Predicting the purebred-crossbred genetic correlation from phenotype and genotype data of parental lines in pigs

P. Duenk[1*], Y. C. J. Wientjes[1], P. Bijma[1], M. S. Lopes[2,3], Mario P. L. Calus[1]

[1] Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [2] Topigs Norsvin Research Center, P.O. Box 43, 6640 AA Beuningen, the Netherlands; [3] Topigs Norsvin, 80420-210 Curitiba, Brazil; [*] pascal.duenk@wur.nl

Abstract

In previous work, we theoretically derived expressions for an upper and lower bound of the correlation between purebred and crossbred performance (r_{pc}), using only variance components of the parental purebred lines. In the current study, we validated these expressions in real data of pigs by comparing predicted bounds of r_{pc} with the estimated r_{pc}. We compared three methods to approximate the required variance components. The results suggest that the most useful method is to use ordinary REML estimates. If confirmed in other datasets, this approach may help breeders to predict the value of r_{pc} based only on parental line information, or to determine the relative contributions of genotype by genotype and genotype by environment interactions to the value of r_{pc}. We therefore advise studies estimating r_{pc} with genotype data to also estimate and report genetic variance components within and between the parental lines, estimated as described in this study.

Introduction

The genetic correlation between purebred (PB) and crossbred (CB) performance (r_{pc}) is an important parameter in crossbreeding programs, because the r_{pc} determines the response in CB performance when selection is based on PB performance in the parental lines (Smith 1964; Wei *et al.* 1991). A low value of r_{pc} may therefore indicate the need for collecting CB information when the aim is to improve CB performance.

The r_{pc} can be estimated from phenotypes measured on PB and CB animals that are genetically related, and when the pedigree-based relationships between these animals are known (Wei and van der Werf 1995; Lutaaya *et al.* 2001). In practice, however, collecting CB phenotypes can be costly, and the pedigree of CB animals is often difficult to record. The need for a pedigree can be alleviated by genotyping the PB and CB animals that have phenotypes. Although this approach can lead to accurate estimates of r_{pc} (Duenk *et al.* 2019), it may not justify the large investment that is needed for phenotyping and genotyping the required number of CB animals, in particular because the value of CB information for genetic improvement has not yet been determined. Breeders may therefore benefit from predicting the r_{pc} beforehand, based only on information from the parental lines, which is readily available in ongoing breeding programs.

In a previous study, we derived expressions for bounds of r_{pc} based on true variance components computed from QTL effects and genotypes in the parental lines (Duenk *et al.* 2021). These expressions predict the bounds of r_{pc} when only genotype by genotype interaction (GxG) due to dominance and epistasis, and no genotype by environment interaction (GxE) is present. Although obtaining the required variance components is straightforward when QTL effects and genotypes in the parental lines are known (Duenk *et al.* 2021), it is not yet clear how they can be estimated from phenotypes and marker genotypes in the parental lines. In this study, we compare three methods to approximate the variance components required to predict bounds of r_{pc}. For this purpose, we compared predicted bounds of r_{pc} with the estimated r_{pc} in empirical data of pigs.

Materials & Methods

Data. We used phenotypic and genotypic data from three PB lines and their three-way CB, provided by Topigs Norsvin and Norsvin. In total, we used 17,100 animals from a synthetic sire line (S), 6,611 animals from a Landrace (LR) line, 8,587 animals from a Large-White (LW) line, and 4,173 three-way CB. The LR and LW lines were crossed to produce F1 sows, which were crossed with the sires from S to produce the three-way CB. Purebreds were housed in a nucleus environment, and crossbreds in a commercial environment. Phenotypes were pre-corrected for fixed effects using a larger dataset during the routine genetic evaluation of Topigs Norsvin. Traits included were test growth (TGR), lifetime daily gain (LGR), daily feed intake (DFI), backfat (BFE) and loin depth (LDE). All animals were genotyped with the Illumina 50K SNP chip or the Illumina Custom 25K SNP chip. After quality control, all genotypes were imputed to 50K using Fimpute v2.2 (Sargolzaei *et al.* 2014). Markers with MAF < 0.01 in any of the lines were excluded, yielding a total of 35,595 markers that were used for analyses.

Estimating r_{pc}. We estimated r_{pc} in line S using a bivariate model that treats PB and CB performance as different, but genetically correlated traits. The statistical model was

$$\begin{bmatrix}\mathbf{y}_{\mathrm{PB}}\\ \mathbf{y}_{\mathrm{CB}}\end{bmatrix} = \begin{bmatrix}\mathbf{1}_{\mathrm{PB}} & 0\\ 0 & \mathbf{1}_{\mathrm{CB}}\end{bmatrix}\begin{bmatrix}\mu_{\mathrm{PB}}\\ \mu_{\mathrm{CB}}\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_{\mathrm{PB}} & 0\\ 0 & \mathbf{Z}_{\mathrm{CB}}\end{bmatrix}\begin{bmatrix}\mathbf{u}_{\mathrm{PB}}\\ \mathbf{u}_{\mathrm{CB}}\end{bmatrix} + \begin{bmatrix}\mathbf{e}_{\mathrm{PB}}\\ \mathbf{e}_{\mathrm{CB}}\end{bmatrix}, \quad (1)$$

where **y** is a column vector of corrected phenotypes, μ is the mean, **1** is column vector of 1's, **u** is a column vector of additive genetic values with incidence matrix **Z**, and **e** is a column vector of random residuals. The distribution of additive genetic values for PB and CB performance was

$$\begin{bmatrix}\mathbf{u}_{\mathrm{PB}}\\ \mathbf{u}_{\mathrm{CB}}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma^2_{u_{PB}} & \sigma_{u_{PB}u_{CB}}\\ \sigma_{u_{PB}u_{CB}} & \sigma^2_{u_{CB}}\end{bmatrix} \otimes \mathbf{G}\right), \quad (2)$$

where **G** is the genomic relationship matrix, constructed following Wientjes *et al.* (2017). Estimated variance components were obtained using restricted maximum likelihood (REML) in MTG2 (Lee and van der Werf 2016).

Predicted bounds of r_{pc}. Bounds of r_{pc} were predicted based on parental line information only, using the expressions derived in Duenk *et al.* (2021). The expression for the lower bound of r_{pc} in line S for a three-way CB is

$$r^L_{pc} = \frac{\sigma_{S(S),S(LW)} + \sigma_{S(S),S(LR)}}{\sigma_S\sqrt{\left(\sigma^2_{S(LW)} + \sigma^2_{S(LR)} + 2\sigma_{S(LW),S(LR)}\right)}}, \quad (3)$$

where σ_S is the standard deviation of additive genetic values of individuals in line S for the trait expressed in line S, $\sigma^2_{S(b)}$ is the variance of additive genetic values of individuals in line S for the trait expressed in line *b*, where *b* is LW or LR, and $\sigma_{S(b),S(c)}$ is the covariance between additive genetic values of individuals in line S for the trait expressed in line *b* and *c*. The expression for the upper bound of r_{pc} uses the same parameters, and is

$$r^U_{pc} = \frac{\sigma^2_S + 0.5\sigma_{S(S),S(LW)} + 0.5\sigma_{S(S),S(LR)}}{\sigma_S\sqrt{\left(\sigma^2_S + 0.25\sigma^2_{S(LW)} + 0.25\sigma^2_{S(LR)} + \sigma_{S(S),S(LW)} + \sigma_{S(S),S(LR)} + 0.5\sigma_{S(LW),S(LR)}\right)}}. \quad (4)$$

Approximating variance components. Equations (3) and (4) show that we need genetic variance components of parental lines that are usually not available, such as the additive genetic standard deviation in line S for the trait expressed in line LW ($\sigma_{S(LW)}$). Such a component cannot be estimated directly, because there are no phenotypes of individuals from line S for the trait expressed in line LR. Instead, we approximated these components using three different methods. With the first method (M-GEBV), we obtained genomic estimated breeding values (GEBV) of individuals in line S using univariate genomic analyses for each line separately, and computed the (co)variances between these GEBV to approximate the parameters in (3) and (4). For example, $\sigma_{S(LW)}$ was approximated as the standard deviation of GEBV of individuals in line S for the trait expressed in line LR, and $\sigma_{S(LW),S(LR)}$ was approximated as the covariance between GEBV of individuals in line S for the traits expressed in line LW and line LR. With the second method (M-GEBV-S), we used the same approach, but we corrected for shrinkage of GEBV by dividing the GEBV by the square-root of their reliabilities before using them to approximate the variance components. With the third method (M-REML), we used ordinary genomic REML estimates of variance components within and between lines. For example, $\sigma_{S(LW)}$ was approximated as the ordinary REML estimate of the additive genetic standard deviation in line LW, and $\sigma_{S(LW),S(LR)}$ was approximated as the ordinary REML estimate of the additive genetic covariance between line LW and LR.

Results

The estimated r_{pc} ranged from 0.76 (LGR and DFI) to 0.79 (BFE), with standard errors that ranged from 0.04 (DFI and BFE) to 0.07 (LDE) (Figure 1). With M-GEBV, the estimated r_{pc} was always between the predicted lower and upper bound. However, the difference between the lower and upper bound of r_{pc} (i.e. the range) was large (0.57 on average). Correcting GEBV for shrinkage (M-GEBV-S) resulted in a smaller range (0.40 on average), but the estimated r_{pc} was sometimes higher than the upper bound. Finally, with M-REML, the estimated r_{pc} was usually between the bounds, except for BFE. Furthermore, M-REML resulted in the smallest range (0.22 on average).

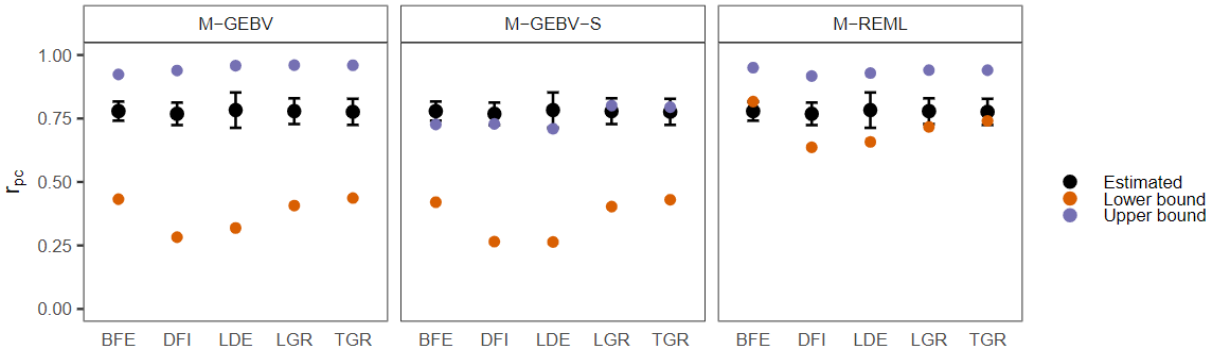

Figure 1 Estimated r_{pc} and predicted bounds of r_{pc} in line S. Facets indicate the method that was used to approximate variance components. BFE=backfat, DFI=daily feed intake, LDE=loin depth, LGR=lifetime daily gain, TGR=test growth rate.

Discussion

In this study, we predicted bounds of r_{pc} based on approximated genetic variance components of parental lines. We compared three methods to approximate the required variance components. The results suggest that the most useful method is M-REML, because this method resulted in the smallest range between the lower and upper bound, and in correct bounds for four out of five traits. For one trait (BFE), the estimated r_{pc} was lower than the predicted lower bound, but this difference was smaller than the standard error of r_{pc}.

Based on theory, the variances and covariances in (3) and (4) should be obtained from true breeding values of S individuals, for the trait expressed in the three parental lines. We therefore expected that method M-GEBV-S would yield the most useful predictions, because this method ensures that (1) approximated variance components refer to the individuals in the focal line, and (2) the GEBV are corrected for shrinkage. The difficulty of this method lies in accurate estimation of GEBV (and their reliabilities) for traits expressed in the mated lines. As a result, estimates of variances and covariances may be inaccurate. In contrast, with M-REML, variance components are estimated directly from phenotypes of the trait of interest, making them more accurate. However, with M-REML, the variance components do not refer to the individuals in the focal line, but to the individuals in the mated line. For the line and traits that were used in this study, M-REML seemed to give more accurate predictions than M-GEBV-S, suggesting that the assumptions of M-REML have a smaller impact on predicted r_{pc} than those of M-GEBV-S.

The expressions to predict bounds of r_{pc} assume that the value of r_{pc} is determined only by non-additive effects in combination with allele frequency differences between parental lines (i.e. GxG interaction), and not by GxE interaction. In this study, GxE interaction may be present, because the PB animals were housed in a nucleus environment, while the CB animals were housed in a commercial environment. Hence, the value of r_{pc} may be lower than 1 due to both GxE and GxG.

In conclusion, we empirically validated that bounds of r_{pc} can be predicted from ordinary REML estimates of variance components in the parental lines. If confirmed in other datasets, this approach may help breeders to predict the value of r_{pc} based only on parental line information, or to determine the relative contributions of GxG and GxE to the value of r_{pc}.

Acknowledgements

The authors thank the Netherlands Organisation of Scientific Research (NWO) and the Breed4Food consortium partners Cobb Europe, CRV, Hendrix Genetics, and Topigs Norsvin for their financial support. Special thanks go out to Norsvin and Topigs Norsvin for providing the data.

References

Duenk, P., P. Bijma, Y.C.J. Wientjes, and M.P.L. Calus, 2021 Genet Sel Evol 53(1):10. https://doi.org/10.1186/s12711-021-00601-w

Duenk, P., M.P.L. Calus, Y.C.J. Wientjes, V.P. Breen, J.M. Henshall *et al.*, 2019 Genet Sel Evol 51(1):6. https://doi.org/10.1186/s12711-019-0447-9

Lee, S.H., and J.H.J. van der Werf, 2016 Bioinformatics 32(9):1420-1422. https://doi.org/10.1093/bioinformatics/btw012

Lutaaya, E., I. Misztal, J.W. Mabry, T. Short, H.H. Timm *et al.*, 2001 J Anim Sci 79(12):3002-3007. https://doi.org/10.2527/2001.79123002x

Sargolzaei, M., J.P. Chesnais, and F.S. Schenkel, 2014 BMC Genomics 15(478. https://doi.org/10.1186/1471-2164-15-478

Smith, C., 1964 Anim Sci 6(3):337-344. https://doi.org/10.1017/S0003356100022133

Wei, M., H.A.M. van der Steen, J.H.J. van der Werf, and E.W. Brascamp, 1991 J Anim Breed Genet 108(1-6):253-261. https://doi.org/10.1111/j.1439-0388.1991.tb00183.x

Wei, M., and J.H. van der Werf, 1995 J Anim Sci 73(8):2220-2226. https://doi.org/10.2527/1995.7382220x

Wientjes, Y.C.J., P. Bijma, J. Vandenplas, and M.P.L. Calus, 2017 Genetics 207(2):503-515. https://doi.org/10.1534/genetics.117.300152