An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments

Yvonne C.J. Wientjes*, Piter Bijma§, Roel F. Veerkamp*, Mario P.L. Calus*

*Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 6700 AH Wageningen, The Netherlands

§Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, The Netherlands
Key words: genomic prediction, multi-population, accuracy, prediction equation

Author information:

Yvonne Wientjes
Animal Breeding and Genomics Centre
Wageningen UR Livestock Research
P.O. box 338, 6700 AH Wageningen, the Netherlands
E-mail: yvonne.wientjes@wur.nl
Phone: +31 317 482207
ABSTRACT

Predicting the accuracy of estimated genomic values using genome-wide marker information is an important step in designing training populations. Currently, different deterministic equations are available to predict accuracy within populations, but not for multi-population scenarios where data from multiple breeds, lines or environments is combined. Therefore, our objective was to develop and validate a deterministic equation to predict the accuracy of genomic values when different populations are combined in one training population. The input parameters of the derived prediction equation are the number of individuals and the heritability from each of the populations in the training population, the genetic correlations between the populations, i.e., the correlation between allele substitution effects of quantitative trait loci, the effective number of chromosome segments across predicted and training populations, and the proportion of the genetic variance in the predicted population captured by the markers in each of the training populations. Validation was performed based on real genotype information of 1033 Holstein Friesian cows that were divided in three different populations by combining half-sib families in the same population. Phenotypes were simulated for multiple scenarios, differing in heritability within populations and in genetic correlations between the populations. Results showed that the derived equation can accurately predict the accuracy of estimating genomic values for different scenarios of multi-population genomic prediction. Therefore, the derived equation can be used to investigate the potential accuracy of different multi-population genomic prediction scenarios and to decide on the most optimal design of training populations.
Genomic markers can be used to estimate genomic values of individuals, also known as additive genetic values or breeding values, that are used to select animals (e.g., Dekkers 2007; De Roos et al. 2011) and plants for breeding (e.g., Heffner et al. 2009; Jannink et al. 2010), and in humans to predict the genetic risk of diseases (e.g., Wray et al. 2007; De Los Campos et al. 2010). In genomic prediction, genome-wide single-nucleotide polymorphism (SNP) marker information is used to predict genomic values based on SNP effects estimated in a training population consisting of individuals with known SNP genotypes and phenotypes (Meuwissen et al. 2001). The accuracy of estimating genomic values is in general higher when the size of the training population is larger, when the level of linkage disequilibrium (LD) between the SNPs and the quantitative trait loci (QTL) underlying the trait is higher, and when the predicted individuals are more related to the individuals in the training population (e.g., Daetwyler et al. 2008; Zhong et al. 2009; De los Campos et al. 2013; Wientjes et al. 2013).

For numerically small populations, the size of the training population is limited which restricts the accuracy of genomic prediction. Therefore, combining different populations in one training population for estimating SNP effects is an appealing approach to increase the size of the training population and, thereby, the accuracy of predicting genomic values. The potential accuracy of combining different populations in one training population has been investigated by combining populations from different breeds (e.g., Hayes et al. 2009a; Harris and Johnson 2010), lines (e.g., Zhong et al. 2009; Calus et al. 2014; Lehermeier et al. 2014), subpopulations (e.g., De los Campos et al. 2013), or countries (e.g., Lund et al. 2011; Haile-Mariam et al. 2015). The increase in accuracy by adding individuals from another population to the training population is in most cases much lower than the increase in accuracy obtained by adding an equal number of individuals from the same population. This is a result of
differences that exist between populations, like differences in allele frequencies, LD patterns (De Roos et al. 2008; Zhong et al. 2009; De los Campos et al. 2012), allele substitution effects of QTL (Spelman et al. 2002; Thaller et al. 2003; Wientjes et al. 2015b), environments in combination with genotype by environment interactions (Lund et al. 2011; Haile-Mariam et al. 2015), the presence of QTL that are only segregating in one population (Kemper et al. 2015), and the absence of close family relationships across populations.

Different deterministic equations are available to calculate the accuracy of genomic prediction when the training population is a subset from the same population as the predicted individuals (Daetwyler et al. 2008; VanRaden 2008; Goddard 2009). One type of deterministic equation is based on prediction error variance of the mixed model equation and is using the genomic relationships within the training population and between training and predicted individuals (VanRaden 2008). This equation has been extended to enable the calculation of the accuracy when different populations are combined in one training population (Wientjes et al. 2015b). A disadvantage of this equation is, however, that individuals have to be genotyped before the accuracy can be calculated. Therefore, this equation cannot be used to decide on the most optimal design of training populations. Another type of deterministic equation is able to predict the accuracy before genotype information is available and is based on population parameters, such as the size of the training population, the heritability of the trait and the effective number of chromosome segments (Daetwyler et al. 2008; Daetwyler et al. 2010). This equation can be used to investigate the accuracy of different training population designs, however, the equation is not applicable for situations with more than one population in the training population.

The first objective of this study is to develop a deterministic equation using population parameters to predict the accuracy of genomic values when different populations are combined in one training population. The different combined populations might for example...
be populations from different lines or environments, or populations measured for different traits. The second objective is to validate the derived equation. For the validation, different scenarios of multi-population genomic prediction were considered by dividing 1033 Holstein Friesian cows with real genotypes and simulated phenotypes in three populations, assuming different heritabilities within populations and different genetic correlations between populations. Moreover, the equation was used to investigate the potential accuracy for one specific dairy cattle scenario and one specific human scenario.
MATERIALS AND METHODS

Theory

The accuracy of estimated genomic values ($r_{EGV}$) is defined as the correlation between estimated and true genomic values. The overall accuracy depends on the square root of the proportion of genetic variance captured by the SNPs ($r_{LD}$) and on the accuracy of estimating SNP effects ($r_{effect}$) (Daetwyler 2009; Goddard 2009). The $r_{LD}$ depends on the strength of LD between QTL and SNPs; the stronger LD, the higher the proportion of the genetic variance that is captured by the SNPs. The $r_{effect}$ depends on the characteristics of the trait, the population in which the effects are estimated and the population in which the effects are used to predict genomic values. First, we will derive $r_{effect}$ for a training population consisting of two distinct populations, based on the same assumptions as underlying a commonly used prediction equation for single-population genomic prediction. Thereafter, $r_{effect}$ is combined with $r_{LD}$ to account for the proportion of the genetic variance captured by the SNPs to derive the accuracy of multi-population genomic prediction.

Using the assumptions that $M$ independent loci are underlying the trait and that each locus is explaining an equal amount of the genetic variance, Daetwyler et al. (2008) derived the following prediction equation for $r_{effect}$ when considering single-population genomic prediction:

$$r_{effect} = \frac{h^2 N}{h^2 N + M},$$

(1)

in which $h^2$ is the heritability of the trait and $N$ is the number of individuals with phenotypes and genotypes included in the training population. The original derivation of this equation is rather complex and difficult to extend to multi-population genomic prediction. As shown by Wientjes et al. (2015b), the same equation can also be derived by partitioning the variance of the average phenotype of $N$ individuals into a part explained by one locus ($\sigma_a^2 / M$) and a part
not explained by that locus \( \left( \frac{\sigma_p^2 - (\sigma_a^2 / M)}{N} \right) \), in which \( \sigma_a^2 \) is the total genetic variance and \( \sigma_p^2 \) is the phenotypic variance. In general, the accuracy of predicting an effect is equal to the square root of the proportion of the total variance explained by that effect (Appendix A provides a formal proof that this result applies to estimation of gene effects). So, the accuracy of predicting the effect of one locus equals:

\[
 r_{\text{locus}} = \sqrt{\frac{\left( \frac{\sigma_a^2}{M} \right)}{\sigma_a^2 / M} \left( \frac{\sigma_p^2 - (\sigma_a^2 / M)}{N} \right)}.
\]  

(2)

Since each locus is assumed to explain only very little variance, \( \sigma_p^2 - (\sigma_a^2 / M) \approx \sigma_p^2 \). Due to the assumption that each locus explains an equal amount of the genetic variance, the accuracy of estimating the effect of one locus is the same for each of the loci, and represents the overall accuracy of estimating SNP effects (see Appendix A):

\[
 r_{\text{effect}} = \sqrt{\frac{\sigma_a^2 / M}{\sigma_a^2 / M + \sigma_p^2 / N}} = \sqrt{\frac{h^2 N}{h^2 N + M}}.
\]  

(3)

Thus, this approach results in the same equation to predict the accuracy as derived by Daetwyler et al. (2008). The derivation described in Equations 2 and 3 is, however, much simpler, and this derivation will be extended to derive the accuracy of multi-population genomic prediction.

Similar to Daetwyler et al. (2008), we assume that \( M \) independent loci are underlying the trait and that each locus explains an equal amount of the genetic variance. The effects of the loci might be different in each population, which is measured by the genetic correlation between populations. Furthermore, we will assume that \( N_A \) individuals from population A and \( N_B \) individuals from population B with phenotype and genotype information are combined
into one training population to estimate SNP effects. Those estimated SNP effects are then
used to predict genomic values of individuals from population C, that could be a sample from
one of the training populations or could be from a different population. The information from
populations A and B, used to estimate SNP effects, is combined in a selection index approach
(Hazel 1943), using the average phenotype of \( N_A \) individuals from population A (\( x_A \)) and the
average phenotype of \( N_B \) individuals from population B (\( x_B \)) as records, and the genomic
values of individuals from population C as breeding goal traits:

\[
I_i = \hat{g}_{C_i} = b_A x_A + b_B x_B, \tag{4}
\]

in which \( b_A \) and \( b_B \) are the regression coefficients on the average phenotype of individuals
from population A (\( x_A \)) and B (\( x_B \)) to predict genomic values for individual \( i \) from population
C (\( \hat{g}_{C_i} \)).

The regression coefficients of genomic values of individuals from population C on the
average phenotype of population A and B can be calculated as:

\[
b = \begin{bmatrix} b_A \\ b_B \end{bmatrix} = P^{-1} g, \tag{5}
\]

in which \( P \) is the (co)variance-matrix of \( x_A \) and \( x_B \) and \( g \) is a vector with covariances between
\( x_A \) and \( x_B \) and the true genomic value of individual \( i \) from population C (\( g_{C_i} \)):

\[
P = \begin{bmatrix} \text{Var}(x_A) & \text{Cov}(x_A, x_B) \\ \text{Cov}(x_A, x_B) & \text{Var}(x_B) \end{bmatrix}, \tag{6}
\]

and:

\[
g = \begin{bmatrix} \text{Cov}(x_A, g_{C_i}) \\ \text{Cov}(x_B, g_{C_i}) \end{bmatrix}. \tag{7}
\]

In analogy with Wientjes et al. (2015b), the variance of the average phenotype of \( N_A \)
individuals can be partitioned into a part explained by one locus \( \left( \sigma_{a_A}^2 / M \right) \) and a part not
explained by that locus $\left( \frac{\sigma^2_{p_A} - \left( \sigma^2_{a_A} / M \right)}{N_A} \right) \approx \frac{\sigma^2_{p_A}}{N_A}$, in which $\sigma^2_{p_A}$ is the total genetic variance in population $A$ and $\sigma^2_{p_A}$ is the total phenotypic variance in population $A$. So, the total variance of $x_A$ can be written as:

$$Var(x_A) = \frac{\sigma^2_{a_A}}{M} + \frac{\sigma^2_{p_A}}{N_A}.$$  \hfill (8)

Note that $\frac{\sigma^2_{p_A}}{N_A}$ represents the part of the phenotypic variance not explained by that locus, i.e., the residual variance ($\sigma^2_{r_A,j}$) for one locus $j$.

The covariance between the average phenotypes in the two populations can be partitioned in a part explained by one locus, a part not explained by that locus and twice the covariance between the two parts. In an additive model, $Cov(a,e) = 0$ and the parts not explained by a locus, i.e. the residual variances, are expected to be independent across populations, indicating that only the covariance between the populations of the part explained by one locus is assumed to differ from zero. Therefore, the covariance can be written as:

$$Cov(x_A, x_B) = r_{G_{A,B}} \cdot \frac{\sigma^2_{a_A} \sigma^2_{a_B}}{M},$$ \hfill (9)

in which $\sigma^2_{a_A}$ and $\sigma^2_{a_B}$ are the genetic standard deviations in respectively population $A$ and $B$ and $r_{G_{A,B}}$ is the genetic correlation between population $A$ and $B$. Hence:

$$P = \begin{bmatrix} \frac{\sigma^2_{a_A}}{M} & \frac{\sigma^2_{a_A}}{M} & \frac{\sigma^2_{a_A} \sigma_{a_B}}{M} \\ \frac{\sigma^2_{a_B}}{M} & \frac{\sigma^2_{a_B}}{M} & \frac{\sigma^2_{a_B} \sigma_{a_A}}{M} \\ r_{G_{A,B}} \frac{\sigma^2_{a_A} \sigma_{a_B}}{M} & r_{G_{A,B}} \frac{\sigma^2_{a_B} \sigma_{a_A}}{M} & r_{G_{A,B}} \frac{\sigma^2_{a_A} \sigma_{a_B}}{M} + \frac{\sigma^2_{p_B}}{N_B} \end{bmatrix},$$ \hfill (10)

in which $\sigma^2_{a_B}$ is the total genetic variance in population $B$ and $\sigma^2_{p_B}$ is the total phenotypic variance in population $B$. 


Since an additive model is assumed, the covariance between the average phenotype of population \( A \) and the true genomic value of individual \( i \) from population \( C \) is also equal to the covariance between the populations of the part explained by one locus:

\[
\text{Cov}(x_A, g_{Ci}) = r_{G_{A,C}} \frac{\sigma_{a_c}}{M},
\]

in which \( \sigma_{a_c} \) is the genetic standard deviation in population \( C \) and \( r_{G_{A,C}} \) is the genetic correlation between population \( A \) and \( C \). Hence:

\[
g = \begin{bmatrix} r_{G_{A,C}} \frac{\sigma_{a_A} \sigma_{a_C}}{M} \\ r_{G_{B,C}} \frac{\sigma_{a_B} \sigma_{a_C}}{M} \\ \end{bmatrix},
\]

in which \( r_{G_{B,C}} \) is the genetic correlation between population \( B \) and \( C \). Substituting Equations 10 and 12 in Equation 5 results in:

\[
b = P^{-1} g = \begin{bmatrix}
\frac{\sigma_{a_A}^2}{M} + \frac{\sigma_{p_A}^2}{N_A} & \frac{\sigma_{a_A} \sigma_{a_B}}{M} \\
\frac{\sigma_{a_A} \sigma_{a_B}}{M} & \frac{\sigma_{a_B}^2}{M} + \frac{\sigma_{p_B}^2}{N_B}
\end{bmatrix}^{-1}
\begin{bmatrix}
r_{G_{A,C}} \frac{\sigma_{a_A} \sigma_{a_C}}{M} \\
r_{G_{B,C}} \frac{\sigma_{a_B} \sigma_{a_C}}{M}
\end{bmatrix}.
\]

With some algebra (see Appendix B), it can be shown that the accuracy of this selection index, representing the accuracy of estimating SNP effects, equals:

\[
r_{HI} = r_{effect} = \frac{b'g}{\sigma^2_H} = \frac{g'P^{-1}g}{\sigma^2_{a_c}/M}.
\]

When only one population is included in the training population, Equation 14 reduces to:
This equation is equivalent to the equation of Wientjes et al. (2015b) for across-population genomic prediction. When estimated SNP effects are applied in another subset of the same population as the training population, i.e., $r_{G_{A,C}}$ is 1, Equation 15 becomes equivalent to the equation derived by Daetwyler et al. (2008) to predict the accuracy of estimating SNP effects within a population (Equation 1).

As explained before, the accuracy of genomic prediction depends on $r_{\text{effect}}$ as well as on $r_{LD}$, accounting for the proportion of the genetic variance captured by the SNPs. It might for example be that the SNP effects are accurately estimated ($r_{\text{effect}}$=1), but when LD between QTL and SNPs is not complete, not all genetic variance can be captured by the SNPs and the accuracy of genomic prediction is still not 1. Moreover, when a number of QTL is segregating in the predicted population and not in the training population, part of the genetic variance in the predicted population can never be captured by the SNPs in the training population.

Altogether, this indicates that the proportion of the genetic variance in the predicted population that can be captured by the SNPs in the training population is specific for a combination of training and predicted population. Therefore, $r_{LD}$ affects the covariance between the phenotypes in the training population and the aggregated genotype of the predicted individuals (Equation 12), which results in:

$$
\mathbf{g} = \begin{bmatrix}
  r_{LD_{A,C}} & r_{G_{A,C}} \frac{\sigma_{gA} \sigma_{gC}}{M} \\
  r_{LD_{B,C}} & r_{G_{B,C}} \frac{\sigma_{gB} \sigma_{gC}}{M}
\end{bmatrix},
$$

in which $r_{LD_{A,C}}$ is the square root of the proportion of the genetic variance in predicted population $C$ captured by the SNPs in training population $A$, and $r_{LD_{B,C}}$ is the square root of...
the proportion of the genetic variance in predicted population $C$ captured by the SNPs in training population $B$. Using Equation 16 instead of Equation 12 in the remaining part of the derivation results in the following equation to predict the accuracy of genomic prediction:

$$r_{EGV} = \left( \frac{h_A^2 + 1}{M} \frac{N_A}{r_{G_{A,B}}} \frac{1}{M} \frac{r_{G_{A,B}}}{N_B} \right)^{-1} \left( \frac{h_A^2}{M} \frac{r_{LD_{A,C}}}{M} \frac{r_{G_{A,C}}}{h_A^2} \right) \cdot \left( \frac{h_B^2}{M} \frac{r_{LD_{B,C}}}{M} \frac{r_{G_{B,C}}}{h_B^2} \right).$$

(17)

In this study, $r_{LD_{A,C}}$ and $r_{LD_{B,C}}$ were assumed to be characteristics of the training and predicted populations, and depending on the SNP density and the properties of the QTL underlying the trait. Therefore, an empirical approach was needed to estimate values for $r_{LD_{A,C}}$ and $r_{LD_{B,C}}$. The values were estimated in the scenarios when only one population ($A$ or $B$) was used as training population, by calculating $r_{LD}$ as $r_{LD} = \frac{r_{EGV}}{r_{effect}}$, in which $r_{EGV}$ was the empirical accuracy and $r_{effect}$ the predicted accuracy assuming all genetic variance in the predicted population was captured by the SNPs. The empirically estimated values for $r_{LD_{A,C}}$ and $r_{LD_{B,C}}$ were used to predict the accuracy when population $A$ and $B$ were combined in the training population to predict genomic values for individuals from population $C$.

**Derivation of $M_e$ to replace $M$**

An important assumption underlying the derived equation is that $M$ independent loci are underlying the trait. In a finite population, loci do not segregate independently due to linkage disequilibrium between loci. The equation predicting the accuracy of SNP effects using a single population (Equation 1), derived by Daetwyler et al. (2008), accounts for that by replacing $M$ by the effective number of chromosome segments, $M_e$, in the population.
(Daetwyler et al. 2010). The $M_e$ within a population is a statistical concept, and can be interpreted as the effective number of chromosome segments that are independently segregating in that population. In other words, it represents the effective number of effects that has to be estimated to predict genomic values for individuals from that population. In the derived equation for multi-population genomic prediction, different populations are combined in the training population, each with different values for $M_e$. For predicting genomic values for individuals from population $C$, using estimated SNP effects in population $A$, the effective number of estimated effects is equal to the effective number of chromosome segments shared between population $A$ and $C$ ($M_{e,A,C}$). Equivalently, when estimated SNP effects in population $B$ are used, the effective number of estimated effects is equal to the effective number of chromosome segments shared between population $B$ and $C$ ($M_{e,B,C}$). In analogy of $M_e$ within a population, the $M_e$ across populations can be interpreted as the effective number of segments that are segregating in a combined population, when considering the differences in LD between the populations. Therefore, we propose the following adjustment to Equation 17:

$$r_{EGV} = \frac{h_A^2}{M_{e,A,C}} + \frac{1}{N_A}r_{G,A,B} \sqrt{\frac{h_B^2}{M_{e,B,C}}} \left[ r_{LD_{A,C}}r_{G,A,C} \frac{h_A^2}{M_{e,A,C}} + r_{LD_{B,C}}r_{G,B,C} \frac{h_B^2}{M_{e,B,C}} \right]^{-1}.$$ (18)

The same equation can also be derived when a selection index is used combining estimated genomic values for individuals from population $C$ based on training populations of respectively population $A$ or $B$, as is shown in Appendix C.

The $M_e$ within a population can be calculated as (Goddard et al. 2011):

$$M_e = \frac{1}{\text{Var}(G_y - E(G_y))},$$ (19)
in which $G_{ij}$ contains the genomic relationship and $E(G_{ij})$ the expected values for the genomic relationships between all individuals $i$ and $j$ from that population, with the variance taken over all pair-wise relationships between individuals $i$ and $j$. In analogy to Equation 19, the values for $M_e$ across populations can be calculated using (Wientjes et al. 2015b):

$$M_{e_{ij}} = \frac{1}{\text{Var}(G_{\text{Pop.1, Pop.2},ij} - E(G_{\text{Pop.1, Pop.2},ij}))}. \quad (20)$$

in which $G_{\text{Pop.1, Pop.2},ij}$ contains the genomic relationships and $E(G_{\text{Pop.1, Pop.2},ij})$ contains the expected genomic relationships between all individuals $i$ from population 1 and individuals $j$ from population 2, again with the variance taken over all pair-wise relationships between individuals $i$ and $j$. The genomic relationships can be calculated following Yang et al. (2010), by calculating the genomic relationships between individual $i$ from population $y$ and individual $j$ from population $z$ as

$$G_{yij} = \frac{1}{n} \sum_k G_{(y, z),k} = \frac{1}{n} \sum_k \frac{(x_{yk} - 2p_{yk})(x_{zik} - 2p_{zk})}{\sqrt{2p_{yk}(1 - p_{yk})\sqrt{2p_{zk}(1 - p_{zk})}}}$$

and the genomic relationship of individual $i$ from population $y$ with itself as

$$G_{yi} = \frac{1}{n} \sum_k G_{(y),k} = 1 + \frac{1}{n} \sum_k \frac{x_{yk}^2 - (1 + 2p_{yk})x_{yk} + 2p_{yk}^2}{2p_{yk}(1 - p_{yk})},$$

in which $n$ is the number of SNPs, $x_{yk}$ and $x_{zik}$ are the genotypes at locus $k$ coded as 0, 1, and 2, and $p_{yk}$ and $p_{zk}$ are the allele frequencies for the second allele (with homozygote genotype coded as 2) at locus $k$ for respectively population $y$ and $z$. The genomic relationships used to calculate $M_e$ are based on population-specific allele frequencies to ensure that unrelated individuals have an expected genomic relationship of 0, which is an underlying assumption of the equation to calculate $M_e$ (Goddard et al. 2011).

In most human studies, individuals included in the data are unrelated (e.g., Yang et al. 2010; Lee et al. 2012; Maier et al. 2015). This indicates that all expected genomic relationships ($E(G)$) would approximately be zero and Equation 20 simplifies to
\[ M_{e_{1,2}} = \frac{1}{\text{Var}(G_{\text{pop.1}, \text{pop.2}})} \]. In most livestock studies, individuals are related, and \( E(G) \) could be approximated by the pedigree relationship matrix \( A \), i.e.,
\[ M_{e_{1,2}} = \frac{1}{\text{Var}(G_{\text{pop.1}, \text{pop.2}} - A_{\text{pop.1}, \text{pop.2}})} \]. When the \( G \) and \( A \) matrix are used to calculate \( M_e \), both matrices should be scaled to the same base population. This can be achieved by rescaling the inbreeding level in \( G \) to the inbreeding in \( A \), for example by using the following adjustment separately for each of the within-population and across-population blocks (Powell et al. 2010):
\[ G^* = (1 - F_b)G + 2F_bJ, \] (21)

in which \( F_b \) is the average pedigree inbreeding level of individuals in population \( b \) and \( J \) is a matrix filled with ones.

The \( G\)-\( E(G) \) values are expected to follow a normal distribution around zero for each value of \( E(G) \). The pedigree relationships between individuals in \( A \), however, depend on the depth of the pedigree for both individuals. In general, the pedigree relationships will more closely resemble \( E(G) \) when the pedigree is deeper. When the pedigree is not deep or complete enough for all or a subset of the individuals, extra variation in \( G\)-\( A \) is introduced, resulting in an underestimation of \( M_e \) when \( A \) is used to represent \( E(G) \). The impact of an insufficient pedigree depth on the calculated \( M_e \) can be reduced by only taking the relationships of individuals with the most complete pedigree into account to calculate \( M_e \). To check if selecting those individuals indeed minimized the impact of an insufficient pedigree depth, values of \( G\)-\( A \) can be plotted versus values of \( A \). When the values for \( G\)-\( A \) are lower for higher \( A \) values, as is shown in Figure 1, an insufficient pedigree depth is still influencing the calculation of \( M_e \). To account for this particular pattern, an exponential function was fitted through the data. For all values of \( A \) in the data, the parameters of the function were estimated.
in R (R Development Core Team 2011) and the fitted values of the function were subtracted from the values of G-A before calculating $M_e$.

**Validation**

After deriving the equation, the aim was to validate it for a broad range of scenarios, differing in heritabilities within populations and genetic correlations between populations. Those scenarios resemble the combining of populations from different environments or measured for different traits. For the validation, real genotypes and simulated phenotypes were used. A pedigree with on average 3.5 complete generations per individual was available, with a minimum of 1 complete generation and a maximum of 9 complete generations. In each of the scenarios, an empirical accuracy was calculated and compared with the predicted accuracy using the derived equation to investigate how accurate the accuracy was predicted. The genotype and pedigree information from all individuals, as well as the simulated phenotypes are available on doi:10.5061/dryad.1525t.

**Genotypes:** Genotypes were available for 1033 dairy cows from the Netherlands, each originating for at least 87.5% from the Holstein Friesian breed, i.e., all animals were pure-bred Holstein Friesians. Genotyping was done using the Illumina BovineSNP50 Beadchip (50k, Illumina, San Diego, CA), after which genotypes were imputed to higher density (777k) using 3150 Holstein Friesian animals as reference population (Pryce et al. 2014). The accuracy of imputation across imputed loci, as reflected by the Beagle $R^2$ value, was on average 0.96, indicating high imputation accuracy. As quality control, SNPs with a call rate smaller than 95%, an unknown mapping position, located on the sex chromosomes, a minor allele frequency (MAF) <0.005, for which only two genotypes were observed, and in complete linkage disequilibrium with a neighboring SNP were deleted. This quality control step reduced the number of SNPs for this study to 422,405.
A total of 50,000 candidate QTL were selected from the 422,405 SNPs, and in each replicate QTL were randomly sampled from the candidate QTL to simulate phenotypes for each individual. The candidate QTL were selected from the SNPs using two different approaches: 1) Candidate QTL were randomly selected (RANDOM), and 2) Candidate QTL were selected from the SNPs with a MAF below 0.2 (LOW MAF), since the MAF of QTL underlying complex traits is expected to be lower than the MAF of SNPs (Goddard and Hayes 2009; Yang et al. 2010; Kemper and Goddard 2012) due to ascertainment bias of the SNPs on the SNP chips (Matukumalli et al. 2009). For each of the two approaches, the remaining 372,405 SNPs were used as markers. In this way, the QTL underlying a trait could be randomly sampled from the candidate QTL in each of the replicates, while the subset of SNP markers was constant across replicates for both RANDOM and LOW MAF.

**Phenotypes:** The 1033 individuals were divided into three groups to represent different populations. The first two groups (population 1 and 2) contained 450 individuals and represented the different training populations (population A and B in the derived equation). The last group (population 3) contained 133 individuals and represented the group of predicted individuals for which genomic values were estimated (population C in the derived equation). The division over the groups was performed using pedigree information, by allocating paternal and maternal half-sib families to the same population. In this way, relationships within a population were higher than between populations, as usually would be expected for distinct populations.

For both the RANDOM and LOW MAF approach of selecting candidate QTL, phenotypes were simulated by randomly sampling 4000 QTL from the group of 50,000 candidate QTL. The QTL underlying the trait were the same in each of the populations. For each QTL, allele substitution effects were sampled from a multivariate normal distribution, with a mean of 0 and standard deviation of 1, using different genetic correlations between the populations.
Only additive effects and no dominance or epistatic interactions were assumed. True genomic values (TGVs) were calculated by multiplying the QTL genotypes, coded as 0, 1 and 2, by the simulated allele substitution effects of the population to which the individual belonged. Across populations, the TGVs were rescaled to a mean of 0 and variance of 1. In each of the populations, the genetic variance was calculated as the variance of the TGVs for the individuals from that population. For all individuals, the environmental effect was sampled from $N(0, \left( \frac{1}{h^2} - 1 \right) \cdot \text{Var}(TGV_i))$, in which $\text{Var}(TGV_i)$ is the variance of TGV in population $i$ to which the individual belonged. For each individual, the simulated TGV and environmental effect was summed to calculate the phenotype.

Scenarios: Seven different scenarios of multi-population genomic prediction were investigated, differing in heritabilities and genetic correlations between the populations (Table 1). The first four scenarios represent multi-environment genomic prediction, where populations in different environments were combined in one training population in which SNP effects were estimated. In those scenarios, the variances were assumed to be homogeneous, i.e., heritability was assumed to be the same in each population (0.95), but genetic correlations between populations varied from 0.4 to 1. The last three scenarios represent multi-trait genomic prediction, where populations measured for different traits are combined in one training population. In those scenarios, variances were assumed to be heterogeneous, i.e., each population had a different heritability of 0.3 or 0.95, and genetic correlations between populations were 0.6 or 1. The values for the heritabilities of 0.3 and 0.95 were chosen to have a clear contrast between the populations.

In each scenario, population 1, population 2, or population 1 and 2 were used as training population and population 3 contained the predicted individuals. Each scenario was analyzed using both approaches of selecting QTL; RANDOM and LOW MAF. Simulations were replicated 100 times in each scenario.
Calculating $M_e$: Values for $M_e$ across the different populations were calculated based on the difference between the genomic and pedigree relationship matrix. Since the subset of SNPs slightly differed between the two approaches of selecting candidate QTL, RANDOM and LOW MAF, values for $M_e$ were calculated for each of the approaches. To reduce the impact of incompleteness of the pedigree, only individuals with at least 3 generations of complete pedigree were taken into account, resulting in 329 individuals in population 1, 270 individuals in population 2, and 90 individuals in population 3. Thereafter, an exponential function was fitted through the data to further reduce the impact of an insufficient pedigree depth, as explained before. The $G$ matrix was the same for all replicates, since the subset of 372,405 SNPs was constant for all replicates while QTL were re-sampled every replicate, resulting in the same $M_e$ for all replicates. Therefore, only one accuracy could be predicted for all replicates of the same approach of selecting candidate QTL, representing the expected average accuracy of estimating SNP effects.

Empirical accuracy of genomic prediction: The empirical accuracies of genomic prediction were obtained both with a single-trait and a multi-trait GBLUP type of model run in ASReml (Gilmour et al. 2009) using the simulated phenotypes and including population as fixed effect. Genomic values for the predicted individuals were estimated using a genomic relationship matrix, $G$, containing all training and predicted individuals, and simulated phenotypes of the training individuals. The $G$ matrix included in the models was calculated using the allele frequencies across all individuals without taking the population into account. The other steps in calculating $G$ were the same as explained above.

In the single-trait model, variances were estimated using REML. Therefore, the model used was termed GREML instead of GBLUP, where variances are assumed to be known. In the single-trait model, the phenotypes of the different populations were pooled in one population, without taking the genetic correlations between the populations into account. The
differences in heritability were, however, taken into account by weighting the phenotypes differently and in this way acknowledging that the phenotypes in one population were more accurately representing the genomic values of the individuals compared to the phenotypes in the other population. It was assumed that the heritability of the phenotypes from the population with the lowest heritability, i.e., a heritability of 0.3, represented the trait heritability based on one measurement. The phenotypes of individuals from this population were given a weight of 1. The heritability of the other population, i.e., a heritability of 0.95, represented the heritability based on multiple measurements of the same trait. In other words, it represented the reliability of the phenotype based on more than one record. This indicates that the genetic variance can be assumed to be the same in both populations. The weight for the phenotypes of individuals from the population with the highest reliability ($r^2$) was equal to the ratio of the residual variances in both populations, which can be calculated as:

$$w = \frac{1-h^2}{h^2/r^2-h^2}.$$  (22)

Following Equation 22, a weight of 44.33 was given to the phenotypes from the population with a heritability of 0.95. One possible scenario where phenotypes could be weighted differently is in dairy cattle populations, where phenotypes of cows are generally based on one single measurement and phenotypes of bulls are based on different numbers of progeny, for which the same weights can be obtained following Garrick et al. (2009).

The multi-trait model considered the phenotypes for the same trait in the different populations as different traits with a genetic correlation between the traits. Estimating all genetic correlations in the multi-trait model was not possible, since phenotypes of the predicted individuals were not included in the model. Therefore, genetic correlations and variance components were assumed to be known and fixed to the simulated values, and the multi-trait model was termed GBLUP.
For each of the models, the accuracy of genomic prediction was calculated as the correlation between the simulated TGVs and predicted genomic values. Note that the single- and multi-trait GBLUP models use both SNP information and simulated phenotypes, that differed across the replicates. Therefore, averages and standard errors across the replicates were calculated and compared to the predicted accuracies.

**Evaluating the potential accuracies of two scenarios**

The derived equation can be used to investigate the accuracy of different scenarios of multi-population genomic prediction. To show this, we used Equation 18 to evaluate the potential accuracy for two specific scenarios, assuming that all genetic variance in the predicted population was captured by the SNPs in the training population ($r_{LD_{A,C}} = r_{LD_{B,C}} = 1$).

The first scenario is relevant for dairy cattle breeding, where bulls with deregressed estimated genetic values based on daughter information are in general used in the training population, with a heritability equal to the reliability of the estimated genetic values. Different studies have investigated the potential to increase the accuracy of genomic prediction by adding cows to the training population with their own phenotypes, that are in general less reliable than estimated genetic values (e.g., Calus et al. 2013; Cooper et al. 2015). In Equation 18 different numbers of cows (range 0 to 50,000) were added to a training population of 10,000 bulls, assuming a heritability of 0.05 for the phenotypes of cows which is representing the heritability of a fertility trait in dairy cattle (e.g., Karoui et al. 2012), different reliabilities (range 0 to 1) for the estimated genetic values of bulls, and a genetic correlation of 1 between the estimated genetic values of bulls and own phenotypes of cows. The values for $M_e$ were set to the values derived from the cattle genotype data used in this study.

The second scenario is based on human studies, in which it was assumed that different numbers of individuals from a population from African descent (range 0 to 100,000) were
added to a training population of 5000 individuals from European descent to increase the accuracy of predicting genetic risk for the European population. As an example, parameters for the trait schizophrenia were used, with a heritability of 0.28 in the European population, a heritability of 0.24 in the African population and a genetic correlation of 0.66 between the populations (De Candia et al. 2013). The \( M_e \) in the European population (\( M_{e,A,C} \) in Equation 18) was set to 43,000, based on the equation \( M_e = \frac{2N_e L}{\ln(4N_e L)} \) (Goddard 2009), an effective population size (\( N_e \)) of 10,000 (McEvoy et al. 2011), and a genome length (\( L \)) of 30 Morgan (Venter et al. 2001). The \( M_e \) across the populations (\( M_{e,b,c} \) in Equation 18) was varied (range 43,000 to 2,000,000).
RESULTS

In this section, the results of the prediction equation are first presented assuming that all genetic variance in the predicted population (population 3) is captured by the SNPs in the training population. Those predicted accuracies were used to calculate $r_{LD_{h,3}}$ and $r_{LD_{z,3}}$ based on the ratio between the empirical and predicted accuracy of genomic prediction when only one of the populations, population 1 or population 2, was used as training population. As a next step, the calculated values for $r_{LD_{h,3}}$ and $r_{LD_{z,3}}$ were used to predict the accuracy of genomic prediction when population 1 and 2 were combined in the training population.

Calculating $M_e$: In Table 2, the different estimated $M_e$ values across populations are shown. Due to only small differences in the subset of SNPs used to calculate $G$, estimated $M_e$ values were very similar for the scenarios with QTL randomly sampled (RANDOM) and QTL sampled with a low MAF (LOW MAF). Using population-specific allele frequencies or allele frequencies across populations only had a very small effect on the estimated values for $M_e$, as well as on the predicted accuracies (range -0.9% - + 1.3%). This indicates that, for this study, the use of population-specific allele frequencies or the allele frequency across populations did not influence the results, due to the very similar allele frequencies across the three populations. Therefore, the predicted accuracies are only shown for the $M_e$ values calculated based on a $G$ matrix using the allele frequencies across the populations.

Scenarios with QTL randomly sampled (RANDOM): In this part, results are presented for the RANDOM scenarios of simulating phenotypes. For those scenarios, the predicted accuracies and average empirical accuracies of genomic prediction obtained with a single-trait model using either a single or combined training population and different scenarios of simulated phenotypes, are shown in Figure 2. The first four scenarios show the accuracies when different genetic correlations between the populations were simulated, with the same heritability in each of the populations. Those scenarios show that when only one population
was used as training population, predicted and empirical accuracies were, as expected, higher when the genetic correlation between training and predicted individuals was higher. There was only a small difference between the accuracies obtained using population 1 or 2 as training population when the genetic correlation with the predicted individuals was the same, because both populations were about equally related to the predicted individuals. Combining the two populations in one training population always resulted in an increase in both predicted and empirical accuracy. The magnitude of the increase in accuracy depended on the genetic correlation between the predicted individuals and the added population; the higher the genetic correlation, the higher the increase in accuracy.

The last three scenarios show the predicted and empirical accuracies using different heritabilities in each of the populations and genetic correlations of 1 or 0.6 between populations. Those scenarios show that when only one population was used as training population, predicted and empirical accuracies were, as expected, higher when the heritability in the training population was higher. For this study, a heritability of 0.3 resulted in approximately 60% of the accuracy obtained with a heritability of 0.95. Adding 450 individuals from the population with a low heritability to a training population of 450 individuals from the population with a high heritability, however, still resulted in an increase in accuracy. The increase in both predicted and empirical accuracy was again lower when the genetic correlation was lower, similar to the scenarios with the same heritability in each population.

For each of the scenarios, the predicted accuracy of genomic prediction shown in Figure 2 is assuming that $r_{LD_{1,3}} = r_{LD_{2,3}} = 1$. In general, predicted accuracies were very slightly overestimating the empirical accuracies of genomic prediction (±1%), both when the heritability was the same in each population and when the heritability was different. When population 1 was used as training population, the overestimation was on average 4% (range
1% – 11%). When population 2 was used as training population, the empirical accuracy was slightly underestimated by the predicted accuracy with on average 8% (range -20% – -2%). When both populations were combined in the training population, the overestimation was on average 6% (range 3% – 12%). Those results indicate that when QTL were randomly sampled from the SNPs, most of the genetic variance in the predicted individuals was tagged by the SNPs in the training population, especially when population 2 was used as training population, and the estimated value for \( r_{LD_{1,3}} \) was 0.96 and for \( r_{LD_{2,3}} \) 1. Using those calculated values to predict the accuracy of genomic prediction for the combined training population reduced the overestimation of the empirical accuracy to 3%.

**Scenarios sampling QTL with low MAF (LOW MAF):** In this part, results are presented for the LOW MAF scenarios of simulating phenotypes. For those scenarios, the predicted and average empirical accuracies of genomic prediction obtained with a single-trait model using either a single or combined training population are shown in Figure 3, assuming \( r_{LD_{1,3}} = r_{LD_{2,3}} = 1 \). All empirical accuracies for the LOW MAF scenarios were lower than the accuracies obtained for the RANDOM scenarios. The predicted accuracies, however, were similar to the predicted accuracies for the RANDOM scenarios. So, the predicted accuracies for the LOW MAF scenarios overestimated the empirical accuracies to a greater extent. On average, the overestimation was ±15%, and again higher when population 1 was used as training population, compared to using population 2 as training population (population 1: 20%; population 2: 7%; combined training population: 20%). Those results indicate that, as expected, a smaller proportion of the genetic variance in the predicted individuals was tagged by the SNPs in the training population when QTL were sampled with a low MAF and the estimated value for \( r_{LD_{1,3}} \) was 0.84 and for \( r_{LD_{2,3}} \) 0.94. Using those calculated values to predict the accuracy of genomic prediction for the combined training population, reduced the overestimation of the empirical accuracy to 5%.
**Single-trait vs multi-trait model:** The analyses using a combined training population were performed using both a single-trait model as well as a multi-trait model, where the same trait in the different populations was modelled as a different correlated trait. The accuracies from both models are shown in Figure 4, for the (A) RANDOM, as well as the (B) LOW MAF scenarios. In this figure, the predicted accuracies for the combined training populations use the estimated values of $r_{LD_{1,3}}$ and $r_{LD_{2,3}}$, estimated when only population 1 or 2 was included in the training population. In general, accuracies obtained with the multi-trait model were equal to or higher than accuracies obtained with the single-trait model, depending on the genetic correlations. When the genetic correlations between both training populations and the predicted population were the same, accuracies obtained with the single- and multi-trait model were similar. When the genetic correlations were different, accuracies obtained with the multi-trait model were higher than accuracies obtained with the single-trait model. Due to those higher empirical accuracies, the overestimation of the empirical accuracy obtained with the multi-trait model by the predicted accuracy of genomic prediction using the estimated values of $r_{LD_{1,3}}$ and $r_{LD_{2,3}}$ reduced on average across replicates to 0% (range -2% to +2%) for the RANDOM scenarios and to 1% (range -2% to +3%) for the LOW MAF scenarios. This indicates that the equation can accurately predict the accuracy of genomic prediction when the proportion of the genetic variance in the predicted population not captured by the SNPs in the training population is known and taken into account.

**The potential accuracies of two scenarios:** The potential accuracies when cows with own phenotypes were added to a training population of 10,000 bulls with deregressed estimated genetic values, is shown in Figure 5, for different numbers of cows added to the training population and different reliabilities for the estimated genetic values. This figure shows that when the reliability of the estimated genetic values of the bulls was low, relatively a small amount of cows had to be added to the training population to see a substantial increase in
accuracy. When the reliability of the estimated genetic values was high (above 0.7), a high accuracy was already obtained with 10,000 bulls in the training population (accuracies were above 0.9), and enlarging the training population by adding cows with own phenotypes only resulted in a minor increase in accuracy.

The potential accuracies for the human scenario where a population from African descent was added to a training population of European descent to predict the genetic risk of individuals from the European population is shown in Figure 6, with different numbers of individuals from the African population added to the training population and different values for $M_e$ across the populations. This figure shows that when $M_e$ across the two populations was low, adding individuals from another population could substantially improve the accuracy of predicting genetic risk. When the $M_e$ across the two populations was large (>20 times the $M_e$ within the European population), adding individuals from the other population only resulted in a minor increase in accuracy. This indicates that to improve the accuracy of predicting genomic values, using training individuals from populations that are more closely related and have a more consistent LD pattern, resulting in lower values for $M_e$ across populations, is more beneficial than using training individuals from populations that are only distantly related.
DISCUSSION

In this paper, a deterministic equation was derived using population parameters to predict the accuracy of genomic values when different populations are combined in the training population. The equation was able to accurately predict the accuracy of multi-environment and multi-trait genomic prediction when the proportion of the genetic variance in the predicted population captured by the SNPs in the training population was known and taken into account. Next to being able to deal with differences in heritability in each population and genetic correlations between populations different from 1, the equation can in principle handle data from more divergent populations as well, such as populations from different environments, breeds or lines. The proportion of the genetic variance captured by the SNPs can, however, be expected to be lower across more divergent populations, as will be discussed later. To confirm that the equation indeed gives accurate predictions for those other scenarios when the proportion of the genetic variance captured by the SNPs is known, further validation of the equation is required using a broader range of populations, preferably with real genotype and phenotype information.

Potential of the derived equation: The equation gives insight in important parameters for multi-population genomic prediction and can be used to compare different scenarios. The equation for example shows that when the $M_e$ across populations is two times higher than $M_e$ within a population, two times more individuals from the other population have to be added to obtain the same increase in accuracy when the heritabilities are the same, the genetic correlations between populations is 1, and all genetic variance can be captured. When those last criteria are not met, even more individuals from the other population have to be added to obtain the same increase in accuracy.

The equation can also be used to investigate the potential accuracy of different scenarios, as was done in Figure 5 and 6. In Figure 6, the equation was applied to a scenario where
human populations from European and African descent were combined in one training
to predict Schizophrenia risk for the European population, a scenario that was
suggested by de Candia et al. (2013). The results show that when the LD pattern is very
different across populations, resulting in a high $M_e$ across populations, it is very unlikely to
see an increase in prediction accuracy, even when a lot of individuals from the other
population are added. Moreover, it shows that the sensitivity of the accuracy for $M_e$ is much
smaller at larger values of $M_e$ across populations compared to small values of $M_e$, which is in
agreement with the results found within a population (Brard and Ricard 2015). Evaluation of
such scenarios requires that estimates for the input parameters, such as the $M_e$ across
predicted and training populations, the heritability of the trait in each of the training
populations, the genetic correlations between the populations ($r_G$), and the part of the genetic
variance in the predicted population captured by the SNPs in the training population ($r_{LD}$)
should, however, be known. Apart from the heritability, for which estimates are
straightforward to calculate, each of the input parameters and how to estimate values for those
parameters will be discussed in more detail in the following paragraphs.

**Effective number of chromosome segments ($M_e$):** In the derived prediction equation, $M_e$
across populations is an important parameter. This parameter can be interpreted as a statistical
concept and represents the effective number of segments that are segregating in a combined
population, which is a measure for the effective number of effects that has to be estimated in
one population to predict genomic values for individuals from another population. It depends
on the consistency in LD between the populations; when the LD pattern is completely
different between the populations, each of the segments has to be very small to segregate in
both populations, resulting in a large $M_e$ across the populations.

It is good to note that the derived equation assumes that $M_e$ segments are underlying the
trait and that each segment explains an equal part of the genetic variance. This indicates that
the equation is basically assuming an infinitesimal model. The GBLUP model is also
assuming an infinitesimal model, therefore the $M_e$ represents the number of effects that have
to be estimated in a GBLUP model and the prediction equation is able to accurately predict
the accuracy from a GBLUP type of model. In a Bayesian variable selection model, the
number of effects that have to be estimated can be lower than $M_e$ for traits where the effective
number of QTL underlying that trait is lower than $M_e$ (Daetwyler et al. 2010; Van den Berg et
al. 2015). This indicates that when the number of QTL is substantially lower than $M_e$ and a
Bayesian variable selection model is used, the number of estimated effects is equal to the
effective number of QTL, which is the value that should be used in the equation to predict the
accuracy of genomic values.

Within a population, the value for $M_e$ can be estimated based on the effective population
size (Goddard 2009; Hayes et al. 2009b; Goddard et al. 2011), as well as using the
relationship matrices based on genomic information and pedigree information (Goddard et al.
2011; Wientjes et al. 2013). For the $M_e$ across populations, it is not possible to use the
equations based on effective population size and a value for $M_e$ can only be estimated based
on the genomic and pedigree relationship matrices. In the prediction equation, however, the
$M_e$ across populations should be known for predicting the accuracy of genetic values before
individuals are genotyped. For those scenarios, it is possible to estimate $M_e$ based on a small
subset of individuals, for example 100 individuals from both populations, for which pedigree
and genotype information is available. Another approach would be to estimate $M_e$ based on
the differences between the populations, since the value for $M_e$ across populations is
depending on the strength of LD between loci (Goddard et al. 2011), which is at least partly
different across populations (Sawyer et al. 2005; De Roos et al. 2008; Veroneze et al. 2013;
Wientjes et al. 2015c). The more divergent the populations are, the higher the value for $M_e$
across populations. In this study, the estimated $M_e$ within a population was around 1350 for all
three populations and the values for $M_e$ across populations were approximately 20% higher. In a study using different closely related cattle breeds, the $M_e$ values across populations were reported to be around 10 times larger than $M_e$ within a population (Wientjes et al. 2015b). This indicates that when very closely related populations are investigated, the $M_e$ across populations can be expected to be around two times the $M_e$ within a population. For closely related breeds, the $M_e$ across populations can be expected to be 10 times the $M_e$ within a population. For distantly related populations, the value for $M_e$ across populations can be even higher.

**Genetic correlation between populations ($r_G$):** Another input parameter is the genetic correlation between the populations, which is the correlation between the allele substitution effects of the QTL. In a simulation study with at least 100 individuals in each of the populations, it was shown that this parameter can accurately be estimated using a genomic multi-trait model, where the same trait in different populations was treated as a different trait (Wientjes et al. 2015b). For closely related populations with an overlapping pedigree, such as populations in different countries that have some common co-ancestry, the genetic correlation can also be estimated using a pedigree relationship matrix (Schaeffer 1994). For more distantly related populations, such as different breeds or lines, the pedigree would probably not be deep enough to capture the relationships across populations and a relationship matrix based on genomic information is required (Karoui et al. 2012; Huang et al. 2014).

**Genetic variance captured by the SNPs ($r_{LD}$):** Results of this study show that the empirical accuracy of genomic prediction was depending on the MAF of the QTL underlying the simulated trait; when QTL had on average a lower MAF than the SNPs, the accuracy reduced. This is in agreement with results of other studies using single- or multi-population genomic prediction (Daetwyler et al. 2013; Wientjes et al. 2015a). The reason for this is a decrease in the strength of LD between QTL and SNPs when the MAF of QTL is lower than
the MAF of SNPs (Khatkar et al. 2008; Yan et al. 2009; Wientjes et al. 2015c), reducing the proportion of the genetic variance captured by the SNPs. As stated before, the MAF of QTL underlying complex traits is expected to be lower than the MAF of SNPs (Goddard and Hayes 2009; Yang et al. 2010; Kemper and Goddard 2012), indicating that it is highly likely that not all the genetic variance can be captured by the SNPs in real data.

The square root of the proportion of the genetic variance captured by the SNPs is represented in the prediction equation as $r_{LD}$, and is depending on the density of the SNP chip, the characteristics of the QTL underlying the trait, and the investigated populations (Daetwyler 2009; Erbe et al. 2013). This parameter can only be estimated based on empirical data, by comparing the predicted and empirical accuracy. Using this approach, $r_{LD}$ was estimated to be around 1 when QTL were randomly sampled from the SNPs and around 0.85 when QTL had a low MAF in this study. In other studies using real data, the square of $r_{LD}$, i.e., $r_{LD}^2$, was estimated to be around 0.8 using a 50k chip in Holstein Friesian dairy populations for Net Merit (Daetwyler 2009) and production traits (Erbe et al. 2013), and slightly lower in Brown Swiss dairy populations for production traits (Erbe et al. 2013; Román-Ponce et al. 2014). The studies estimating $r_{LD}^2$ only focused on one population. Across populations, the value for $r_{LD}$ is supposed to be lower and depending on the number of generations since the separation of the populations; the higher the number of generations, the lower the consistency in LD (e.g., Andreescu et al. 2007; De Roos et al. 2008) and the higher the chance on QTL segregating in only one population (Kemper et al. 2015). Therefore, the values of $\sqrt{0.8} = 0.89$ for $r_{LD}$ found in the empirical studies can probably be seen as the upper limit of $r_{LD}$, which can only be obtained when the predicted and training population are subsets from the same population. The more divergent the predicted and training population are, the lower the value of $r_{LD}$ and the further away the value is from the upper limit of $r_{LD}$ within a population.
Single-trait vs multi-trait model: Empirical accuracies were obtained using both a single-trait model as well as a multi-trait model. The results showed that the use of a multi-trait model was beneficial when the genetic correlation between the two training populations and the predicted population was different. In an empirical study with three different chicken lines with different genetic correlations between populations, a multi-trait model resulted in more or less similar accuracies than a single-trait model (Huang et al. 2014). In an empirical study with three dairy cattle breeds, a multi-trait model using estimated genetic correlations resulted in more or less similar accuracies than a multi-trait model with genetic correlations fixed at 0.95 (Karoui et al. 2012). The combining of dairy cattle populations from three different countries, however, showed a higher accuracy for a multi-trait model compared to a single-trait model (De Haas et al. 2012). So, empirical studies have shown that multi-trait models yield similar or slightly higher accuracies than single-trait models, however, genetic correlations were generally estimated with large standard errors.

The observed increase in accuracy of using a multi-trait model when genetic correlations between the two training populations and the predicted population were different can be explained as follows. When the genetic correlations are different, it is beneficial to take into account that estimated SNP effects from one training population are more related to SNP effects in the predicted population than estimated SNP effects from the other training population. When the genetic correlation was the same, the use of a multi-trait model was not beneficial, even not when the genetic correlation among the training populations was different from 1. This can be explained by the fact that estimated SNP effects in each of the training populations are equally related to SNP effects in the predicted population. In the single-trait model, averages of the SNP effects in both training populations are estimated, which have the same correlation with the SNP effects in the predicted population as the SNP effects in each
of the training populations. Therefore, taking the genetic correlation between the training populations into account had no effect on the obtained accuracy for those scenarios.

**Conclusion:** A deterministic equation is derived to predict the accuracy of genomic values when the training population comprises individuals of different populations, such as populations from different lines or environments, or populations measured for different traits. In this study, the equation was validated for different multi-environment and multi-trait scenarios. Results showed that the accuracy of estimating genomic values can be accurately predicted for those scenarios, provided that the effective number of chromosome segments across predicted and training populations, the heritability of the trait in each of the training populations, the genetic correlations between the populations, and the proportion of the genetic variance in the predicted population captured by the SNPs in the training population are known. Therefore, the derived equation can be used to investigate the potential accuracy of different multi-population genomic prediction scenarios and to decide on the most optimal design of training populations.
ACKNOWLEDGEMENTS

The authors are thankful for useful comments from Chris Schrooten and Henk Bovenhuis.

This study was financially supported by Breed4Food (KB-12-006.03-005-ASG-LR), a public-private partnership in the domain of animal breeding and genomics, and CRV (Arnhem, The Netherlands). The RobustMilk project and the National Institute of Food and Agriculture (NIFA) are acknowledged for providing the 50k genotypes of the HF cows, and the gDMI consortium is acknowledged for imputing those to 777k genotypes.
LITERATURE CITED


Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan, et al., 2014 Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction. Genetics 198: 3-16.


Pryce, J. E., J. Johnston, B. J. Hayes, G. Sahana, K. A. Weigel, et al., 2014 Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations. J. Dairy Sci. 97: 1799-1811.


Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt, et al., 2003 Effects of DGAT1
variants on milk production traits in German cattle breeds. J. Anim. Sci. 81: 1911-1918.

population genomic prediction scenarios in which Bayesian variable selection outperforms
GBLUP. Submitted to BMC Genet.


Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, et al., 2001 The sequence of
the human genome. Science 291: 1304-1351.

Veroneze, R., P. S. Lopes, S. E. F. Guimarães, F. F. Silva, M. S. Lopes, et al., 2013 Linkage

disequilibrium and family relationships on the reliability of genomic prediction. Genetics
193: 621-631.


Empirical and deterministic accuracies of across-population genomic prediction. Genet.
Sel. Evol. 47: 5.


FIGURE 1
The genomic minus pedigree relationships ($G-A$) versus the pedigree relationships ($A$) for across population elements between individuals of two populations. The red line is the fitted exponential function ($f = a + 1/e^{bx+c}$) used to correct $G-A$ values to reduce the impact of an insufficient pedigree depth.
FIGURE 2

Predicted and empirical accuracies of genomic prediction (± standard errors) using a single-trait model, one or two populations in the training population, QTL randomly sampled from the SNPs, and assuming in the prediction equation that all genetic variance in the predicted population was captured by the SNPs in the training population. The different scenarios represent the different genetic correlations and heritabilities used to simulate phenotypes. The scenarios starting with HOM have homogeneous variances in both training populations, the scenarios starting with HET have heterogeneous variances. For each scenario, HOM or HET is followed by the genetic correlation between population 1 and 3, and the genetic correlation between population 2 and 3.
Predicted and empirical accuracies of genomic prediction (± standard errors) using a single-trait model, one or multiple populations in the training population, QTL sampled with a low minor allele frequency (MAF), and assuming in the prediction equation that all genetic variance in the predicted population was captured by the SNPs in the training population. The different scenarios represent the different genetic correlations and heritabilities used to simulate phenotypes. The scenarios starting with HOM have homogeneous variances in both training populations, the scenarios starting with HET have heterogeneous variances. For each scenario, HOM or HET is followed by the genetic correlation between population 1 and 3, and the genetic correlation between population 2 and 3.
Predicted and empirical accuracies of genomic prediction (± standard errors) using a training population consisting of two populations and QTL (A) randomly sampled, or (B) with a low minor allele frequency, and accounting for the proportion of genetic variance in the predicted population captured by the SNPs in the training population in the prediction equation. Empirical accuracies were either obtained with a single-trait model or a multi-trait model. The different scenarios represent the different genetic correlations and heritabilities used to simulate phenotypes. The scenarios starting with HOM have homogeneous variances in both.
training populations, the scenarios starting with HET have heterogeneous variances. For each scenario, HOM or HET is followed by the genetic correlation between population 1 and 3, and the genetic correlation between population 2 and 3.
Predicted accuracies with different numbers of individuals from population 2 added to a training population consisting of 10,000 individuals from population 1 with different heritabilities for the trait. The input parameters represent a scenario in dairy cattle where a cow population with own phenotypes (population 2) was added to a bull population with estimated genetic values based on daughter information (population 1). Due to different numbers of daughters used to estimate genetic values for the bulls, the heritability or reliability of the phenotype in population 1 ranged between 0 and 1. The heritability for the trait in population 2 was 0.05, and genetic correlations between the training populations and between both training populations and the predicted population were 1. The values for $M_e$ were equal to the values in the simulations ($M_{e,1} = 1620, M_{e,2} = 1694$).
FIGURE 6

Predicted accuracies with different numbers of individuals from population 2 added to a training population consisting of individuals from population 1 with different values for the effective number of chromosome segments, $M_e$, across population 1 and 2. The input parameters represent a human scenario where a population from African descent (population 2) was added to a population from European descent (population 1) to predict the genetic risk for Schizophrenia in the European population (population 3 = population 1), with heritabilities of 0.28 in population 1 and 0.24 in population 2 and a genetic correlation of 0.66 between populations 1 and 2 (De Candia et al. 2013). The $M_e$ in population 1 was set to 43,000, based on the equation $M_e = \frac{2N_e L}{\ln(4N_e L)}$ (Goddard 2009) and an effective population size of 10,000 (McEvoy et al. 2011).
### Table 1 - Overview of the different scenarios to simulate phenotypes

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Heritability</th>
<th>Genetic correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pop. 1</td>
<td>Pop. 2</td>
</tr>
<tr>
<td><strong>Homogeneous variances:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOM_1.0-0.6</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>HOM_0.8-0.6</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>HOM_0.8-0.4</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>HOM_0.4-0.4</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td><strong>Heterogeneous variances:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HET_1.0-1.0</td>
<td>0.95</td>
<td>0.30</td>
</tr>
<tr>
<td>HET_1.0-0.6</td>
<td>0.95</td>
<td>0.30</td>
</tr>
<tr>
<td>HET_0.6-1.0</td>
<td>0.95</td>
<td>0.30</td>
</tr>
</tbody>
</table>

1 Scenarios are labeled as follows: The names of the scenarios assuming homogeneous variances in both training populations start with **HOM**, followed by the genetic correlation between population 1 and 3, and the genetic correlation between population 2 and 3. The names of scenarios with heterogeneous variances in the training populations start with **HET**, followed by the genetic correlation between population 1 and 3, and the genetic correlation between population 2 and 3.
Table 2 – Estimated $M_e$ values across populations using population-specific allele frequencies or the allele frequency across populations to set-up $G$.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Population-specific allele frequency</th>
<th>Allele frequency across populations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>QTL with low MAF</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population 1 - 3</td>
<td>1541</td>
<td>1515</td>
</tr>
<tr>
<td>Population 2 - 3</td>
<td>1616</td>
<td>1652</td>
</tr>
<tr>
<td><strong>QTL randomly sampled</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population 1 - 3</td>
<td>1620</td>
<td>1585</td>
</tr>
<tr>
<td>Population 2 - 3</td>
<td>1694</td>
<td>1741</td>
</tr>
</tbody>
</table>
**APPENDIX A**

### Derivation based on random effects model

In the main text, Equation 2 and further were derived by analogy, based on the idea that the accuracy is the square root of the proportion of variance explained by a locus. In this appendix, we provide a proof based on first principles for estimating a random effect.

Consider an additive trait determined by $M$ independently segregating loci, where each locus explains an equal amount of additive genetic variance. The total additive genetic variance equals $\sigma_a^2 = 2M \ p_i (1 - p_i) \ \sigma_{a_i}^2$, where $p_i$ is the allele frequency at the $i^{th}$ locus, and $\sigma_{a_i}^2$ is the variance of the average effect at that locus (this expression is valid, since $p_i (1 - p_i) \ \sigma_{a_i}^2$ is the same for all loci). Thus the variance of the average effect at a locus can be written as:

$$\sigma_{a_i}^2 = \frac{\sigma_a^2}{2M \ p_i (1 - p_i)}.$$  \hspace{1cm} (A.1)

Since loci are independent, the effects at each of the loci can be estimated one at a time. Thus, the average effect at the $i^{th}$ locus can be estimated using a random-effects model,

$$y = z_i a_i + e,$$ \hspace{1cm} (A.2)

in which $y$ is an $N \times 1$ vector with phenotypes corrected for fixed effects for $N$ individuals, $a_i$ is a random genetic effect for locus $i$, $z_i$ is an $N \times 1$ incidence vector with genotypes for all $N$ individuals at locus $i$. Elements of $z_i$ are 0-$2p_i$, 1-$2p_i$, and 2-$2p_i$ for the three genotype classes, and $e$ is a vector of residuals. Since each locus explains only a small part of the variance, the residual variance can be approximated as: $\sigma_e^2 = \sigma_p^2 - (\sigma_a^2 / M) \approx \sigma_p^2$, where $\sigma_p^2$ is the total phenotypic variance.

The variance of $y$ follows from:
\[ \text{Var}(y) \approx z_i z_i \sigma_{a_i}^2 + I \sigma_p^2 = z_i z_i \sigma_{a_i}^2 + I \frac{2p_i(1-p_i)M \sigma_a^2}{h^2}, \]  \hspace{1cm} (A.3)\\

in which \(I\) is an \(N \times N\) identity matrix, and \(h^2\) is the heritability.

Following the mixed model equations, the effect of one locus is estimated as:

\[ \hat{a}_i = \left( z_i z_i + \sigma_p^2 \sigma_{a_i}^2 \right)^{-1} z_i y = \left( 2p_i(1-p_i)N + \sigma_p^2 \frac{2p_i(1-p_i)M}{\sigma_a^2} \right)^{-1} z_i y. \]  \hspace{1cm} (A.4)\\

Thus the variance of the estimated effect for one locus equals:

\[ \text{Var}(\hat{a}_i) = \text{Var} \left( \frac{1}{2p_i(1-p_i) \left( N + \frac{M}{h^2} \right)} z_i y \right), \]
\[ = \left[ \frac{1}{2p_i(1-p_i) \left( N + \frac{M}{h^2} \right)} \right]^{-2} z_i \left( z_i z_i \sigma_{a_i}^2 + I \frac{2p_i(1-p_i)M \sigma_a^2}{h^2} \right) z_i, \]
\[ = \left[ \frac{1}{2p_i(1-p_i) \left( N + \frac{M}{h^2} \right)} \right]^{-2} \left[ 2p_i(1-p_i)N \sigma_{a_i}^2 + [2p_i(1-p_i)]^2 NM \sigma_a^2 \right], \]  \hspace{1cm} (A.5)\\

\[ = \frac{N \sigma_{a_i}^2}{N + \frac{M}{h^2}}. \]

With Best Linear Prediction, the accuracy of an estimated random effect follows from the variances of the estimated and true effects (Falconer and Mackay 1996),

\[ r_{\text{effect}} = \frac{\text{Var}(\hat{a}_i)}{\text{Var}(a_i)} = \frac{\left( \frac{N \sigma_{a_i}^2}{N + \frac{M}{h^2}} \right)}{\sigma_{a_i}^2} = \sqrt{\frac{N}{N + \frac{M}{h^2}}} = \sqrt{\frac{Nh^2}{Nh^2 + M}} = \sqrt{\frac{\left( \frac{\sigma_a^2}{M} \right)}{\left( \frac{\sigma_a^2}{M} \right) + \left( \frac{\sigma_p^2}{N} \right)}}. \]  \hspace{1cm} (A.6)
where $\sigma^2_a/M$ is the variance explained by a single locus. This result is equivalent to Equation 3 from the main text, and shows that the accuracy of an estimated gene effect follows from the proportion of variance explained by the locus.

The estimated effects can be used to calculate an estimated genomic value for individual $j$:

$$EGV_j = z_j \hat{a},$$

(A.7)

in which $z_j$ is an $M \times 1$ vector with genotypes for individual $j$ for all $M$ loci (modelled similarly to $z_i$ above), and $\hat{a}$ is an $M \times 1$ vector with estimated effects for all loci.

The true genomic value of an individual equals:

$$TGV_j = z_j a,$$

(A.8)

in which $a$ is a vector with true effects for all loci.

The accuracy of the $EGV$ equals:

$$r_{TGV, EGV} = \frac{\text{Cov}(TGV, EGV)}{\sqrt{\text{Var}(TGV) \text{Var}(EGV)}} = \frac{\text{Cov}(z_j a, \hat{z}_j \hat{a})}{\sqrt{\text{Var}(z_j a) \text{Var}(\hat{z}_j \hat{a})}} = \frac{z_j^t z_j \sigma^2_a}{\sqrt{z_j^t z_j \sigma^2_a \hat{z}_j^t \hat{z}_j \sigma^2_a}} = \frac{\sigma^2_a}{\sigma^2_a} = r_{\text{Effect}}$$

(A.9)

This result shows that, when all loci explain an equal amount of the genetic variance, the accuracy of the $EGV$ is equal to the accuracy of estimating a single locus effect.

The above represents an alternative derivation of the result of Daetwyler et al. (2008), and is conceptually simpler than the original derivation that treats estimated gene effects as both fixed and random.
Deriving the accuracy of estimating SNP effects in a combined training population

The accuracy of the selection index, representing the accuracy of estimating the effect of one locus, can be calculated as:

$$r_{HI} = r_{effect} = \frac{\mathbf{b}' \mathbf{g}}{\sqrt{\mathbf{g}' \mathbf{P}^{-1} \mathbf{g}}}$$

$$= \sqrt{\begin{bmatrix} \frac{\sigma_{aA}^2}{M} & \frac{\sigma_{ab}^2}{M} \\ \frac{\sigma_{ab}^2}{M} & \frac{\sigma_{bb}^2}{M} + \frac{\sigma_{pb}^2}{N_B} \end{bmatrix} \begin{bmatrix} r_{G_{A,C}} & \frac{\sigma_{aA}}{M} \\ \frac{\sigma_{ab}}{M} & \frac{\sigma_{bb}}{M} + \frac{\sigma_{pb}}{N_B} \end{bmatrix}^{-1}}$$

$$= \sqrt{\begin{bmatrix} \frac{\sigma_{aA}^2}{M} & \frac{\sigma_{ab}^2}{M} \\ \frac{\sigma_{ab}^2}{M} & \frac{\sigma_{bb}^2}{M} + \frac{\sigma_{pb}^2}{N_B} \end{bmatrix} \begin{bmatrix} r_{G_{A,C}} & \frac{\sigma_{aA}}{\sqrt{M}} \\ \frac{\sigma_{ab}}{\sqrt{M}} & \frac{\sigma_{bb}}{\sqrt{M}} + \frac{\sigma_{pb}}{\sqrt{N_B}} \end{bmatrix}^{-1}}$$

For simplicity, we will start by referring to the first element of this inversed $\mathbf{P}$ matrix as $A$, to the off-diagonal elements as $B$ and to the last element as $C$. Hence, Equation B.1 can be written as:

$$r_{effect} = \sqrt{\begin{bmatrix} r_{G_{A,C}} & \frac{\sigma_{aA}}{\sqrt{M}} \\ \frac{\sigma_{ab}}{\sqrt{M}} & \frac{\sigma_{bb}}{\sqrt{M}} + \frac{\sigma_{pb}}{\sqrt{N_B}} \end{bmatrix} \begin{bmatrix} A & B \\ B & C \end{bmatrix}^{-1}}$$

$$= \sqrt{\left( r_{G_{A,C}} \frac{\sigma_{aA}}{\sqrt{M}} A + r_{G_{B,C}} \frac{\sigma_{ab}}{\sqrt{M}} B \right) \left( r_{G_{A,C}} \frac{\sigma_{aA}}{\sqrt{M}} B + r_{G_{B,C}} \frac{\sigma_{ab}}{\sqrt{M}} C \right) \left( r_{G_{A,C}} \frac{\sigma_{aA}}{\sqrt{M}} C \right)}$$

The inverse of the $\mathbf{P}$ matrix can be written as:
\[
\begin{align*}
\begin{bmatrix}
\frac{\sigma_{a_a}^2}{M} + \frac{\sigma_{p_a}^2}{N_A} & \frac{\sigma_{a_a} \sigma_{a_b}}{M} \\
\frac{r_{G_a,b}}{M} & \frac{\sigma_{a_b}^2}{M} + \frac{\sigma_{p_b}^2}{N_B}
\end{bmatrix}^{-1}
&= \\
1 & \left\{ \frac{\sigma_{a_a}^2}{M} + \frac{\sigma_{p_a}^2}{N_A} \right\} \left( \frac{\sigma_{a_b}^2}{M} + \frac{\sigma_{p_b}^2}{N_B} \right) - \left( \frac{r_{G_a,b}}{M} \right)^2 \\
&= \left\{ \frac{\sigma_{a_a}^2}{M} + \frac{\sigma_{p_a}^2}{N_A} \right\} \left( \frac{\sigma_{a_b}^2}{M} + \frac{\sigma_{p_b}^2}{N_B} \right) - \left( \frac{r_{G_a,b}}{M} \right)^2
\end{align*}
\]

1059

\[
1060
\begin{align*}
\left[ \frac{\sigma_{a_a}^2}{M} + \frac{\sigma_{p_a}^2}{N_A} \right] & \left( \frac{\sigma_{a_b}^2}{M} + \frac{\sigma_{p_b}^2}{N_B} \right) - \left( \frac{r_{G_a,b}}{M} \right)^2 \\
&= \left[ \frac{\sigma_{a_a}^2}{M} + \frac{\sigma_{p_a}^2}{N_A} \right] \left( \frac{\sigma_{a_b}^2}{M} + \frac{\sigma_{p_b}^2}{N_B} \right) - \left( \frac{r_{G_a,b}}{M} \right)^2
\end{align*}
\]

Hence, Equation B.2 can be written as:

\[
r_{\text{effect}} = \sqrt{ \left( \frac{\sigma_{a_a}^2}{M} + \frac{\sigma_{p_a}^2}{N_A} \right) \left( \frac{\sigma_{a_b}^2}{M} + \frac{\sigma_{p_b}^2}{N_B} \right) - \left( \frac{r_{G_a,b}}{M} \right)^2 } \\
= \sqrt{ \left( \frac{\sigma_{a_a}^2}{M} + \frac{\sigma_{p_a}^2}{N_A} \right) \left( \frac{\sigma_{a_b}^2}{M} + \frac{\sigma_{p_b}^2}{N_B} \right) - \left( \frac{r_{G_a,b}}{M} \right)^2 }.
\]

1063

Dividing both the numerator and the denominator by \( \sigma_{p_a}^2 \) and \( \sigma_{p_b}^2 \), results in:
\[ r_{\text{effect}} = \frac{\sqrt{\frac{h_A^2}{M} \left( \frac{h_B^2}{M} + \frac{1}{N_B} \right) - 2r_{G_{A,C}} \frac{\sqrt{h_B^2}}{\sqrt{M}} r_{G_{A,B}} \frac{\sqrt{h_A^2}}{\sqrt{M}} + r_{G_{B,C}} \frac{h_B^2}{M} \left( \frac{h_A^2}{M} + \frac{1}{N_A} \right)}}{r_{G_{A,C}} \sqrt{\frac{h_A^2}{M}} - r_{G_{B,C}} \sqrt{\frac{h_B^2}{M}}} \]

Since each locus is assumed to explain an equal amount of the genetic variance, the accuracy of estimating the effect of one SNP is the same for each of the SNPs, and represents the overall accuracy of estimating SNP effects \((r_{\text{effect}})\).
Alternative way of deriving the prediction equation

In this section, an alternative derivation of the prediction equation is presented. In this derivation, the estimated genomic values for population C based on two different training populations (population A and population B), are combined in a selection index to calculate the estimated genomic values for population C when the two populations would be combined in one training population. The estimated genomic value for individual $i$ from population C ($EGV_{A,C_i}$) can be calculated using the estimated marker effects in a training population of population A, following:

$$EGV_{A,C_i} = r_{G_{A,C}} \sum_j X_{C,i,j} \hat{\beta}_{A,j}, \quad (C.1)$$

in which $r_{G_{A,C}}$ is the genetic correlation between population A and C, $X_{C,i,j}$ is the genotype of individual $i$ from population C for marker $j$, and $\hat{\beta}_{A,j}$ is the estimated effect of marker $j$ in population A. In an equivalent way, the estimated genomic value for individual $i$ from population C can be calculated using the estimated markers effects in a training population of population B, i.e., $EGV_{B,C_i}$.

Both estimated genomic values, $EGV_{A,C_i}$ and $EGV_{B,C_i}$, can be combined in a selection index to estimate the genomic value for individual $i$ from population C when both population A and B would be combined in the training population ($EGV_{A+B,C_i}$), following:

$$EGV_{A+B,C_i} = b_A EGV_{A,C_i} + b_B EGV_{B,C_i}, \quad (C.2)$$

in which $b_A$ and $b_B$ are the regression coefficients on $EGV_{A,C_i}$ and $EGV_{B,C_i}$ to predict the estimated genomic value for individual $i$ from population C for the combined training population ($EGV_{A+B,C_i}$).
The regression coefficients on $EGV_{AC_i}$ and $EGV_{BC_i}$ that would maximize the estimation of the genomic value for individual $i$ from population $C$ can be calculated as:

$$\mathbf{b} = \begin{bmatrix} b_A \\ b_B \end{bmatrix} = \mathbf{P}^{-1} \mathbf{g},$$

(C.3)

in which $\mathbf{P}$ is the (co)variance-matrix between the information sources $EGV_{AC_i}$ and $EGV_{BC_i}$, and $\mathbf{g}$ is a vector with covariances between the information sources, $EGV_{AC_i}$ and $EGV_{BC_i}$, and the true genomic value for individual $i$ from population $C$ ($TGV_{C_i}$):

$$\mathbf{P} = \begin{bmatrix} \text{Var}(EGV_{AC_i}) & \text{Cov}(EGV_{AC_i}, EGV_{BC_i}) \\ \text{Cov}(EGV_{AC_i}, EGV_{BC_i}) & \text{Var}(EGV_{BC_i}) \end{bmatrix},$$

(C.4)

and:

$$\mathbf{g} = \begin{bmatrix} \text{Cov}(EGV_{AC_i}, TGV_{C_i}) \\ \text{Cov}(EGV_{BC_i}, TGV_{C_i}) \end{bmatrix}.$$  

(C.5)

In the following part, we will assume that the variances of the estimated and true genomic values are scaled, such that the true genomic values in population $C$ have a variance of 1. The variance of the estimated genomic values for population $C$ using population $A$ in the training population is then equal to the reliability of predicting genomic values for population $C$:

$$\text{Var}(EGV_{AC_i}) = r_{EGV_{AC}}^2.$$  

(C.6)

The covariance between $EGV_{AC_i}$ and $EGV_{BC_i}$ can be written as:

$$\text{Cov}(EGV_{AC_i}, EGV_{BC_i}) = \text{Cov} \left( r_{GAC} \sum_{j} X_{C_{ij}} \hat{\beta}_A, r_{GBC} \sum_{j} X_{C_{ij}} \hat{\beta}_B \right)$$

(C.7)

$$= r_{GAC} r_{GBC} \text{Cov} \left( \sum_{j} X_{C_{ij}} \hat{\beta}_A, \sum_{j} X_{C_{ij}} \hat{\beta}_B \right) = r_{GAC} r_{GBC} \text{Cov} \left( \sum \hat{\beta}_A, \sum \hat{\beta}_B \right).$$

The covariance between the effects marker estimated in population $A$ and $B$ can be written as:
Using the path coefficient method as described by Dekkers (2007), it can be shown that the correlation between the estimated marker effects is equal to:

\[ r_{\hat{\beta}_A, \hat{\beta}_B} = r_{G_{A,B}} r_{\text{effect}_A} r_{\text{effect}_B}, \]  

in which \( r_{G_{A,B}} \) is the genetic correlation between population A and B, and \( r_{\text{effect}_A} \) and \( r_{\text{effect}_B} \) are the accuracies of estimating the marker effects in respectively population A and B. The square root of the variance of the estimated marker effects in each of the populations is equal to the accuracy of the estimated marker effects, i.e., \( \sqrt{\text{Var}(\hat{\beta}_A)} = r_{\text{effect}_A} \), therefore:

\[ \text{Cov}\left( \sum_j \hat{\beta}_{A,j}, \sum_j \hat{\beta}_{B,j} \right) = r_{G_{A,B}} r_{\text{effect}_A} r_{\text{effect}_B} r_{\text{effect}_B} = r_{G_{A,B}} r_{\text{effect}_A}^2 r_{\text{effect}_B}^2. \]  

And:

\[ \text{Cov}\left( \text{EGV}_{A,C}, \text{EGV}_{B,C} \right) = r_{G_{A,C}} r_{G_{B,C}} r_{G_{A,B}}^2 r_{\text{effect}_A}^2 r_{\text{effect}_B}^2. \]  

The accuracy of estimating marker effects in population A multiplied by the genetic correlation between population A and C equals the accuracy of the estimated genomic values, i.e., \( r_{\text{EGV}_{A,C}} = r_{G_{A,C}} r_{\text{Effect}_A} \), under the assumption that all genetic variance of the predicted population is captured by the training populations. Hence, the covariance can be written as:

\[ \text{Cov}\left( \text{EGV}_{A,C}, \text{EGV}_{B,C} \right) = r_{G_{A,B}}^2 r_{\text{EGV}_{A,C}}^2 r_{\text{EGV}_{B,C}}. \]  

Hence, \( P \) can be written as:

\[
P = \begin{bmatrix}
r_{\text{EGV}_{A,C}}^2 & r_{G_{A,B}}^2 r_{\text{EGV}_{A,C}}^2 r_{\text{EGV}_{B,C}} \\
r_{G_{A,B}}^2 r_{\text{EGV}_{A,C}}^2 r_{\text{EGV}_{B,C}} & r_{G_{A,C}}^2 r_{G_{B,C}}
\end{bmatrix}. \]
The covariance between the estimated genomic values for individual $i$ from population $C$ using population $A$ as training population is also equal to the reliability of predicting genomic values for population $C$, i.e., 

$$
\text{Cov}(\text{EGV}_{A,C}, \text{TGV}_{C}) = r_{\text{EGV}_{A,C}}^2.
$$

Hence, $g$ can be written as:

$$
g = \begin{bmatrix}
r_{\text{EGV}_{A,C}}^2 \\
r_{\text{EGV}_{B,C}}^2
\end{bmatrix}.
$$

(C.14)

Since it is assumed that the variance of the true genomic values in population $C$ is scaled to 1, the accuracy of this selection index, representing the accuracy of estimating genomic values for population $C$ based on a training population of population $A$ and $B$, can be calculated as:

$$
r_{\text{EGV}_{A,B,C}} = \left( \frac{g^TP^{-1}g}{\sigma_{a_c}^2} \right) = \sqrt{g^TP^{-1}g}.
$$

(C.15)

For simplicity, we will start by referring to the first element of matrix $P^{-1}$ as $A$, to the off-diagonal elements as $B$ and to the last element as $C$. Hence, Equation C.15 can be written as:

$$
r_{\text{EGV}_{A,B,C}} = \sqrt{\left( \begin{array}{ccc}
r_{\text{EGV}_{A,C}}^2 & r_{\text{EGV}_{B,C}}^2 \\
r_{\text{EGV}_{A,C}}^2 & r_{\text{EGV}_{B,C}}^2
\end{array} \right)^{-1} \begin{bmatrix}
A & B \\
B & C
\end{bmatrix} \begin{bmatrix}
r_{\text{EGV}_{A,C}}^2 \\
r_{\text{EGV}_{B,C}}^2
\end{bmatrix}}.
$$

(C.16)

The matrix $P^{-1}$ can be written as:
If we assume that all genetic variance in population C can be captured by the SNPs in the training population, the accuracies for each of the populations can be replaced by the corresponding equation to predict the accuracy of genomic prediction (Daetwyler et al. 2008; Daetwyler et al. 2010; Wientjes et al. 2015b):
\[ r_{EGV,A,C} = \sqrt{r_{G,A,C}^2 \frac{h_A^2 N_A}{h_A^2 N_A + M_{e,A,C}}} = \sqrt{\frac{h_A^2}{M_{e,A,C}}} \cdot \frac{h_A^2}{M_{e,A,C}} + \frac{1}{N_A}. \]  

(C.19)

And:

\[ r_{EGV,B,C} = \sqrt{r_{G,B,C}^2 \frac{h_B^2}{h_B^2 + \frac{1}{N_B}}}. \]  

(C.20)

Using this in Equation C.18 results in:

\[
\begin{align*}
    \frac{r_{EGV,A,B,C}}{r_{G,A,C}} & = \frac{r_{G,A,C}^2 \left( \frac{h_A^2}{M_{e,A,C}} \right) - 2r_{G,A,B} r_{G,B,C} + r_{G,B,C}^2}{1 - r_{G,A,B}^2} \\
    & = \frac{r_{G,A,C}^2 \left( \frac{h_A^2}{M_{e,A,C}} + \frac{1}{N_A} \right) + r_{G,B,C}^2}{1 - r_{G,A,B}^2}.
\end{align*}
\]

(B.21)
Multiplying both the numerator and the denominator by \( \frac{h_A^2}{M_{eA,C}} + \frac{1}{N_A} \) and \( \frac{h_B^2}{M_{eB,C}} + \frac{1}{N_B} \), results in:

\[
\frac{r_{E_{G_{V_{A,B,C}}}}}{r_{G_{A,C}}^2 \left( \frac{h_A^2}{M_{eA,C}} \right)^2 + \frac{1}{N_A} r_{G_{A,B,C}}^2 \left( \frac{h_A^2}{M_{eA,C}} r_{G_{B,C}}^2 \left( \frac{h_B^2}{M_{eB,C}} \right) + \frac{1}{N_B} r_{G_{A,B}}^2 \left( \frac{h_B^2}{M_{eB,C}} \right) \right) \right] = \sqrt{ \frac{\sqrt{h_A^2} r_{G_{A,C}} \sqrt{\frac{h_B^2}{M_{eB,C}}}}{\sqrt{M_{eA,C}}} } \left[ \frac{h_A^2}{M_{eA,C}} + \frac{1}{N_A} r_{G_{A,B}}^2 \left( \frac{\sqrt{h_A^2 h_B^2}}{\sqrt{M_{eA,C} M_{eB,C}}} \right) \right]^{-1} \left[ \frac{h_B^2}{M_{eB,C}} + \frac{1}{N_B} r_{G_{B,C}} \left( \frac{\sqrt{h_B^2}}{\sqrt{M_{eB,C}}} \right) \right]. \quad (B.22)
\]

This last equation is equivalent to the equation derived before, using the same assumption that all genetic variance of the predicted population is captured by the SNPs in the training populations.