations are considered**.**

In this paper, we assumed that causal loci were known and were used to calculate **G**. In this way, differences in linkage disequilibrium (LD) between markers and causal loci across populations did not affect the results and all genetic variance was explained by **G**. When genomic markers are used to calculate **G**, differences in LD can affect the results, since the LD pattern is known to differ across populations in humans (Sawyer *et al.* 2005) as well as in livestock (e.g., Heifetz *et al.* 2005; Gautier *et al.* 2007; Veroneze *et al.* 2013). This difference

in LD is likely to affect the estimated genetic correlation, since it reduces the correlation of marker effects (Gianola *et al.* 2015). Moreover, markers might not explain all genetic variance when there is no complete LD between a causal locus and at least one marker (e.g., Yang *et al.* 2010; Daetwyler *et al.* 2013). This can affect the estimated genetic correlation when the variance explained by the markers shows either a higher or lower genetic correlation than the part not explained (Bulik-Sullivan *et al.* 2015). Therefore, it is difficult to predict the effect of not explaining all genetic variance by markers on the estimated genetic correlation. In a follow-up study, we will investigate the effect of using marker genotypes on the estimated genetic correlation between populations.

Other approaches to estimate the genetic correlation between populations

We focused on using genomic relationships in a multi-trait model to estimate genetic correlations between populations. Genetic correlations can also be estimated using summary statistics of genome-wide association studies (GWAS; Bulik-Sullivan *et al.* 2015; Brown *et al.* 2016) or using random regression on genotypes (Sørensen *et al.* 2012; Krag *et al.* 2013). The method based on summary statistics of GWAS combines information from different studies and weights estimated marker effects by LD overlap and corresponding *z* score (Bulik-Sullivan *et al.* 2015; Brown *et al.* 2016). This method is beneficial when the costs of collecting enough data are high and data sharing is not possible. It is, however, not known whether this method estimates the correct genetic correlation. The method using random regression on genotypes is equivalent to the multi-trait GREML method used in this study, since both estimate the same additive genetic values when the genotypes are centered and scaled in the same way (Habier *et al.* 2007; VanRaden 2008; Goddard 2009). Variance components estimated with random regression on marker genotypes represent variances of marker effects (Meuwissen *et al.* 2001), similar to **G**_1, when the same centered genotypes

are used as input. Hence, random regression on centered genotypes can also be used to estimate genetic correlations between populations. When genotypes for the random regression are centered and scaled, the estimated genetic correlation becomes equal to the estimated genetic correlation using **G** based on method 2 of VanRaden (VanRaden 2008; Yang *et al.* 2010). Therefore, the interpretation of this estimated genetic correlation remains unclear as well.

Importance of the genetic correlation between populations

The genetic correlation between populations is an important parameter for genomic prediction, since it determines the usefulness of combining information from multiple populations. A low genetic correlation means that it is very unlikely that combining populations will increase the accuracy of estimated genetic values. Therefore, the genetic correlation partly determines the accuracy of across- or multi-population genomic prediction. For predicting the accuracy in those scenarios, an accurate estimation of genetic correlations is essential (Wientjes *et al.* 2015; Wientjes *et al.* 2016). For predicting response to selection, both the accuracy as well as current genetic variances are needed (Falconer and Mackay 1996). Even though the accuracy of estimated genetic values is quite consistent across methods for calculating **G** (Makgahlela *et al.* 2013, 2014; Lourenco *et al.* 2016), for estimating genetic (co)variances and correlations it is important to use the **G**_New matrix.

Genetic correlation versus genic correlation

The genetic correlation is defined based on additive genetic (co)variances. Under selection, however, additive genetic (co)variances change over generations, since selection creates transient gametic phase disequilibrium (i.e., correlations between allele substitution effects at different loci). This process is also known as the Bulmer effect (Bulmer 1971). Therefore,

357 genetic (co)variances and correlations depend not only on the genetic background of the traits,
358 but also on transient processes like the type and intensity of selection. Apart from additive
359 genetic (co)variances, quantitative genetics also describes genic (co)variances (e.g., Bulmer
360 1980; Bulmer 1989), defined as the additive genetic (co)variance in the absence of gametic
361 phase disequilibrium. In contrast to genetic variances, genic variances are independent from
362 selection and are always equal to twice the Mendelian sampling variance (Hill 2014). In
363 analogy to genic (co)variances, genic correlations can be defined as well. We believe that
364 genic correlations are more relevant than additive genetic correlations, since genic
365 correlations are not influenced by transient processes and, therefore, more constant across
366 generations.

367 In our simulation study, allele substitution effects were randomly sampled, so no transient
368 gametic phase disequilibrium was present and genic (co)variances were equal to the additive
369 genetic (co)variances. In all situations, genic variances can be estimated when the base
370 population of the relationship matrix is unselected and phenotypic records on which selection
371 decisions are based are available (Henderson 1985). It is also shown that even when
372 phenotypic records from the base population are absent, the genic variance can be estimated
373 when phenotypic records for several generations are available and the base population is
374 unselected (Henderson 1985; Van der Werf and de Boer 1990). It can be expected that as long
375 as several generations of phenotypic data is available in combination with the relationships
376 between all those individuals, variances are corrected for selection and effectively genic
377 variances are estimated. Therefore, genic correlations can likely be calculated using **G**_New,
378 provided that data is available for several generations.
379

Conclusion

The properties of the genomic relationship matrix affect estimates of genetic variances within as well as genetic correlations between populations. For estimating current genetic variances, allele frequencies of the current population should be used to calculate relationships within that population. For estimating genetic correlations between populations, scaling factors of the different blocks of the relationship matrix, based on method 1 of VanRaden (2008), should fulfill the property $k_{12} = \sqrt{k_1}\sqrt{k_2}$. When this property is not fulfilled, the estimated genetic correlation can be corrected by multiplying the estimate by $\frac{\sqrt{k_1}\sqrt{k_2}}{k_{12}}$. In this study we present a genomic relationship matrix, **G**_New, which directly results in current genetic variances as well as genetic correlations between populations.

393 **ACKNOWLEDGMENTS**

394 This research is supported by the Netherlands Organisation of Scientic Research (NWO)
395 and the Breed4Food consortium partners Cobb Europe, CRV, Hendrix Genetics and
396 TopigsNorsvin.

397

398

399 **LITERATURE CITED**

400

401 Bohren, B. B., W. G. Hill and A. Robertson, 1966 Some observations on asymmetrical
402 correlated responses to selection. Genet. Res. 7: 44-57.
403 Brown, B. C., C. J. Ye, A. L. Price and N. Zaitlen, 2016 Transethnic genetic-correlation
404 estimates from summary statistics. Am. J. Hum. Genet. 99: 76-88.
405 Bulik-Sullivan, B., H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, *et al.*, 2015 An atlas of
406 genetic correlations across human diseases and traits. Nat. Genet. 47: 1236-1241.
407 Bulmer, M., 1989 Maintenance of genetic variability by mutation-selection balance: a child's
408 guide through the jungle. Genome 31: 761-767.
409 Bulmer, M. G., 1971 The effect of selection on genetic variability. Am. Nat. 105: 201-211.
410 Bulmer, M. G., 1980 *The mathematical theory of quantitative genetics*. Clarendon Press.,
411 Oxford.
412 Chen, L., F. Schenkel, M. Vinsky, D. Crews and C. Li, 2013 Accuracy of predicting genomic
413 breeding values for residual feed intake in Angus and Charolais beef cattle. J. Anim. Sci.
414 91: 4669-4678.
415 Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. De los Campos and J. M. Hickey, 2013
416 Genomic prediction in animals and plants: Simulation of data, validation, reporting, and
417 benchmarking. Genetics 193: 347-365.
418 De Candia, T. R., S. H. Lee, J. Yang, B. L. Browning, P. V. Gejman, *et al.*, 2013 Additive
419 genetic variation in schizophrenia risk is shared by populations of African and European
420 descent. Am. J. Hum. Genet. 93: 463-470.
421 Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, *et al.*, 2012 Improving
422 accuracy of genomic predictions within and between dairy cattle breeds with imputed high-
423 density single nucleotide polymorphism panels. J. Dairy Sci. 95: 4114-4129.

424 Falconer, D. S., 1952 The problem of environment and selection. Amer. Nat. 86: 293-298.

425 Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*. Pearson
426 Education Limited, Harlow.

427 Fisher, R. A., 1918 The Correlation between relatives on the supposition of Mendelian
428 inheritance. Trans. Roy. Soc. Edinburgh 52: 399-433.

429 Fisher, R. A., 1930 *The genetical theory of natural selection: a complete variorum edition*.
430 Oxford University Press.

431 Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, *et al.*, 2007 Genetic
432 and haplotypic structure in 14 European and African cattle breeds. Genetics 177: 1059-
433 1070.

434 Gianola, D., G. De los Campos, M. A. Toro, H. Naya, C.-C. Schön, *et al.*, 2015 Do molecular
435 markers inform about pleiotropy? Genetics 201: 23-29.

436 Goddard, M. E., 2009 Genomic selection: Prediction of accuracy and maximisation of long
437 term response. Genetica 136: 245-257.

438 Habier, D., R. L. Fernando and J. C. M. Dekkers, 2007 The impact of genetic relationship
439 information on genome-assisted breeding values. Genetics 177: 2389-2397.

440 Haldane, J. B. S., 1924 A mathematical theory of natural and artificial selection—I. Trans.
441 Camb. Phil. Soc. 23: 19-41.

442 Harris, B. L. and D. L. Johnson, 2010 Genomic predictions for New Zealand dairy bulls and
443 integration with national genetic evaluation. J. Dairy Sci. 93: 1243-1252.

444 Hayes, B. J., P. M. Visscher and M. E. Goddard, 2009 Increased accuracy of artificial
445 selection by using the realized relationship matrix. Genet. Res. 91: 47-60.

446 Heifetz, E. M., J. E. Fulton, N. O'Sullivan, H. Zhao, J. C. M. Dekkers, *et al.*, 2005 Extent and
447 consistency across generations of linkage disequilibrium in commercial layer chicken
448 breeding populations. Genetics 171: 1173-1181.

449 Henderson, C. R., 1985 Best linear unbiased prediction using relationship matrices derived
450 from selected base populations. J. Dairy Sci. 68: 443-448.
451 Hill, W. G., 2014 Applications of population genetics to animal breeding, from Wright, Fisher
452 and Lush to genomic prediction. Genetics 196: 1-16.
453 Karoui, S., M. Carabaño, C. Díaz and A. Legarra, 2012 Joint genomic evaluation of French
454 dairy cattle breeds using multiple-trait models. Genet. Sel. Evol. 44: 39.
455 Krag, K., N. A. Poulsen, M. K. Larsen, L. B. Larsen, L. L. Janss, *et al.*, 2013 Genetic
456 parameters for milk fatty acids in Danish Holstein cattle based on SNP markers using a
457 Bayesian approach. BMC Genet. 14: 79.
458 Legarra, A., 2016 Comparing estimates of genetic variance across different relationship
459 models. Theor. Popul. Biol. 107: 26-30.
460 Lehermeier, C., C.-C. Schön and G. De los Campos, 2015 Assessment of genetic
461 heterogeneity in structured plant populations using multivariate whole-genome regression
462 models. Genetics 201: 323-337.
463 Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, C. Y. Chen, W. O. Herring, *et al.*, 2016
464 Crossbreed evaluations in single-step genomic best linear unbiased predictor using
465 adjusted realized relationship matrices. J. Anim. Sci. 94: 909-919.
466 Makgahlela, M. L., I. Strandén, U. S. Nielsen, M. J. Sillanpää and E. A. Mäntysaari, 2013 The
467 estimation of genomic relationships using breedwise allele frequencies among animals in
468 multibreed populations. J. Dairy Sci. 96: 5364-5375.
469 Makgahlela, M. L., I. Strandén, U. S. Nielsen, M. J. Sillanpää and E. A. Mäntysaari, 2014
470 Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic
471 evaluation of a multibreed population. J. Dairy Sci. 97: 1117-1127.
472 Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2001 Prediction of total genetic value
473 using genome-wide dense marker maps. Genetics 157: 1819-1829.

Olson, K. M., P. M. VanRaden and M. E. Tooker, 2012 Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 95: 5378-5383.

Powell, J. E., P. M. Visscher and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Gen. 11: 800-805.

Sawyer, S. L., N. Mukherjee, A. J. Pakstis, L. Feuk, J. R. Kidd, *et al.*, 2005 Linkage disequilibrium patterns vary substantially among populations. Europ. J. Hum. Genet. 13: 677-686.

Sørensen, L. P., L. Janss, P. Madsen, T. Mark and M. S. Lund, 2012 Estimation of (co)variances for genomic regions of flexible sizes: application to complex infectious udder diseases in dairy cattle. Genet. Sel. Evol. 44: 18.

Speed, D. and D. J. Balding, 2015 Relatedness in the post-genomic era: is it still useful? Nat. Rev. Genet. 16: 33-44.

Strandén, I. and O. F. Christensen, 2011 Allele coding in genomic evaluation. Genet. Sel. Evol. 43: 25.

Van der Werf, J. and I. de Boer, 1990 Estimation of additive genetic variance when base populations are selected. J. Anim. Sci. 68: 3124-3132.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414-4423.

Veroneze, R., P. S. Lopes, S. E. F. Guimarães, F. F. Silva, M. S. Lopes, *et al.*, 2013 Linkage disequilibrium and haplotype block structure in six commercial pig lines. J. Anim. Sci. 91: 3493-3501.

Visscher, P. M., G. Hemani, A. A. E. Vinkhuyzen, G.-B. Chen, S. H. Lee, *et al.*, 2014 Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. PLoS Genet 10: e1004269.

Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten, *et al.*, 2015 Empirical and deterministic accuracies of across-population genomic prediction. Genet. Sel. Evol. 47: 5.

Wientjes, Y. C. J., P. Bijma, R. F. Veerkamp and M. P. L. Calus, 2016 An equation to predict the accuracy of genomic values by combining data from multiple traits, breeds, lines, or environments. Genetics 202: 799-823.

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97-159.

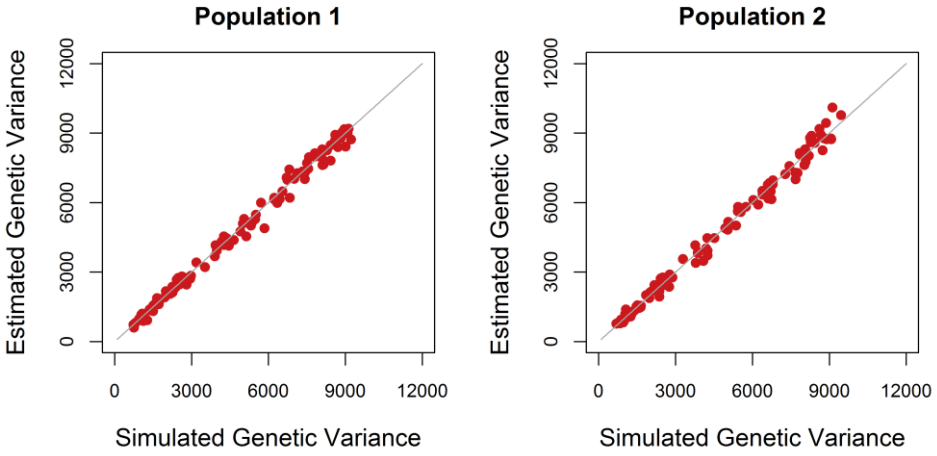
Wright, S., 1937 The distribution of gene frequencies in populations. Proc. Nat. Acad. Sci. USA 23: 307-320.

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.

Yang, J., A. Bakshi, Z. Zhu, G. Hemani, A. A. E. Vinkhuyzen, *et al.*, 2015 Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. 47: 1114-1120.

513 **FIGURES**

514



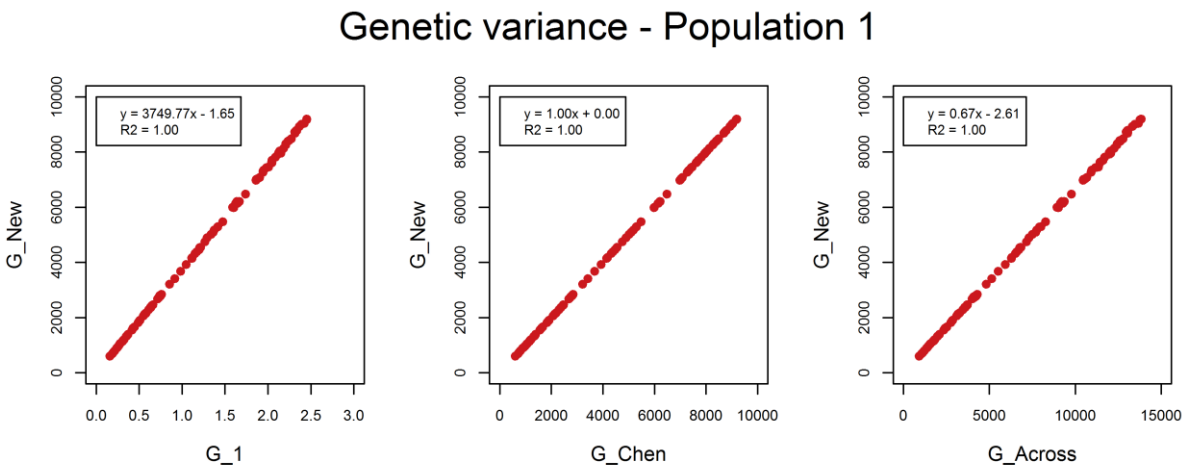
515

516 **Figure 1 – Estimated versus simulated genetic variance.** The estimated genetic variance in
517 both populations in each of the 100 replicates using the genomic relationship matrix derived
518 in this study (**G**_New) versus the simulated genetic variance. The grey line represents the line
519 y=x.

520

521

522



523

524 **Figure 2 – Estimated genetic variance in population 1.** The estimated genetic variance in
525 population 1 in each of the 100 replicates using the genomic relationship matrix derived in
526 this study (**G**_New) versus the estimated genetic variance using population-specific allele
527 frequencies and either a genomic relationship matrix without scaling factors (**G**_1) or based
528 on the method of Chen *et al.* (2013; **G**_Chen), or using allele frequencies across populations
529 (**G**_Across).

530

531

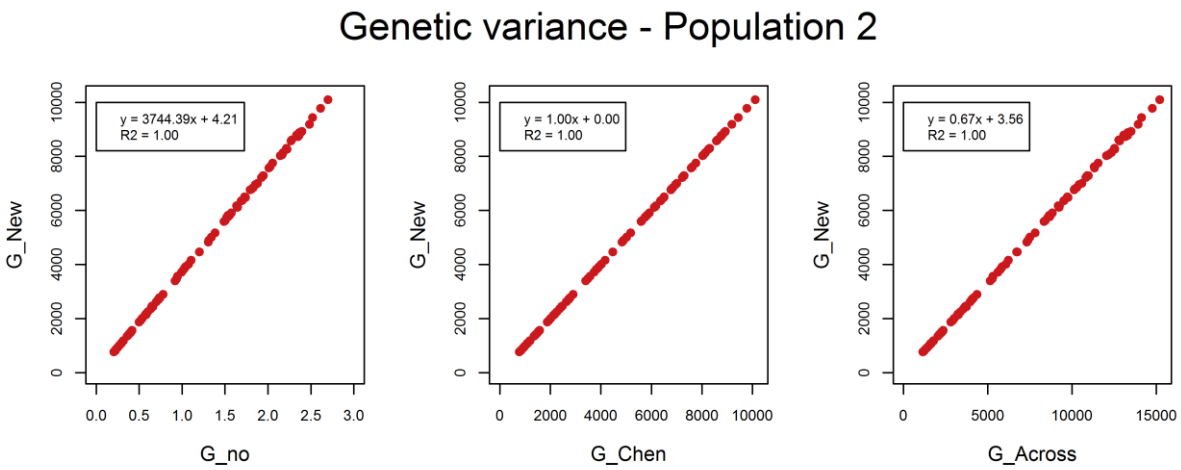
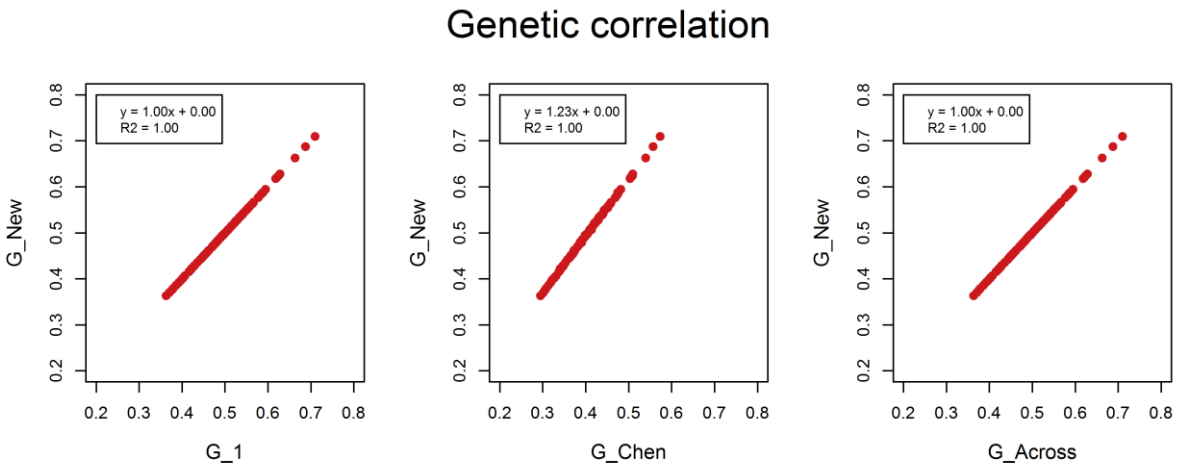
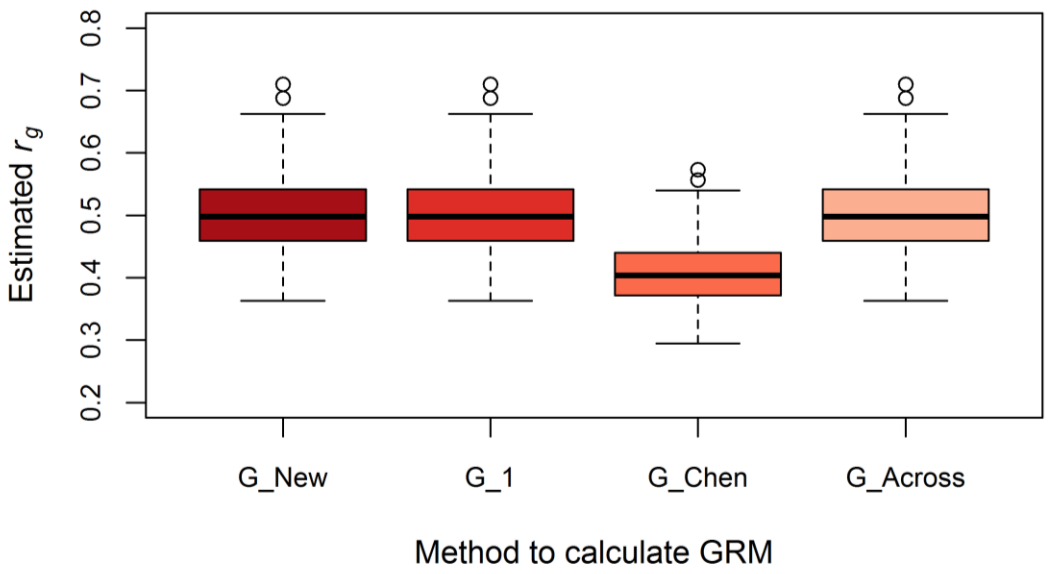


Figure 3 – Estimated genetic variance in population 2. The estimated genetic variance in population 2 in each of the 100 replicates using the genomic relationship matrix derived in this study (**G**_New) versus the estimated genetic variance using population-specific allele frequencies and either a genomic relationship matrix without scaling factors (**G**_1) or based on the method of Chen *et al.* (2013; **G**_Chen), or using allele frequencies across populations (**G**_Across).



542
543 **Figure 4 – Estimated genetic correlation between population 1 and 2.** The estimated
544 genetic correlation between population 1 and 2 in each of the 100 replicates using the
545 genomic relationship derived in this study (**G**_New) versus the estimated genetic correlation
546 using population-specific allele frequencies and either a genomic relationship matrix without
547 scaling factors (**G**_1), based on the method of Chen *et al.* (2013; **G**_Chen), or using allele
548 frequencies across populations (**G**_Across).
549



550

551 **Figure 5 – Boxplot of the estimated genetic correlation using different methods to**
552 **calculate the genomic relationship matrix.** The estimated genetic correlation between
553 population 1 and 2 in each of the 100 replicates using the genomic relationship matrix derived
554 in this study (**G**_New), using population-specific allele frequencies and either a genomic
555 relationship matrix without scaling factors (**G**_1), or based on the method of Chen *et al.*
556 (2013; **G**_Chen), or using allele frequencies across populations (**G**_Across). The simulated
557 genetic correlation was 0.5.

Table 1 – Overview of frequently used method to calculate **G** across populations with scaling and correction factors.

Method of calculating G[a]	Described by	Used scaling factors of the different blocks in G[b]			Correction factor to correct the genetic correlation
		k_1[c]	k_2[c]	k_{12}[c]	
G_New	This study	$\sum 2p_{1i}(1-p_{1i})$	$\sum 2p_{2i}(1-p_{2i})$	$\sqrt{\sum 2p_{1i}(1-p_{1i})}\sqrt{\sum 2p_{2i}(1-p_{2i})}$	Not needed
G_1	This study	1	1	1	Not needed
G_Chen	Chen *et al.* (2013)	$\sum 2p_{1i}(1-p_{1i})$	$\sum 2p_{2i}(1-p_{2i})$	$\sum 2\sqrt{p_{1i}(1-p_{1i})p_{2i}(1-p_{2i})}$	$\frac{\sqrt{\sum 2p_{1i}(1-p_{1i})}\sqrt{\sum 2p_{2i}(1-p_{2i})}}{\sum 2\sqrt{p_{1i}(1-p_{1i})p_{2i}(1-p_{2i})}}$
G_Across	VanRaden (2008)/ Makgahlela *et al.* (2013)	$\sum 2\bar{p}_i(1-\bar{p}_i)$	$\sum 2\bar{p}_i(1-\bar{p}_i)$	$\sum 2\bar{p}_i(1-\bar{p}_i)$	Not needed
Erbe	Erbe *et al.* (2012)	$\sum 2p_i^*(1-p_i^*)$	$\sum 2p_i^*(1-p_i^*)$	$\sum 2p_i^*(1-p_i^*)$	Not needed
VanRaden2/ Yang	VanRaden (2008); Yang *et al.* (2010)	Nr. of markers[d]	Nr. of markers[d]	Nr. of markers[d]	Unknown

[a] Methods were compared assuming that no adjustment for inbreeding or regression back to the pedigree relationship matrix was performed.

[b] k_1 is the scaling factor of the block containing relationships in population 1, k_2 is the scaling factor of the block containing relationships in population 2, and k_{12} is the scaling factor of the block containing relationship between population 1 and 2.

[c] p_{1i} is the allele frequency in population 1, p_{2i} is the allele frequency in population 2, $\bar{p}_i$ is the average allele frequency across populations, p_i^* is the estimated allele frequency when the populations separated.

[d] Per marker *i*, genotypes are scaled by $\sqrt{2p_i(1-p_i)}$.

APPENDIX

The **G** matrix based on method 2 of VanRaden (2008) and Yang *et al.* (2010), **G**_VR2, weights markers by the reciprocal of the square root of the variance of its genotypes. In this Appendix, it is shown that this is only correct under the assumption that the variance of the average effect (α) at a locus, say *l*, is inversely proportional to expected heterozygosity at that locus,

$$\sigma^2_{\alpha_l} = \frac{c}{2p_l(1-p_l)}, \tag{A1}$$

where *c* is a constant, and p_l the allele frequency at locus *l*.

Consider the single-trait mixed model $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$, where **a** is the vector of random additive genetic effects, with $\text{var}(\mathbf{a}) = \mathbf{G}\sigma^2_A$. This mixed model is valid only when $\mathbf{G}\sigma^2_A$ indeed represents the covariances between additive genetic effects (*A*) of individuals. This requires that

$$\mathbf{G}_{ij} = \text{cov}(A_i, A_j) / \text{var}(A), \tag{A2}$$

where *i* and *j* are individuals.

By definition, the additive genetic effect of an individual is the sum of the average effects at its loci, weighted by the centred allele count (Fisher 1918; Falconer and Mackay 1996),

$$A_i = \sum_l (x_{il} - 2p_l)\alpha_l, \tag{A3}$$

where x_{il} is the allele count of individual *i* at locus *l*, taking values 0, 1 or 2. Thus

$$\text{cov}(A_i, A_j) = \text{cov}\left[\sum_l (x_{il} - 2p_l)\alpha_l, \sum_l (x_{jl} - 2p_l)\alpha_l\right]. \tag{A4}$$

For the genic covariance, the $(x_{il} - 2p_l)\alpha_l$ terms are independent between loci by definition (Bulmer 1971), so that the covariance reduces to

$$\text{cov}(A_i, A_j) = \sum_l (x_{il} - 2p_l)(x_{jl} - 2p_l)\sigma^2_{\alpha_l}. \tag{A5}$$

589 Substituting the relationship between average effects and allele frequency given by
590 Equation A1 yields

591 $$\mathrm{cov}(A_i, A_j) = c\sum_l \left[\frac{(x_{il} - 2p_l)(x_{jl} - 2p_l)}{2p_l(1-p_l)} \right]. \quad \text{(A6)}$$

592 Analogously, the genic variance equals

593 $$\mathrm{var}(A) = \sum_l 2p_l(1-p_l)\sigma^2_{\alpha_l} = \sum_l c = n_l c,$$

594 where n_l is the number of loci. Finally, from Equation A2,

595 $$\mathbf{G}_{ij} = \mathrm{cov}(A_i, A_j)/\mathrm{var}(A) = \frac{1}{n_l}\sum_l \left[\frac{(x_{il} - 2p_l)(x_{jl} - 2p_l)}{2p_l(1-p_l)} \right], \quad \text{(A7)}$$

596 which is **G**_VR2. Thus obtaining **G**_VR2 requires Equation A1.

597 Hence, **G**_VR2 is valid under the assumption that the magnitude of the average effect at a
598 locus is proportional to the reciprocal of the square root of expected heterozygosity at that
599 locus,

600 $$\alpha_l \propto \frac{1}{\sqrt{2p_l(1-p_l)}}. \quad \text{(A8)}$$

601 Equation A7 shows that elements of **G**_VR2 are the genome-wide average of the
602 correlations at individual loci; the term in square-brackets is the correlation between additive
603 genetic effects at locus *l*, and the sum of these terms is divided by the number of loci. Thus
604 **G**_VR2 may have been motivated as the genome-wide average of relationships at individual
605 loci.

606 However, relatedness refers to the correlation between the total additive genetic effects of
607 individuals (Equation A2), which are sums of additive genetic effects at individual loci. In
608 general, the correlation between sums does not equal the average correlation between
609 components of the sums,

610 $$\mathbf{G}_{ij} \neq \frac{1}{n_l}\sum_{l}\mathbf{G}_{ijl} \quad \text{(A9)}$$

611 but is defined as the ratio of the covariance and variance of the sum,

612 $$\mathbf{G}_{ij} = \mathrm{cov}(A_i, A_j)/\sigma_A^2. \quad \text{(A10)}$$

613 Equations A9 and A10 are only equal to each other under the assumption given in Equation

614 A1.

615

616

617