

RESEARCH ARTICLE

Open Access



# Predicting the purebred-crossbred genetic correlation from the genetic variance components in the parental lines

Pascal Duenk<sup>\*</sup> , Piter Bijma, Yvonne C. J. Wientjes and Mario P. L. Calus

## Abstract

**Background:** The genetic correlation between purebred and crossbred performance ( $r_{pc}$ ) is an important parameter in pig and poultry breeding, because response to selection in crossbred performance depends on the value of  $r_{pc}$  when selection is based on purebred (PB) performance. The value of  $r_{pc}$  can be substantially lower than 1, which is partly due to differences in allele frequencies between parental lines when non-additive genetic effects are present. This relationship between  $r_{pc}$  and parental allele frequencies suggests that  $r_{pc}$  can be expressed as a function of genetic parameters for the trait in the parental lines. In this study, we derived expressions for  $r_{pc}$  based on genetic variances within, and the genetic covariance between parental lines. It is important to note that the variance components used in our expressions are not the components that are typically estimated in empirical data. The expressions were derived for a genetic model with additive and dominance effects (D), and additive and epistatic additive-by-additive effects ( $E_{AA}$ ). We validated our expressions using simulations of purebred parental lines and their crosses, where the parental lines were either selected or not. Finally, using these simulations, we investigated the value of  $r_{pc}$  for genetic models with both dominance and epistasis or with other types of epistasis, for which expressions could not be derived.

**Results:** Our simulations show that when non-additive effects are present,  $r_{pc}$  decreases with increasing differences in allele frequencies between the parental lines. Genetic models that involve dominance result in lower values of  $r_{pc}$  than genetic models that involve epistasis only. Using information of parental lines only, our expressions provide exact estimates of  $r_{pc}$  for models D and  $E_{AA}$ , and accurate upper and lower bounds of  $r_{pc}$  for two other genetic models.

**Conclusion:** This work lays the foundation to enable estimation of  $r_{pc}$  from information collected in PB parental lines only.

## Background

Pig and poultry breeders benefit from heterosis and breed complementarity by mating animals from genetically-distinct purebred parental lines to produce crossbred production animals [1, 2]. The aim of such breeding programs is to improve crossbred (CB) performance, while

selection is within the parental lines, usually based on measurements of purebred (PB) performance. As a result, response to selection in CB performance depends partly on the genetic correlation between PB and CB performance ( $r_{pc}$ ), which is generally lower than 1 for most traits in livestock populations [3–9]. Hence,  $r_{pc}$  is an important parameter in breeding programs of pig and poultry.

Estimates of  $r_{pc}$  can be obtained with models that use phenotypic information on both PB and CB performance. Such models require either pedigree information

\*Correspondence: pascal.duenk@wur.nl  
Animal Breeding and Genomics, Wageningen University and Research,  
P.O. Box 338, 6700 AH Wageningen, The Netherlands



© The Author(s) 2021. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

that links the CB to the PB animals [4, 10], or genotype information on both PB and CB animals [11]. Tracking pedigree in a crossbreeding system is often impractical, and collecting phenotypes and genotypes on CB animals may be difficult and costly. Furthermore, breeding companies may produce many different crosses between parental lines, which makes the effort of estimating all relevant  $r_{pc}$  even more costly. To overcome these issues, it would be beneficial if  $r_{pc}$  could be estimated based on data from the parental PB lines, instead of requiring CB data.

The  $r_{pc}$  can be lower than 1 due to (1) differences in trait definition between PB and CB performance [12, 13], (2) genotype-by-environment interactions ( $G \times E$ ) [10, 14], and (3) genotype-by-genotype interactions ( $G \times G$ ) in combination with differences in allele frequencies between the parental lines at loci that affect the trait [9, 15, 16], i.e. the quantitative trait loci (QTL).  $G \times G$  interactions result from non-additive genetic effects (i.e. dominance and epistasis). Here, we consider only the impact of  $G \times G$  interactions on  $r_{pc}$ , assuming that there are no  $G \times E$  interactions. The impact of non-additive effects on  $r_{pc}$  has been studied by Wei et al. [15] and Baumung et al. [16], who derived expressions for  $r_{pc}$  in terms of known additive, dominance, and epistatic genetic effects of loci, and as a function of differences in allele frequencies at these loci between the parental lines. These expressions were, however, limited to one- and two-locus models, and thus cannot be used to predict  $r_{pc}$  for traits that are highly polygenic. Furthermore, genetic effects and allele frequencies at the QTL are usually unknown. Thus, for polygenic traits, there is a need for expressions of  $r_{pc}$  that are based on observable parameters in the parental lines.

Previously, we investigated the impact of non-additive effects on the additive genetic correlation ( $r_g$ ) for a trait between breeding lines [17] and showed that  $r_g$  decreases with increasing size of non-additive effects, and with increasing differences in allele frequencies at QTL between the lines. In the current study, we investigate the impact of non-additive effects on the relationship between  $r_g$  and  $r_{pc}$ .

While estimation of  $r_g$  between two PB lines is relatively straightforward with a genomic relationship matrix [18], the interpretation of the resulting estimates requires careful consideration. Following Duenk et al. [17], we define  $r_g$  between line 1 and line 2 as the correlation between additive genetic values of the individuals in line 1, for the trait expressed in lines 1 and 2. In other words, suppose we know the average effects of all QTL in lines 1 and 2, then we can calculate two additive genetic values for the individuals in line 1; one based on the average effects in line 1, and one based on the

average effects in line 2. The  $r_g$  between lines 1 and 2 for line 1 is the correlation between these two additive genetic values:

$$r_g = \frac{\sigma_{1,1(2)}}{\sqrt{\sigma_1^2} \sqrt{\sigma_{1(2)}^2}} \quad (1)$$

In Eq. (1),  $\sigma_1^2$  is the ordinary additive genetic variance for PB performance in line 1;  $\sigma_{1(2)}^2$  is the additive genetic variance in line 1 for the trait expressed in line 2, which depends on the allele frequencies in line 1 and the average effects for performance in line 2; and, similarly,  $\sigma_{1,1(2)}$  is the additive genetic covariance in line 1 between the trait expressed in line 1 and the trait expressed in line 2, which depends on allele frequencies in line 1 and the average effects for the trait in lines 1 and 2. Note that  $\sigma_{1(2)}^2$  differs from the ordinary additive genetic variance for purebred performance in line 2 (i.e.  $\sigma_2^2$ ). Similarly, the covariance also differs between lines 1 and 2, i.e.  $\sigma_{1,1(2)} \neq \sigma_{2,2(1)}$ ; while both covariances depend on the average effects in both lines,  $\sigma_{1,1(2)}$  depends on the allele frequencies in line 1 and  $\sigma_{2,2(1)}$  depends on the allele frequencies in line 2. Therefore,  $r_g$  for line 2 is a different parameter because it depends on allele frequencies in line 2. Here, we focus on  $r_g$  for line 1, because we are interested in  $r_{pc}$  for line 1.

Our aim was to derive expressions for the prediction of  $r_{pc}$  in a two-way crossbred breeding program, based on genetic variances within the parental lines and the genetic covariance between the parental lines (i.e., the terms in Eq. (1)). The resulting expressions predict the component of  $r_{pc}$  that is due to non-additive effects. Expressions were derived for two genetic models; a genetic model with additive and dominance effects (D), and a genetic model with additive and additive-by-additive ( $A \times A$ ) epistatic effects between pairs of QTL ( $E_{AA}$ ). We validated our expressions using simulations of PB parental lines and their crosses, where the parental lines were either selected or not. Finally, using simulations, we also investigated the value of  $r_{pc}$  for two genetic models for which expressions could not be derived: a model with both dominance and  $A \times A$  epistatic effects ( $D + E_{AA}$ ) and a model with complementary epistatic effects ( $E_C$ ). We compared the results from these models with our predictions of  $r_{pc}$  under models D and  $E_{AA}$ .

## Theory

We consider two PB parental lines (1 and 2) that are mated to produce CB individuals. The additive genetic correlation between PB and CB performance ( $r_{pc}$ ) in line 1 is defined as the correlation between additive genetic values for PB and CB performance of members of line 1 [15, 19]. For PB performance, the additive genetic value of individual  $i$  from line 1 is:

$$v_{i,1} = \mathbf{h}_i' \boldsymbol{\alpha}_1 \tag{2}$$

where  $\mathbf{h}_i$  is a column vector of genotypes of individual  $i$  at all QTL (measured as allele counts minus the average allele count in line 1), and  $\boldsymbol{\alpha}_1$  is a column vector of average effects of allele substitution for PB performance at QTL in line 1. Similarly, the additive genetic value of individual  $i$  for CB performance is:

$$v_{i,C} = \mathbf{h}_i' \boldsymbol{\alpha}_{1(C)} \tag{3}$$

where  $\boldsymbol{\alpha}_{1(C)}$  is a column vector of average effects of allele substitution for CB performance at QTL in line 1. Here and throughout the remainder of this paper, genotypes ( $\mathbf{h}_i$ ) are considered random variables, whereas the average effects ( $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_{1(C)}$ ) are considered fixed.

The  $r_{pc}$  in line 1 is the correlation between the additive genetic values in Eqs. (2) and (3):

$$r_{pc} = cor(v_{i,1}, v_{i,C}) = \frac{cov(v_{i,1}, v_{i,C})}{\sqrt{var(v_{i,1})} \sqrt{var(v_{i,C})}} = \frac{\sigma_{1,1(C)}}{\sigma_1 \sigma_{1(C)}} \tag{4}$$

where  $\sigma_1$  is the additive genetic standard deviation for PB performance in line 1,  $\sigma_{1(C)}$  is the additive genetic standard deviation for CB performance in line 1, and  $\sigma_{1,1(C)}$  is the additive genetic covariance between PB and CB performance in line 1.

Our aim is to express  $r_{pc}$  in terms of genetic parameters in the parental lines. First, we derive expressions for  $\boldsymbol{\alpha}_1$  and  $\boldsymbol{\alpha}_{1(C)}$  for a genetic model with additive and dominance effects (D), and for a model with additive and additive-by-additive epistatic effects ( $E_{AA}$ ). Second, we express  $\boldsymbol{\alpha}_{1(C)}$  in terms of average effects of allele substitution for PB performance in the parental lines ( $\alpha_1$  and  $\alpha_2$ ). Third, we derive expressions for  $\sigma_{1,1(C)}$  and  $\sigma_{1(C)}$ , and finally for  $r_{pc}$ , in terms of genetic parameters in the parental lines.

**Derivation of average effects of allele substitution for PB and CB performance**

The first step is to derive expressions for average effects of allele substitution in line 1 for PB and CB performance. For CB performance, we are interested in the effect of alleles from line 1 on genotypic values of CB offspring (i.e., when line 1 is randomly mated to line 2). Hence, we want to express the average effects of allele substitution in terms of differences between genotypic values of CB offspring. Following Falconer [20], the average effect of an allele is the mean genotypic value of offspring produced by transmitting that allele, minus the mean genotypic value of the population. The average effect of allele substitution at a bi-allelic locus is equal to the difference between the average effects of its alleles.

Strictly speaking, this is the definition of average excess, but it is equivalent to the average effect under random mating [20]. Hence, if individuals of line 1 are mated at random to individuals of line 2, then the average effect and average excess are identical, even though the resulting CB population is not in Hardy–Weinberg equilibrium. In the following, we will use the term ‘average effect’ to refer to the average effect of allele substitution at a locus. Furthermore, we assume that the genetic additive, dominance, and epistatic effects are the same for PB and CB performance, and that for CB performance, these effects are independent of line origin. Statistical additive, dominance and epistatic effects, however, are line-dependent due to differences in allele frequencies. In other words,  $r_{pc}$  values lower than one are the result of  $G \times G$  interaction.

**Dominance model (D)**

Consider a locus that has an additive effect ( $a$ ), a dominance effect ( $d$ ), and no epistatic interactions with other loci. The average effect for PB performance in line 1 under this genetic model (D) is equal to:

$$\alpha_1^D = a + (1 - 2p_1)d \tag{5}$$

where  $p_1$  is the frequency of the focal allele in line 1. The full derivation leading to this result can be found in Falconer and Mackay [19], and the average effect for CB performance in line 1 when mated to line 2 can be derived in a similar way. In contrast to alleles transmitted to PB animals, alleles from line 1 transmitted to crossbreds will always pair with an allele from line 2. Thus, the average effect for CB performance in line 1 under genetic model D depends on the allele frequency in line 2 only [21, 22]:

$$\alpha_{1(C)}^D = a + (1 - 2p_2)d \tag{6}$$

where  $p_2$  is the frequency of the focal allele in line 2. Thus, under genetic model D, the average effect for CB performance in line 1 when mated to line 2 is equal to the average effect for PB performance in line 2 ( $\alpha_2^D$ ) (see also Zeng et al. [23] and Vitezica et al. [24]).

**Additive-by-additive epistasis model ( $E_{AA}$ )**

With additive-by-additive ( $A \times A$ ) epistasis (i.e. genetic model  $E_{AA}$ ), and without dominance or other types of epistasis, the average effect at a locus does not depend on the allele frequency at the focal locus, but on the allele frequency at the loci it interacts with. Consider a locus F with alleles  $F$  and  $f$ , which has an additive effect ( $a$ ), and an  $A \times A$  epistatic interaction with locus G with alleles  $G$  and  $g$ . The epistatic effect between F and G is denoted as  $\epsilon$ . For simplicity of presentation, we assume in the following derivation that locus G has no additive

effect, because the result for locus F does not depend on the additive effect at locus G. In addition, we only considered pairwise interactions between loci, as opposed to interactions between more than two loci. Table 1 shows the genotypic values for the two-locus genotypes, e.g. [25]. The genotypic values are the sum of the additive effect at locus F, and the epistatic effect between loci F and G. The sign in front of the additive effect depends on the genotype at locus F, whereas the sign in front of the epistatic effect depends on the genotype at both F and G.

The average effect at locus F for PB performance in line 1 can be derived by computing the difference between the average effects of alleles *F* and *f*. The average effect of an allele for PB performance of the same line is the mean genotypic value of offspring that inherited that allele from line 1, assuming the other allele was drawn at random from line 1. The average effect of allele *F* in line 1 is:

$$\begin{aligned} \alpha^F &= p_1^F \left( P_1^G(a + \epsilon) + H_1^G(a) + Q_1^G(a - \epsilon) \right) \\ &\quad + p_1^f \left( P_1^G(0) + H_1^G(0) + Q_1^G(0) \right) \\ &= p_1^F \left( (P_1^G - Q_1^G)\epsilon + a \right) \end{aligned}$$

where  $P_1^G$ ,  $H_1^G$ , and  $Q_1^G$  are genotype frequencies at locus G in line 1 (Table 1),  $p_1^F$  is the frequency of allele *F* in line 1, and  $p_1^f = 1 - p_1^F$ . Similarly, the average effect of allele *f* is:

$$\begin{aligned} \alpha^f &= p_1^f \left( P_1^G(0) + H_1^G(0) + Q_1^G(0) \right) \\ &\quad + p_1^F \left( P_1^G(-a - \epsilon) + H_1^G(-a) + Q_1^G(a + \epsilon) \right) \\ &= -p_1^f \left( (P_1^G - Q_1^G)\epsilon + a \right) \end{aligned}$$

The average effect at locus F for PB performance in line 1 under genetic model  $E_{AA}$  then is:

$$\begin{aligned} \alpha_1^{AA} &= \alpha^F - \alpha^f = (p_1^F + p_1^f) \left( (P_1^G - Q_1^G)\epsilon + a \right) \\ &= a + (P_1^G - Q_1^G)\epsilon = a - (1 - 2p_1^G)\epsilon \end{aligned} \tag{7}$$

where  $p_1^G$  is the frequency of allele *G* in line 1. The average effect for PB performance in line 2 can be obtained by using the allele frequency in line 2 in Eq. (7).

The average effect for CB performance in line 1 can be derived similarly using the expected genotype frequencies in CB offspring at locus G. This results in:

$$\alpha_{1(C)}^{AA} = a - (1 - 2p_C^G)\epsilon \tag{8}$$

where  $p_C^G$  is the expected frequency of allele *G* in the CB offspring. Given the expressions for  $\alpha_1^{AA}$  and  $\alpha_2^{AA}$  (Eqs. (7) and (8)), and using  $p_C^G = 0.5(p_1^G + p_2^G)$ , the average effect of CB performance in line 1 under genetic model  $E_{AA}$  can be written as the mean of average effects for PB performance in lines 1 and 2:

$$\alpha_{1(C)}^{AA} = 0.5(\alpha_1 + \alpha_2) \tag{9}$$

**Derivation of  $r_{pc}$**

In the following, we use our derivations of  $\alpha_1$  and  $\alpha_{1(C)}$  for genetic models D and  $E_{AA}$  to derive the additive genetic variance for PB and CB performance in line 1 ( $\sigma_1^2$  and  $\sigma_{1(C)}^2$ ), and the additive genetic covariance between PB and CB performance in line 1 ( $\sigma_{1,1(C)}$ ). Then, we use these derivations in an expression for  $r_{pc}$ . In these derivations, we treat the genotypes of individuals as random. We assume that there is no correlation between average effects at different loci, and that the average effect at a locus is independent from the genotypes at that locus as a result of random allele coding (i.e., which allele is the focal allele is independent of the effect of the allele). Regardless of the genetic model, we define the additive genetic variance for PB performance in line 1 as:

$$\begin{aligned} \sigma_1^2 &= var(\mathbf{h}_1\alpha_1) = var\left(\sum_j h_{ij}\alpha_{1j}\right) \\ &= E\left[\left(\sum_j h_{ij}\alpha_{1j}\right)\left(\sum_j h_{ij}\alpha_{1j}\right)\right] \\ &= \sum_j E(h_{ij}h_{ij})\alpha_{1j}\alpha_{1j} = \sum_j 2p_{1j}(1 - p_{1j})\alpha_{1j}^2 \end{aligned} \tag{10}$$

where *j* denotes the locus.

**Dominance model (D)**

With dominance (model D), average effects for CB performance in line 1 are equal to average effects for PB performance in line 2 (Eq. (6)). Hence, with model D, the additive genetic variance for CB performance in line 1 is:

**Table 1 Genotypic values of two locus (F and G) genotypes with additive-by-additive (A × A) epistasis (model E<sub>AA</sub>)**

Frequency		$p^F$	$H^F$	$Q^F$
		$FF$	$Ff$	$ff$
$p^G$	$GG$	$a + \epsilon$	0	$-a - \epsilon$
$H^G$	$Gg$	$a$	0	$-a$
$Q^G$	$gg$	$a - \epsilon$	0	$-a + \epsilon$

The genotypic values are the sum of the additive effect at locus F, and the epistatic effect between loci F and G. The sign in front of the additive effect depends on the genotype at locus F, whereas the sign in front of the epistatic effect depends on the genotype at loci F and G. For simplicity, it is assumed that G has no additive effect

$p^X, H^X,$  and  $Q^X$  denote the genotype frequencies at locus X,  $a$  denotes the additive effect at locus F, and  $\epsilon$  denotes the epistatic effect between loci F and G

$$\begin{aligned}
 \sigma_{1(C)}^2 &= var(\mathbf{h}'_i \alpha_{1(C)}) = var\left(\sum_j h_{ij} \alpha_{2j}\right) \\
 &= E\left[\left(\sum_j h_{ij} \alpha_{2j}\right)\left(\sum_j h_{ij} \alpha_{2j}\right)\right] \\
 &= \sum_j E(h_{ij} h_{ij}) \alpha_{2j} \alpha_{2j} \\
 &= \sum_j 2p_{1j}(1 - p_{1j}) \alpha_{2j}^2 = \sigma_{1(2)}^2
 \end{aligned}
 \tag{11}$$

where  $\sigma_{\alpha_2}^2$  is the variance of average effects for PB performance in line 2. As mentioned in the Background section, it is important to note that  $\sigma_{1(2)}^2$  is the additive genetic variance in line 1 for the trait expressed in line 2. This is evident from Eq. (11), where  $\sum_j (2p_{1j}(1 - p_{1j}))$  contains allele frequencies in line 1 and  $\alpha_{2j}$  is the average effect at locus  $j$  for the trait expressed in line 2 (see also Background and Discussion).

The additive genetic covariance between PB and CB performance in line 1 can be written as:

$$\begin{aligned}
 \sigma_{1,1(C)} &= cov(\mathbf{h}'_i \alpha_1, \mathbf{h}'_i \alpha_{1(C)}) = cov\left(\sum_j h_{ij} \alpha_{1j}, \sum_j h_{ij} \alpha_{2j}\right) \\
 &= E\left[\left(\sum_j h_{ij} \alpha_{1j}\right)\left(\sum_j h_{ij} \alpha_{2j}\right)\right] = \sum_j E(h_{ij} h_{ij}) \alpha_{1j} \alpha_{2j} \\
 &= \sum_j 2p_{1j}(1 - p_{1j}) \alpha_{1j} \alpha_{2j} = \sigma_{1,1(2)},
 \end{aligned}
 \tag{12}$$

where  $\sigma_{1,1(2)}$  is the additive genetic covariance for individuals in line 1 between the trait expressed in lines 1 and 2.

As a result, with model D,  $r_{pc}$  for line 1 can be written as:

$$r_{pc}^D = cor(\mathbf{h}'_i \alpha_1, \mathbf{h}'_i \alpha_2) = \frac{\sigma_{1,1(2)}}{\sigma_1 \sigma_{1(2)}}
 \tag{13}$$

Hence, for model D,  $r_{pc}$  for line 1 is equal to the correlation between additive genetic values of individuals in line 1 for the trait expressed in parental lines 1 and 2, which is equal to the genetic correlation between lines 1 and 2, as defined in the Background section (Eq. (1)). It is important to note, however, that this is different from the correlation between average effects for PB performance in the parental lines (e.g., Xiang et al. [26]), because Eq. (1) is a weighted correlation between average effects, where weights are computed based on genotype frequencies in line 1.

**Additive-by-additive epistasis (model E<sub>AA</sub>)**

With A × A epistatic interactions (model E<sub>AA</sub>), the average effect for CB performance in line 1 is equal to the mean of the average effects for PB performance in the two parental lines (Eq. (9)). Thus, the additive genetic variance for CB performance in line 1 can be written as:

$$\begin{aligned}
 \sigma_{1(C)}^2 &= var(\mathbf{h}'_i \alpha_{1(C)}) = var\left(\sum_j h_{ij} 0.5(\alpha_{1j} + \alpha_{2j})\right) \\
 &= E\left[\left(\sum_j h_{ij} 0.5(\alpha_{1j} + \alpha_{2j})\right)\left(\sum_j h_{ij} 0.5(\alpha_{1j} + \alpha_{2j})\right)\right] \\
 &= \sum_j E(h_{ij} h_{ij}) (0.5(\alpha_{1j} + \alpha_{2j}))^2 \\
 &= 0.25 \sum_j 2p_{1j}(1 - p_{1j}) (\alpha_{1j}^2 + \alpha_{2j}^2 + 2\alpha_{1j} \alpha_{2j}) \\
 &= 0.25(\sigma_1^2 + \sigma_{1(2)}^2) + 2\sigma_{1,1(2)}
 \end{aligned}
 \tag{14}$$

The additive genetic covariance between PB performance and CB performance in line 1 can be derived as:

$$\begin{aligned}
 \sigma_{1,1(C)} &= cov(\mathbf{h}'_1\alpha_1, \mathbf{h}'_1\alpha_{1(C)}) \\
 &= cov\left(\sum_j h_{ij}\alpha_{1j}, h_{ij}0.5(\alpha_{1j} + \alpha_{2j})\right) \\
 &= E\left[\left(\sum_j h_{ij}\alpha_{1j}\right)\left(\sum_j h_{ij}0.5(\alpha_{1j} + \alpha_{2j})\right)\right] \\
 &= \sum_j E(h_{ij}h_{ij})\alpha_{1j}0.5(\alpha_{1j} + \alpha_{2j}) \\
 &= 0.5 \sum_j 2p_{1j}(1 - p_{1j})\left(\alpha_{1j}^2 + \alpha_{1j}\alpha_{2j}\right) \\
 &= 0.5\sigma_1^2 + 0.5\sigma_{1,1(2)}
 \end{aligned}
 \tag{15}$$

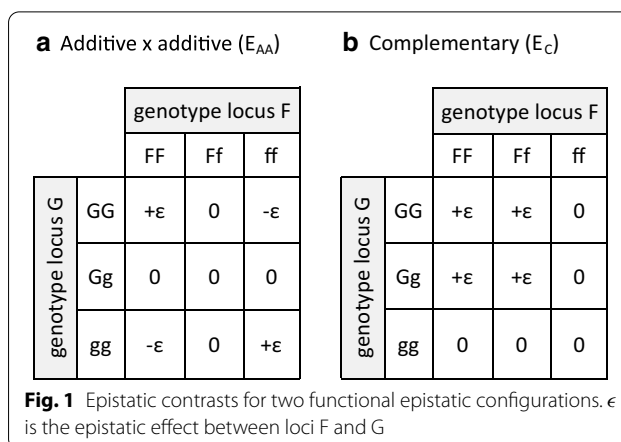
Hence,  $r_{pc}$  for line 1 with genetic model  $E_{AA}$  is equal to:

$$\begin{aligned}
 r_{pc}^{AA} &= \frac{0.5\sigma_1^2 + 0.5\sigma_{1,1(2)}}{\sigma_1\sqrt{0.25(\sigma_1^2 + \sigma_{1(2)}^2 + 2\sigma_{1,1(2)})}} \\
 &= \frac{\sigma_1^2 + \sigma_{1,1(2)}}{\sigma_1\sqrt{(\sigma_1^2 + \sigma_{1(2)}^2 + 2\sigma_{1,1(2)})}}
 \end{aligned}
 \tag{16}$$

Thus, for genetic model  $E_{AA}$ ,  $r_{pc}$  is a function of the additive genetic covariance in line 1 between the trait expressed in line 1 and line 2 ( $\sigma_{1,1(2)}$ ), and the additive genetic variances in line 1 for the trait expressed in line 1 ( $\sigma_1^2$ ) and in line 2 ( $\sigma_{1(2)}^2$ ). The expressions in Eqs. (13) and (16) show that the value of  $r_{pc}$  due to  $G \times G$  interactions can be determined based on genetic parameters within the parental lines.

### Expressions as bounds of $r_{pc}$

As evident from our derivations (e.g. Equations (2) and (3)),  $r_{pc}$  depends on the difference between average effects at QTL for PB and CB performance ( $\Delta_\alpha$ ). With model D, it follows from the difference between Eqs. (5) and (6) that  $\Delta_\alpha$  increases by  $2(p_1 - p_2)$  per unit increase in the magnitude of the dominance effect. This is because with model D,  $\alpha_{1(C)}$  at a locus depends on the allele frequency in the mated line, whereas  $\alpha_1$  depends on the allele frequency in line 1. With model  $E_{AA}$ , in contrast,  $\Delta_\alpha$  increases by  $2(p_1 - p_C) = (p_1 - p_2)$  per unit increase in the epistatic effect, based on the difference between Eqs. (7) and (8). This is because with model  $E_{AA}$ ,  $\alpha_{1(C)}$  of a locus depends on the allele frequency of the interacting locus in the cross, rather than in the mated line. Hence, for each unit increase in the magnitude of non-additive effects,  $\Delta_\alpha$  increases twice as fast with genetic model D as with model  $E_{AA}$ .

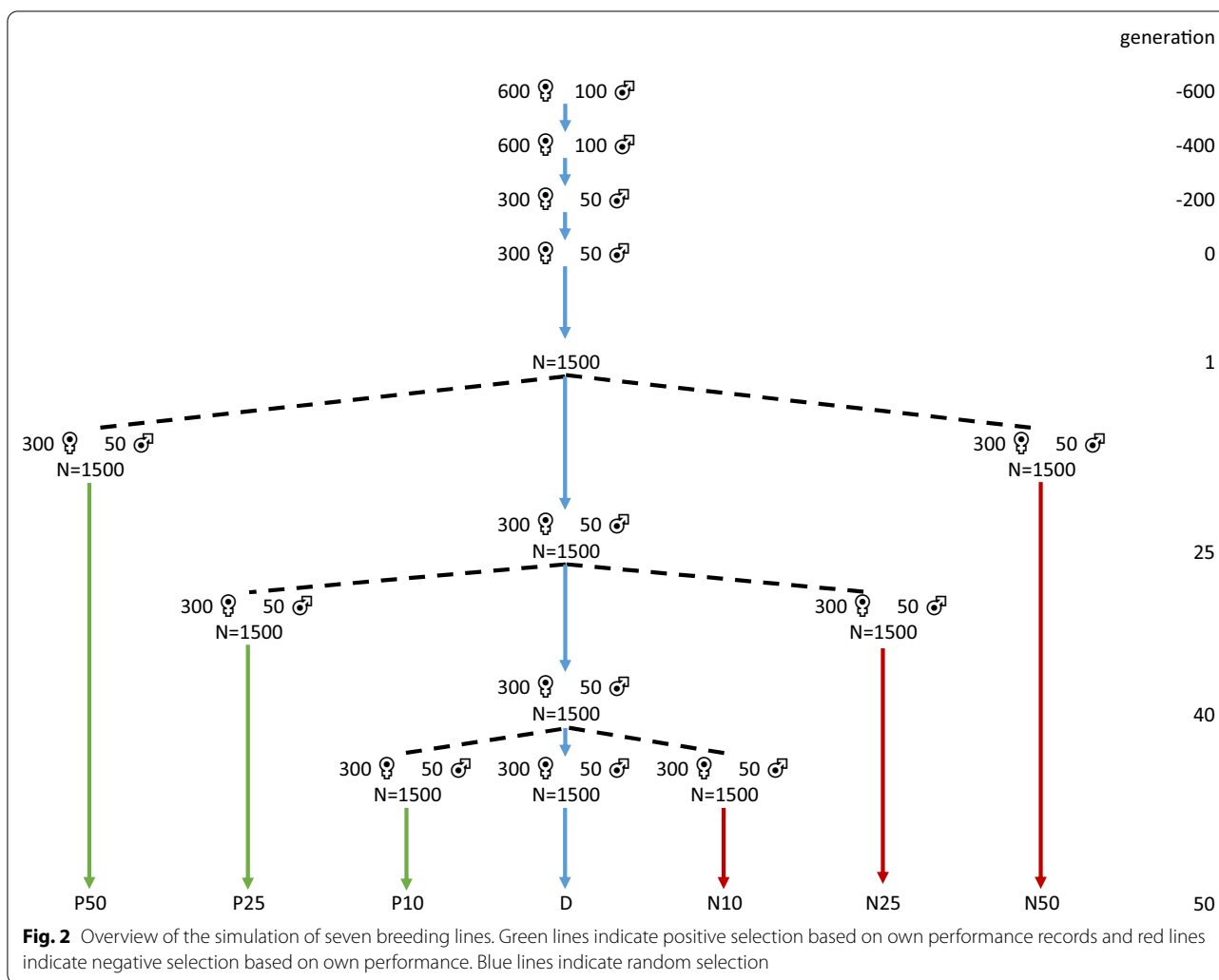


**Fig. 1** Epistatic contrasts for two functional epistatic configurations.  $\epsilon$  is the epistatic effect between loci F and G

Because any non-additive interaction involves either dominance, epistasis, or both, models D and  $E_{AA}$  may represent extremes, where  $r_{pc}$  either depends on (1) the difference in allele frequency between the parental lines (model D), or (2) half of that difference (model  $E_{AA}$ ). With other genetic models,  $r_{pc}$  may depend on (1), (2), or both. However, it is unlikely that other genetic models will lead to much lower  $r_{pc}$  than model D, because the maximum  $\Delta_\alpha$  is bounded by the difference in allele frequencies between parental lines. In addition, it is unlikely that other genetic models will lead to higher  $r_{pc}$  than predicted with model  $E_{AA}$ , because the minimum  $\Delta_\alpha$  is bounded by the difference in allele frequencies between line 1 and the cross. Hence, we can expect that  $r_{pc}$  lies somewhere in between  $r_{pc}^D$  and  $r_{pc}^{AA}$  for other non-additive genetic models. To test this hypothesis, we used simulation to evaluate  $r_{pc}$  for two other non-additive genetic models, as described in the following.

### Methods

Simulation was used to validate the derived expressions for  $r_{pc}$  (i.e. Equations (13) and (16)). For that purpose, we simulated seven purebred lines that were either positively (P), negatively (N), or randomly selected (R) for the trait of interest. Both positive and negative selection were considered, such that pairs of lines were either selected in the same direction (convergent) or in opposite directions (divergent), resulting in pairs of lines with small and large differences in allele frequencies at QTL. We considered four scenarios that differed in the type of non-additive effects simulated (Fig. 1); (1) only dominance (D), (2) only additive-by-additive ( $A \times A$ ) epistasis ( $E_{AA}$ ), (3) both dominance and  $A \times A$  epistasis ( $D + E_{AA}$ ), and (4) complementary epistasis ( $E_C$ ). The latter was chosen because it is expected to result in substantial non-additive variance of all types (i.e. dominance,



additive-by-additive, dominance-by-additive, additive-by-dominance, and dominance-by-dominance) [27]. For each scenario and each pairwise cross between parental lines, we computed the realized (i.e. true)  $r_{pc}$  and compared it with the predicted  $r_{pc}$  based on Eqs. (13) and (16).

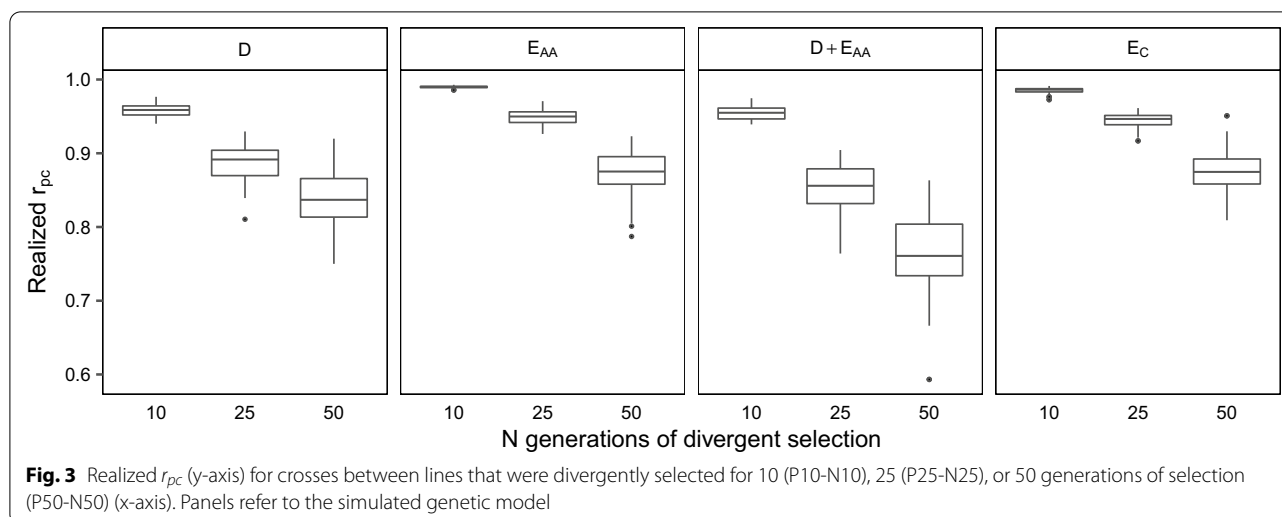
**Simulation**

**Population**

We simulated QTL genotypes of animals from seven breeding lines that originated from a common historical population using QMSim [28], such that the number of generations that separated pairs of lines ranged from 10 to 100 (Fig. 2). First, a historical population was simulated by randomly mating 600 females with 100 males for 200 generations. During the following 200 generations, the population size was gradually decreased to 300 females and 50 males, to generate linkage disequilibrium

(LD). Then, mating continued with a constant population size for another 200 generations. In the last historical generation (generation 0), the population size was increased to 1500 males and 1500 females by creating litters of 10 offspring per mating. The effective population size ( $N_e$ ) between generations - 600 and 0 was ~ 234, calculated as the harmonic mean of  $\frac{4N_mN_f}{N_m+N_f}$  in each historical generation, where  $N_m$  is the number of males and  $N_f$  is the number of females that become parents in a generation [19].

From the last historical generation, three breeding lines (P50, R and N50) were created by sampling 300 females and 50 males for each of the lines, without replacement. Within each line, mating continued for 50 generations, by selecting 50 males and 300 females in each generation using truncation selection based on own performance records with a broad sense heritability of 0.3. In line P50, selection was for high performance



(positive selection, P), in line N50 for low performance (negative selection, N), and in line R, selection was random. Similar to lines P50 and N50, two additional lines (P25 and N25) were created by randomly sampling and mating 50 males and 300 females from line R in generation 25, again without replacement. Within each of these lines, mating continued for 25 generations with positive (P25) or negative (N25) selection. Finally, another two lines (P10 and N10) were created by randomly sampling and mating 50 males and 300 females from line R in generation 40. Within these lines, mating continued for 10 generations with positive (P10) or negative (N10) selection. Litter size was kept constant at 10 offspring (5 male and 5 female) in each of the breeding lines, and mating of selected parents was always at random. The average  $N_e$  within the breeding lines was around  $\sim 115$ , which was calculated as  $1/(2\Delta F)$ , where  $\Delta F$  is the inbreeding rate estimated using pedigree [19]. The simulated lines resemble real livestock breeding lines that are under selection, with a rate of inbreeding of about 0.5%.

### Genome

The genome consisted of 10 chromosomes of 1 Morgan each. Each chromosome had 5000 randomly positioned bi-allelic loci. In the first historical generation, allele frequencies of these loci were sampled from a uniform distribution. During the historical generations, the mutation rate was  $5.0 \times 10^{-5}$ , while there was no mutation after the historical generations. In generation 0 (i.e., after the last historical generation), the distribution of allele frequencies was U-shaped, and we randomly selected 1000 loci from those that segregated to become QTL. We did not simulate markers, because our interest was in the true value of  $r_{pc}$ , not in its estimation.

### Functional genetic effects

The additive effect ( $a$ ) of each of the 1000 QTL was sampled from a normal distribution with mean 0 and variance 1. The size of non-additive effects at QTL was assumed to depend on the size of additive effects at these QTL. To achieve this, independently sampled dominance and epistatic coefficients for a QTL were scaled by the already sampled additive effects. Dominance coefficients ( $\delta$ ) were sampled from a normal distribution with a mean of 0.2 and a standard deviation of 0.3, following empirical observations by Bennewitz and Meuwissen [29] and Sun and Mumm [30]. Dominance effects ( $d$ ) were then computed by element-wise multiplication of  $\delta$  and  $|a|$ . Epistatic interactions limited to pairs of QTL and each QTL had an epistatic interaction with five randomly sampled QTL. Epistatic interactions between pairs of QTL followed either the additive-by-additive ( $E_{AA}$ ) configuration, or the complementary ( $E_C$ ) configuration (Fig. 1), depending on the scenario. Epistatic coefficients ( $\gamma$ ) were sampled from a normal distribution with a mean of 0, because epistatic effects are likely non-directional [31]. The standard deviation of epistatic effects was set such that the total functional epistatic variance per QTL was comparable to the total functional dominance variance per QTL in scenario D. The total functional dominance variance at a QTL is equal to the squared mean dominance coefficient, plus the variance of dominance coefficients. Because each QTL was involved in five epistatic interactions but had only one dominance effect, the standard deviation of epistatic coefficients was set to  $\sqrt{(0.2^2 + 0.3^2)/5} \approx 0.16$ . Epistatic effects ( $\epsilon$ ) were computed as  $\gamma_{kl}\sqrt{|a_k a_l|}$  for all pairwise interactions between QTL  $k$  and  $l$ .



### Average effects and additive genetic values

For a single locus, the average effect for PB performance in line 1 ( $\alpha_1$ ) was computed from the functional genetic effects ( $a$ ,  $d$ , and  $\epsilon$ ), and genotype frequencies in that line, as described in Duenk et al. [17], using the natural and orthogonal interactions (NOIA) model [32, 33]. The average effect for CB performance in line 1 when mated with line 2 ( $\alpha_{1(C)}$ ) was computed by the same procedure but with a small adjustment, as explained in Appendix 1.

Additive genetic values for PB performance in line 1 were computed as:

$$\mathbf{v}_1 = \mathbf{H}_1 \alpha_1 \tag{17}$$

where  $\mathbf{H}_1$  is the ( $n \times m$ ) QTL genotype matrix of animals in line 1 and  $\alpha_1$  is the ( $m \times 1$ ) column vector of average effects for PB performance in line 1, where  $n$  is the number of animals and  $m$  is the number of QTL. Genotypes in  $\mathbf{H}_1$  for individual  $i$  at QTL  $j$  were coded as in  $\mathbf{h}_i$  (Eq. (2)), with elements:

$$h_{ij} = \begin{cases} 2 - 2p_j \\ 1 - 2p_j \\ 0 - 2p_j \end{cases} \text{ for genotypes, } \begin{cases} FF \\ Ff \\ ff \end{cases}, \tag{18}$$

where  $p_j$  is the frequency of allele  $F$  at QTL  $j$  in line 1. Additive genetic values for CB performance of animals in line 1 (when mated to line 2) ( $\mathbf{v}_C$ ) were computed by replacing  $\alpha_1$  with  $\alpha_{1(C)}$  in Eq. (17).

### Parameters of interest

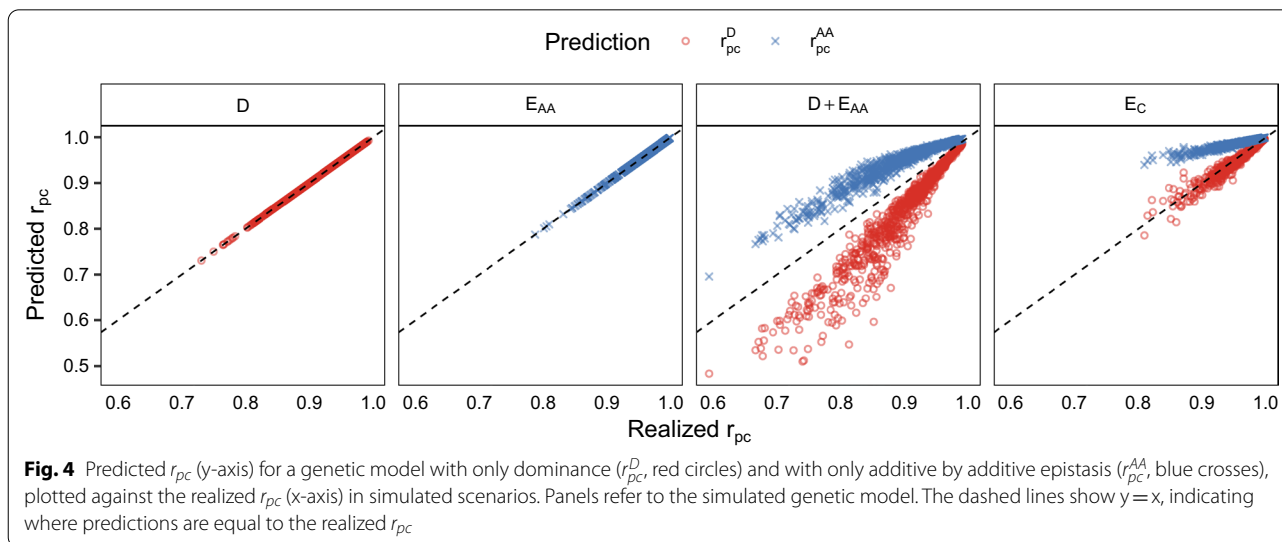
The true value of  $r_{pc}$  in line 1 when it is mated to line 2 was computed as the correlation between additive genetic values for PB ( $\mathbf{v}_1$ ) and CB performance ( $\mathbf{v}_C$ ) of animals in line 1, i.e. we did not estimate  $r_{pc}$  from the

simulated data. Note that this  $r_{pc}$  is not the same as the  $r_{pc}$  in line 2 when it is mated to line 1, because differences in allele frequencies between lines 1 and 2 lead to differences in contributions of QTL to the (co)variance of additive genetic values. In addition, average effects for CB performance in line 1 can differ from those in line 2 (e.g. based on Eqs. (5) and (6) for genetic model D). Thus, we computed  $r_{pc}$  for all  $7 \times (7 - 1) = 42$  combinations of breeding lines. All simulations were replicated 20 times, resulting in  $42 \times 20 = 840$  realized  $r_{pc}$  values for each scenario.

We compared each of the realized  $r_{pc}$  values with the theoretical predictions of  $r_{pc}$  under genetic models D ( $r_{pc}^D$ , Eq. (13)) and  $E_{AA}$  ( $r_{pc}^{AA}$ , Eq. (16)). We expected that  $r_{pc}^D$  and  $r_{pc}^{AA}$  would exactly predict  $r_{pc}$  in scenarios D and  $E_{AA}$ , respectively. For scenarios D +  $E_{AA}$  and  $E_C$ ,  $r_{pc}$  could not be expressed in terms of genetic parameters in the parental lines (see Discussion), but, as argued above, we expected that  $r_{pc}^D$  may represent a lower bound and  $r_{pc}^{AA}$  an upper bound of realized  $r_{pc}$ . Thus, it may be possible to predict the realized  $r_{pc}$  for these two scenarios with a multiple regression model with  $r_{pc}^D$  and  $r_{pc}^{AA}$  as covariates.

### Results

Figure 3 and Table S1 [see Additional file 1: Table S1] show the realized  $r_{pc}$  for all crosses between lines that were divergently selected for either 10, 25, or 50 generations (i.e., crosses P10-N10, P25-N25, and P50-N50). For all scenarios (i.e. the four simulated genetic models), the realized  $r_{pc}$  decreased with increasing generations of divergent selection, as expected. For each cross shown in Fig. 3, the lowest realized  $r_{pc}$  was observed when both dominance and epistasis were simulated (scenario D +  $E_{AA}$ ), and the highest realized  $r_{pc}$  was observed when



**Fig. 4** Predicted  $r_{pc}$  (y-axis) for a genetic model with only dominance ( $r_{pc}^D$ , red circles) and with only additive by additive epistasis ( $r_{pc}^{AA}$ , blue crosses), plotted against the realized  $r_{pc}$  (x-axis) in simulated scenarios. Panels refer to the simulated genetic model. The dashed lines show  $y = x$ , indicating where predictions are equal to the realized  $r_{pc}$

only epistasis was simulated (scenarios  $E_{AA}$  and  $E_C$ ). Differences in  $r_{pc}$  between scenarios were caused by differences in the genetic models between scenarios, rather than by differences in allele frequencies between lines, because mean differences in allele frequencies between lines were similar across scenarios (results not shown). This agrees with Duenk et al. [17].

Figure 4 shows the theoretical predictions of  $r_{pc}$  from our expressions, plotted against the realized  $r_{pc}$  from the simulations, for all replicates and for all combinations of parental lines within each replicate. For scenarios D and  $E_{AA}$ , our expressions for  $r_{pc}$  based on parameters in the purebred parental lines provided exact predictions of realized  $r_{pc}$  (left two panels in Fig. 4). For scenarios D +  $E_{AA}$  and  $E_C$ , our expressions for  $r_{pc}$  were expected to provide upper ( $r_{pc}^{AA}$ ) and lower ( $r_{pc}^D$ ) bounds for realized  $r_{pc}$ . However, for scenario  $E_C$ , realized  $r_{pc}$  was lower than the lower bound ( $r_{pc}^D$ ) in 12% of the cases, for which realized  $r_{pc}$  was about  $\sim 0.01$  lower than  $r_{pc}^D$ . For both scenarios D +  $E_{AA}$  and  $E_C$ , the gap between the predicted lower and upper bounds (i.e. the difference between  $r_{pc}^{AA}$  and  $r_{pc}^D$ ) increased with decreasing realized  $r_{pc}$ .

## Discussion

The aim of this study was to derive expressions for  $r_{pc}$  for a purebred line when it is mated to another purebred line, based on genetic variances within, and the genetic covariance between the two parental lines. These expressions were derived for a genetic model with additive and dominance effects (model D), and for a genetic model with additive and additive-by-additive ( $A \times A$ ) epistatic effects (model  $E_{AA}$ ). The results showed that our expressions provide exact predictions of  $r_{pc}$  for scenarios that were simulated based on models D and  $E_{AA}$ , respectively. For scenarios with both dominance and  $A \times A$  epistasis (D +  $E_{AA}$ ), and for models with complementary epistasis ( $E_C$ ), theoretical predictions could not be derived but the expressions for models D and  $E_{AA}$  provide approximate upper and lower bounds for  $r_{pc}$ , respectively. For the simulated D +  $E_{AA}$  scenario, the realized  $r_{pc}$  always fell between these bounds, while for the simulated  $E_C$  scenario, the realized  $r_{pc}$  was slightly lower than the lower bound in 12% of the cases.

The results of our simulations showed that the realized  $r_{pc}$  decreased for all scenarios (i.e. the four genetic models simulated) when the number of generations of divergent selection between the parental lines increased. This was as expected because, with divergent selection, the difference in allele frequencies between parental lines increases over time, causing an increase in the differences between the average effects for PB and CB performance within each parental line [15, 16]. The realized  $r_{pc}$  was lower for scenarios that involved dominance

compared to scenarios that involved only epistasis. Across scenarios, the realized  $r_{pc}$  ranged from 0.60 to 0.99, covering a large proportion of empirical estimates of  $r_{pc}$  for livestock [3–9].

## Predicting $r_{pc}$ in practice

The expressions derived in this study suggest that the  $r_{pc}$  can be predicted without CB information when certain variance components for the parental lines are known. Recent developments in genome-wide marker panels have made accurate estimation of the required variance components within and the covariances between distantly related lines feasible [18, 34, 35]. To validate our expressions as bounds for the  $r_{pc}$  for a parental line, we attempted to apply them to empirical estimates reported in the literature. However, to our knowledge, only one study presents both an estimate of  $r_{pc}$  based on PB and CB data, and of the corresponding genetic variances within, and of the covariances between the parental lines [36]. In that study, the estimate of  $r_{pc}$  in the Yorkshire breed when mated to the Landrace breed was 0.67 with a standard error of 0.10. Based on the PB estimates presented in that paper, the predicted lower bound of  $r_{pc}$  based on Eq. (13) was 0.30 and the predicted upper bound based on Eq. (16) was 0.84. These results suggest that it is unlikely that dominance is the only reason for the estimate of  $r_{pc}$  in this study to be lower than 1, because the estimated  $r_{pc}$  was higher than the lower bound. Thus, it is likely that epistasis or GxE is present.

Although our expressions appeared to predict lower and upper bounds for the above example, there are two important issues that arise when our expressions are applied to empirical data. First,  $r_{pc}$  may be lower than the predicted lower bound given by Eq. (13), because our expressions do not account for  $G \times E$  interactions.  $G \times E$  interactions may be present in the study of Xiang et al. [36] because the PB and CB animals used in that study were housed in different environments. However, this implies that the results of our study can be used to evaluate the relative contributions of  $G \times E$  and  $G \times G$  to  $r_{pc}$  by comparing estimates of  $r_{pc}$  from PB and CB data with the predicted lower bound. For example, when the estimate of  $r_{pc}$  is much lower than the predicted lower bound of  $r_{pc}$ , the contribution of  $G \times E$  is likely large compared to the contribution of non-additive effects and differences in allele frequencies.

Second, the estimates of variance components obtained from empirical data are usually different from those used in Eqs. (13) and (16). In Xiang et al. [36] for example, a bivariate model was used to estimate genetic parameters within and between the two parental lines (say line 1 and 2). With such a model, the estimate of the genetic variance in line 2 refers to the variance in line 2

for the trait expressed in line 2 (i.e.  $\sigma_2^2$ ). However, when the aim is to predict the bounds of  $r_{pc}$  in line 1, we need the genetic variance in line 1 for the trait expressed in line 2 (i.e.  $\sigma_{1(2)}^2$  in Eqs. (13) and (16)). Similarly, the covariance between lines 1 and 2 that is estimated from PB data from the two lines ( $\sigma_{1,2}$ ) is not the same as the covariance between lines used in our expressions ( $\sigma_{1,1(2)}$ ). Thus, as shown by our expressions,  $r_{pc}$  based on genetic model D is different from the genetic correlation between parental lines that is usually estimated in empirical studies (e.g., Xiang et al. [36]).

To predict  $r_{pc}$  in line 1 when it is mated to line 2, there is no obvious way to directly estimate the variance ( $\sigma_{1(2)}^2$ ) and covariance ( $\sigma_{1,1(2)}$ ) that are required for the expressions for  $r_{pc}$ , simply because the trait expressed in line 2 is not observed for individuals from line 1, but it may be possible to approximate them. One possible approach would be to estimate marker effects using genotype and phenotype data from line 2 and multiply them with the marker genotypes from line 1, resulting in genomic estimated breeding values (GEBV) for the animals in line 1 for the trait expressed in line 2. Parameter  $\sigma_{1(2)}^2$  can then be approximated by the variance of these GEBV and parameter  $\sigma_{1,1(2)}$  by the covariance between these GEBV and the usual GEBV for PB performance in line 1. Although this approach appears straightforward, there are at least two issues that must be addressed. First, the LD between markers and QTL may be different in the parental lines, which leads to differences in estimated marker effects even when the QTL effects are not different [37]. Second, the estimated marker effects are subject to shrinkage to a degree that depends on the number of phenotypic records and the minor allele frequency of the marker. As a result, the variance of the estimated marker effects is smaller than the variance of the “true” marker effects. The effect of these issues on the predicted  $r_{pc}$  requires further investigation.

### Three- and four-way crosses

The predictions of  $r_{pc}$  presented here are valid for purebred parental lines that produce a two-way crossbred (i.e. from mating with one other purebred parental line). However in practice, commercial animals are usually three- or four-way crossbreds. The  $r_{pc}$  for three- and four-way crosses under genetic models D and  $E_{AA}$  can also be expressed in terms of variances and covariances within and between parental lines. Derivation of these expressions is presented in Appendix 2. With genetic model D,  $r_{pc}$  for the two dam lines that are involved in a three-way cross is equal to the  $r_{pc}$  for two-way CB performance when the respective lines are mated directly to the sire line. This is because,

alleles at a locus that are transmitted to a three-way cross from each of the dam lines will always pair with an allele from the sire line. Hence, following Eq. (13),  $r_{pc}$  of a dam line for three-way CB performance is equal to the correlation between additive genetic values of the dams for the trait expressed in the dam line and in the sire line. For all other scenarios with three- or four-way crosses, the derivations of  $r_{pc}$  result in expressions that differ from the expressions for two-way crosses (Appendix 2).

### Validity of predicted bounds of $r_{pc}$

We hypothesized that the predicted  $r_{pc}$  under the dominance model ( $r_{pc}^D$ ) yields a lower bound for realized  $r_{pc}$  because the difference between average effects for PB and CB performance is maximized with model D, since the average effect for CB performance at a QTL is a function of allele frequencies in the mated line. Our simulations showed that this lower bound was indeed correct for most cases, apart from a few replicates for the scenario of complementary epistasis (model  $E_C$ ). This is probably because, with model  $E_C$ , the average effect for CB performance at a QTL involves a multiplication between the allele frequency of the same locus in the mated line, and the allele frequencies of the interacting loci in the cross. In contrast, for model D, the expression for the average effect for CB performance only involves the allele frequency of the same locus in the mated line (Eq. 6). As a consequence, with model  $E_C$ , a difference in the allele frequency between parental lines at a QTL can result in differences in average effects between PB and CB performance at two QTL, instead of at only one QTL with model D.

### Conclusions

We derived expressions for  $r_{pc}$  in purebred parental lines of two-, three-, and four-way crosses based on the genetic variances within and the genetic covariance between parental lines, noting that these variance components are not those that are typically estimated using empirical data. The expressions were derived for a genetic model with additive and dominance effects (model D), and for a model with additive and epistatic additive-by-additive effects (model  $E_{AA}$ ). Results showed that these expressions provide exact predictions of  $r_{pc}$  for models D and  $E_{AA}$ , and accurate upper and lower bounds of  $r_{pc}$  for genetic models with both dominance and additive-by-additive epistatic effects (model  $D + E_{AA}$ ), or with complementary epistatic effects (model  $E_C$ ). This work lays the foundation for estimation of  $r_{pc}$  based on information collected on the PB parental lines, without requiring CB information.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12711-021-00601-w>.

### 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. The use of the HPC cluster has been made possible by CAT-AgroFood (Shared Research Facilities Wageningen UR). The authors also thank Mehdi Sargolzaei for providing an improved version of QMSim.

### Authors' contributions

PD simulated and analysed the data, and wrote the manuscript. PB, YW, and MC supervised the study and assisted with the interpretation of results and writing of the manuscript. All authors read and approved the final manuscript.

### Funding

This research is supported by the Netherlands Organisation of Scientific Research (NWO) and the Breed4Food consortium partners Cobb Europe, CRV, Hendrix Genetics, and Topigs Norsvin.

### Availability of data and materials

The data simulated in this study can be reproduced with the files in the following GitHub repository: [https://git.wageningenur.nl/duenk002/predicting\\_rpc](https://git.wageningenur.nl/duenk002/predicting_rpc).

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## Appendices

### Appendix 1

With non-additive genetic effects (i.e. dominance and epistasis), average effects of QTL are a function of genotype frequencies and functional genetic effects. The procedure to obtain average effects for PB performance from known genotype frequencies and functional additive, dominance, and epistatic effects, was described in Duenk et al. [17]. In short, the procedure involves applying the natural and orthogonal interactions (NOIA) model [32] for each epistatic interaction between two loci in a population with given allele frequencies, resulting in statistically orthogonal terms that contribute to the average effects of the two loci. Consider for example two loci,  $k$  and  $l$ , that have an epistatic interaction between them. The functional epistatic values for each possible two-locus genotype can be partitioned into nine statistical genetic effects, using Eq. 2 in Duenk et al. [17]. For average effects for PB performance, the frequencies used in  $\mathbf{D}_{kl}$ ,  $\mathbf{W}_k$  and  $\mathbf{W}_l$  are those in the PB line. The procedure immediately leads to two terms ( $\alpha_{kl}^k$  and  $\alpha_{kl}^l$ ) that contribute to the average effects of loci  $k$  and  $l$ . For example, in a two-locus model where locus  $k$  has

an interaction with locus  $l$  only, the average effect for PB performance of locus  $k$  in line 1 is:

$$\alpha_1^k = a^k + (1 - 2p_1^k)d^k + \alpha_{kl}^k$$

where  $a^k$  is the functional additive effect of locus  $k$ ,  $p_1^k$  is the allele frequency of locus  $k$  in line 1, and  $d^k$  is the functional dominance effect of locus  $k$ . For PB performance, the average effect at locus  $k$  depends on genotype frequencies of  $k$  in line 1, because alleles of locus  $k$  transmitted to PB animals always pair with an allele from the same line origin. Furthermore, the average effect of  $k$  depends on genotype frequencies of locus  $l$  in line 1, because alleles of locus  $k$  transmitted to PB animals will be expressed in the genetic background of line 1.

The same procedure can be used to obtain average effects for CB performance for line 1, by making a small modification. The alleles of locus  $k$  transmitted to two-way crossbreds always pair with an allele from line 2 and they will be expressed in the genetic background of crossbreds. Hence, the average effect for CB performance of locus  $k$  depends on genotype frequencies of  $k$  in line 2, and on genotype frequencies of  $l$  in the crossbreds. Thus, to obtain the average effect at locus  $k$  for CB performance,  $\mathbf{W}_k$  needs to be constructed using the genotype frequencies of  $k$  in line 2,  $\mathbf{W}_l$  needs to be constructed using the genotype frequencies of  $l$  in crossbreds, and  $\mathbf{D}_{kl}$  is a  $(9 \times 9)$  diagonal matrix of two-locus genotype probabilities, constructed using genotype frequencies of  $k$  in line 2 and frequencies of  $l$  in crossbreds. Then, the average effect at locus  $k$  for CB performance in line 1 is:

$$\alpha_{1(C)}^k = a^k + (1 - 2p_2^k)d^k + \alpha_{kl}^k$$

where  $p_2^k$  is the allele frequency of locus  $k$  in line 2. For the same epistatic interaction, the procedure needs to be repeated for locus  $l$ , because the average effect at locus  $l$  for CB performance depends on genotype frequencies of  $l$  in line 2 and on genotype frequencies of  $k$  in the crossbreds.

### Appendix 2

#### Dam line of a three-way crossbred

Under genetic model D, the average effect at locus F for three-way CB performance for line 2 (or 3) is:

$$\alpha_{2(C)}^D = 0.5(a + (1 - 2p_1)d)$$

where  $p_1$  is the allele frequency of allele  $F$  in the sire line (line 1). The value 0.5 is because an allele from this dam line has a probability of 0.5 to be transmitted to the final

crossbred. Given the expressions for average effects for PB performance, the average effect for CB performance for line 2 is:

$$\alpha_{2(C)}^D = 0.5\alpha_1$$

The additive genetic variance for three-way CB performance for line 2 under model D is:

$$\begin{aligned} \sigma_{2(C)}^2 &= \text{var}(\mathbf{h}'_i \alpha_{2(C)}) = \text{var}\left(\sum_j h_{ij} 0.5\alpha_{1j}\right) \\ &= E\left[\left(\sum_j h_{ij} 0.5\alpha_{1j}\right)\left(\sum_j h_{ij} 0.5\alpha_{1j}\right)\right] \\ &= \sum_j E(h_{ij}h_{ij})(0.5\alpha_{1j})^2 \\ &= 0.25 \sum_j 2p_{2j}(1-p_{2j})\alpha_{1j}^2 = 0.25\sigma_{2(1)}^2 \end{aligned}$$

The additive genetic covariance between PB and three-way CB performance for line 2 under model D is:

$$\begin{aligned} \sigma_{2,2(C)} &= \text{cov}(\mathbf{h}'_i \alpha_2, \mathbf{h}'_i \alpha_{2(C)}) \\ &= \text{cov}\left(\sum_j h_{ij} \alpha_{2j}, \sum_j h_{ij} 0.5\alpha_{1j}\right) \\ &= E\left[\left(\sum_j h_{ij} \alpha_{2j}\right)\left(\sum_j h_{ij} 0.5\alpha_{1j}\right)\right] \\ &= \sum_j E(h_{ij}h_{ij})0.5\alpha_{2j}\alpha_{1j} \\ &= 0.5 \sum_j 2p_{2j}(1-p_{2j})\alpha_{2j}\alpha_{1j} = 0.5\sigma_{2,2(1)} \end{aligned}$$

Hence,  $r_{pc}$  for line 2 with genetic model D is equal to:

$$r_{pc}^D = \frac{0.5\sigma_{2,2(1)}}{\sigma_2 \sqrt{0.25\sigma_{2(1)}^2}} = \frac{\sigma_{2,2(1)}}{\sigma_2 \sigma_{2(1)}}$$

This result shows that the  $r_{pc}$  for a dam line of a three-way cross is equal to the genetic correlation between the traits for lines 1 and 2, as expressed in line 2. This result is similar to the expression of  $r_{pc}$  in a sire line of a two-way cross.

Under genetic model  $E_{AA}$ , the average effect at locus F for three-way CB performance for dam line 2 (or 3) is:

$$\alpha_{2(C)}^{AA} = a - (1 - 2p_C^G)\epsilon$$

where  $p_C^G$  is the allele frequency of allele G in cross 1(23). Given the expressions for average effects for PB performance in lines 1, 2, and 3, and using  $p_C^G = 0.5p_1^G + 0.25p_2^G + 0.25p_3^G$ , the average effect for three-way CB performance for line 2 under genetic model  $E_{AA}$  can be written as:

$$\alpha_{2(C)}^{AA} = 0.5\alpha_1 + 0.25\alpha_2 + 0.25\alpha_3$$

The additive genetic variance for three-way CB performance for line 2 with genetic model  $E_{AA}$  is:

$$\begin{aligned} \sigma_{2(C)}^2 &= \text{var}(\mathbf{h}'_i \alpha_{2(C)}) \\ &= \text{var}\left(\sum_j h_{ij}(0.5\alpha_{1j} + 0.25\alpha_{2j} + 0.25\alpha_{3j})\right) \\ &= E\left[\left(\sum_j h_{ij}(0.5\alpha_{1j} + 0.25\alpha_{2j} + 0.25\alpha_{3j})\right)\right. \\ &\quad \left.\left(\sum_j h_{ij}(0.5\alpha_{1j} + 0.25\alpha_{2j} + 0.25\alpha_{3j})\right)\right] \\ &= \sum_j E(h_{ij}h_{ij})(0.5\alpha_{1j} + 0.25\alpha_{2j} + 0.25\alpha_{3j})^2 \\ &= 0.25 \sum_j 2p_{2j}(1-p_{2j})\left(\alpha_{1j}^2 + 0.25\alpha_{2j}^2\right. \\ &\quad \left.+ 0.25\alpha_{3j}^2 + \alpha_{1j}\alpha_{2j} + \alpha_{1j}\alpha_{3j} + 2 * 0.25\alpha_{2j}\alpha_{3j}\right) \\ &= 0.25\left(\sigma_{2(1)}^2 + 0.25\sigma_2^2 + 0.25\sigma_{2(3)}^2\right. \\ &\quad \left.+ \sigma_{2,2(1)} + \sigma_{2(1),2(3)} + 0.5\sigma_{2,2(3)}\right) \end{aligned}$$

Note that  $\sigma_{2(1),2(3)}$  is the additive genetic covariance for individuals from line 2, between the trait expressed in lines 1 and 3.

The additive genetic covariance between PB performance and three-way CB performance for line 2 is:

$$\begin{aligned}
 \sigma_{2,2(C)} &= cov(\mathbf{h}'_i\alpha_2, \mathbf{h}'_i\alpha_{2(C)}) \\
 &= cov\left(\sum_j h_{ij}\alpha_{2j}, \sum_j h_{ij}\left(0.5\alpha_{1j}^{AA} + 0.25\alpha_{2j}^{AA} + 0.25\alpha_{3j}^{AA}\right)\right) \\
 &= E\left[\left(\sum_j h_{ij}\alpha_{2j}\right)\left(\sum_j h_{ij}\left(0.5\alpha_{1j}^{AA} + 0.25\alpha_{2j}^{AA} + 0.25\alpha_{3j}^{AA}\right)\right)\right] \\
 &= \sum_j E(h_{ij}h_{ij})0.5\alpha_{2j}\left(0.5\alpha_{1j}^{AA} + 0.25\alpha_{2j}^{AA} + 0.25\alpha_{3j}^{AA}\right) \\
 &= 0.5\sum_j 2p_{2j}(1-p_{2j})\alpha_{2j}\left(\alpha_{1j}^{AA} + 0.5\alpha_{2j}^{AA} + 0.5\alpha_{3j}^{AA}\right) \\
 &= 0.5\left(\sigma_{2,2(1)} + 0.5\sigma_2^2 + 0.5\sigma_{2,2(3)}\right)
 \end{aligned}$$

Hence, the  $r_{pc}$  for line 2 under genetic model  $E_{AA}$  is:

$$\begin{aligned}
 r_{pc}^{AA} &= \frac{0.5\left(\sigma_{2,2(1)} + 0.5\sigma_2^2 + 0.5\sigma_{2,2(3)}\right)}{\sigma_2\sqrt{0.25\left(\sigma_{2(1)}^2 + 0.25\sigma_2^2 + 0.25\sigma_{2(3)}^2\right) + \sigma_{2,2(1)} + \sigma_{2(1),2(3)} + 0.5\sigma_{2,2(3)}}} \\
 &= \frac{\sigma_{2,2(1)} + 0.5\sigma_2^2 + 0.5\sigma_{2,2(3)}}{\sigma_2\sqrt{\left(\sigma_{2(1)}^2 + 0.25\sigma_2^2 + 0.25\sigma_{2(3)}^2\right) + \sigma_{2,2(1)} + \sigma_{2(1),2(3)} + 0.5\sigma_{2,2(3)}}}
 \end{aligned}$$

**Sire line of a 3-way crossbred**

Under genetic model D, the average effect at locus F for three-way CB performance for line 1 is:

$$\alpha_{1(C)}^D = a + (1 - 2p_{23})d$$

where  $p_{23}$  is the allele frequency of allele  $F$  in cross 23. Given the expressions for average effects for PB performance, and using  $p_{23} = (p_2 + p_3)/2$ , the average effect for CB performance for line 1 is:

$$\alpha_{1(C)}^D = 0.5(\alpha_2 + \alpha_3)$$

The additive genetic variance for three-way CB performance for line 1 under model D is:

$$\begin{aligned}
 \sigma_{1(C)}^2 &= var(\mathbf{h}'_i\alpha_{1(C)}) = var\left(\sum_j h_{ij}0.5(\alpha_{2j} + \alpha_{3j})\right) \\
 &= E\left[\left(\sum_j h_{ij}0.5(\alpha_{2j} + \alpha_{3j})\right)\left(\sum_j h_{ij}0.5(\alpha_{2j} + \alpha_{3j})\right)\right] \\
 &= \sum_j E(h_{ij}h_{ij})\left(0.5(\alpha_{2j} + \alpha_{3j})\right)^2 \\
 &= 0.25\sum_j 2p_{2j}(1-p_{2j})(\alpha_{2j}\alpha_{2j} + \alpha_{3j}\alpha_{3j} + 2\alpha_{2j}\alpha_{3j}) \\
 &= 0.25(\sigma_{1(2)}^2 + \sigma_{1(3)}^2 + 2\sigma_{1(2),1(3)})
 \end{aligned}$$

The additive genetic covariance between PB and three-way CB performance for line 1 under model D is:

$$\begin{aligned} \sigma_{1,1(C)} &= cov(\mathbf{h}'_i\alpha_1, \mathbf{h}'_i\alpha_{1(C)}) \\ &= cov\left(\sum_j h_{ij}\alpha_{1j}, \sum_j h_{ij}0.5(\alpha_{2j} + \alpha_{3j})\right) \\ &= E\left[\left(\sum_j h_{ij}\alpha_{1j}\right)\left(\sum_j h_{ij}0.5(\alpha_{2j} + \alpha_{3j})\right)\right] \\ &= \sum_j E(h_{ij}h_{ij})(0.5\alpha_{1j}\alpha_{2j} + 0.5\alpha_{1j}\alpha_{3j}) \\ &= 0.5 \sum_j 2p_{1j}(1 - p_{1j})(\alpha_{1j}\alpha_{2j} + \alpha_{1j}\alpha_{3j}) \\ &= 0.5(\sigma_{1,1(2)} + \sigma_{1,1(3)}) \end{aligned}$$

Hence, the  $r_{pc}$  for a sire line of a three-way cross with genetic model D is equal to:

$$\begin{aligned} r_{pc}^D &= \frac{0.5(\sigma_{1,1(2)} + \sigma_{1,1(3)})}{\sigma_1 \sqrt{0.25(\sigma_{1(2)}^2 + \sigma_{1(3)}^2 + 2\sigma_{1(2),1(3)})}} \\ &= \frac{\sigma_{1,1(2)} + \sigma_{1,1(3)}}{\sigma_1 \sqrt{(\sigma_{1(2)}^2 + \sigma_{1(3)}^2 + 2\sigma_{1(2),1(3)})}} \end{aligned}$$

Under genetic model  $E_{AA}$ , the average effect at locus F for three-way CB performance for line 1 is:

$$\alpha_{1(C)}^{AA} = a - (1 - 2p_C^G)\epsilon$$

where  $p_C^G$  is the allele frequency of allele  $G$  in cross 1(23). Given the expressions for average effects for PB performance for lines 1, 2, and 3, and using

$p_C^G = 0.5p_1^G + 0.25p_2^G + 0.25p_3^G$ , the average effect for three-way CB performance for line 1 under genetic model  $E_{AA}$  can be written as:

$$\alpha_{1(C)}^{AA} = 0.5\alpha_1 + 0.25\alpha_2 + 0.25\alpha_3$$

The additive genetic variance for three-way CB performance for line 1 under model  $E_{AA}$  is:

$$\begin{aligned} \sigma_{1(C)}^2 &= var(\mathbf{h}'_i\alpha_{1(C)}) \\ &= var\left(\sum_j h_{ij}(0.5\alpha_{1j} + 0.25\alpha_{2j} + 0.25\alpha_{3j})\right) \\ &= E\left[\left(\sum_j h_{ij}(0.5\alpha_{1j} + 0.25\alpha_{2j} + 0.25\alpha_{3j})\right)^2\right] \\ &= \sum_j E(h_{ij}h_{ij})(0.5\alpha_{1j} + 0.25\alpha_{2j} + 0.25\alpha_{3j})^2 \\ &= \sum_j 2p_{1j}(1 - p_{1j})(0.25\alpha_{1j}^2 + 0.25^2\alpha_{2j}^2 + 0.25^2\alpha_{3j}^2 \\ &\quad + 0.25\alpha_{1j}\alpha_{2j} + 0.25\alpha_{1j}\alpha_{3j} + 0.125\alpha_{2j}\alpha_{3j}) \\ &= 0.25 \sum_j 2p_{1j}(1 - p_{1j})(\alpha_{1j}^2 + 0.25\alpha_{2j}^2 \\ &\quad + 0.25\alpha_{3j}^2 + \alpha_{1j}\alpha_{2j} + \alpha_{1j}\alpha_{3j} + 0.5\alpha_{2j}\alpha_{3j}) \\ &= 0.25(\sigma_1^2 + 0.25\sigma_{1(2)}^2 + 0.25\sigma_{1(3)}^2 \\ &\quad + \sigma_{1,1(2)} + \sigma_{1,1(3)} + 0.5\sigma_{1(2),1(3)}) \end{aligned}$$

The additive genetic covariance between PB and three-way CB performance for line 1 under genetic model  $E_{AA}$  is:

---


$$\begin{aligned} \sigma_{1,1(C)} &= cov(\mathbf{h}'_i\alpha_1, \mathbf{h}'_i\alpha_{1(C)}) \\ &= cov\left(\sum_j h_{ij}\alpha_{1j}, \sum_j h_{ij}(0.5\alpha_{1j} + 0.25\alpha_{2j} + 0.25\alpha_{3j})\right) \\ &= E\left[\left(\sum_j h_{ij}\alpha_{1j}\right)\left(\sum_j h_{ij}(0.5\alpha_{1j} + 0.25\alpha_{2j} + 0.25\alpha_{3j})\right)\right] \\ &= \sum_j E(h_{ij}h_{ij})(0.5\alpha_{1j}^2 + 0.25\alpha_{1j}\alpha_{2j} + 0.25\alpha_{1j}\alpha_{3j}) \\ &= 0.5 \sum_j 2p_{1j}(1 - p_{1j})(\alpha_{1j}^2 + 0.5\alpha_{1j}\alpha_{2j} + 0.5\alpha_{1j}\alpha_{3j}) \\ &= 0.5(\sigma_1^2 + 0.5\sigma_{1,1(2)} + 0.5\sigma_{1,1(3)}) \end{aligned}$$


---

Hence, the  $r_{pc}$  for line 1 for three-way CB performance under genetic model  $E_{AA}$  is equal to:

$$r_{pc}^{AA} = \frac{0.5(\sigma_1^2 + 0.5\sigma_{1,1(2)} + 0.5\sigma_{1,1(3)})}{\sigma_1 \sqrt{0.25(\sigma_1^2 + 0.25\sigma_{1(2)}^2 + 0.25\sigma_{1(3)}^2 + \sigma_{1,1(2)} + \sigma_{1,1(3)} + 0.5\sigma_{1(2),1(3)})}}$$

$$= \frac{\sigma_1^2 + 0.5\sigma_{1,1(2)} + 0.5\sigma_{1,1(3)}}{\sigma_1 \sqrt{(\sigma_1^2 + 0.25\sigma_{1(2)}^2 + 0.25\sigma_{1(3)}^2 + \sigma_{1,1(2)} + \sigma_{1,1(3)} + 0.5\sigma_{1(2),1(3)})}}$$

The additive genetic covariance between PB and four-way CB performance for line 1 under model D is:

**Four-way cross**

Under genetic model D, the average effect at locus F for four-way CB performance for line 1 is:

$$\alpha_{1(C)}^D = 0.5(a + (1 - 2p_{34})d)$$

where  $p_{34}$  is the allele frequency of allele  $F$  in cross 34. The 0.5 is because an allele from this parental line has a probability of 0.5 to be transmitted to the final crossbred. Given the expressions for average effects for PB performance, and using  $p_{34} = (p_3 + p_4)/2$ , the average effect for CB performance for line 1 is:

$$\alpha_{1(C)}^D = 0.25(\alpha_3 + \alpha_4)$$

The additive genetic variance for four-way CB performance for line 1 under model D is:

$$\sigma_{1(C)}^2 = var(\mathbf{h}'_i \alpha_{1(C)})$$

$$= var\left(\sum_j h_{ij} 0.25(\alpha_{3j} + \alpha_{4j})\right)$$

$$= E\left[\left(\sum_j h_{ij} 0.25(\alpha_{3j} + \alpha_{4j})\right)\left(\sum_j h_{ij} 0.25(\alpha_{3j} + \alpha_{4j})\right)\right]$$

$$= \sum_j E(h_{ij} h_{ij}) (0.25(\alpha_{3j} + \alpha_{4j}))^2$$

$$= \sum_j 2p_{1j}(1 - p_{1j}) 0.25^2 (\alpha_{3j}^2 + \alpha_{4j}^2 + 2\alpha_{3j}\alpha_{4j})$$

$$= 0.25^2 \sum_j 2p_{1j}(1 - p_{1j}) (\alpha_{3j}^2 + \alpha_{4j}^2 + 2\alpha_{3j}\alpha_{4j})$$

$$= 0.25^2 (\sigma_{1(3)}^2 + \sigma_{1(4)}^2 + 2\sigma_{1(3),1(4)})$$

$$\sigma_{1,1(C)} = cov(\mathbf{h}'_i \alpha_1, \mathbf{h}'_i \alpha_{1(C)})$$

$$= cov\left(\sum_j h_{ij} \alpha_{1j}, \sum_j h_{ij} 0.25(\alpha_{3j} + \alpha_{4j})\right)$$

$$= E\left[\left(\sum_j h_{ij} \alpha_{1j}\right)\left(\sum_j h_{ij} 0.25(\alpha_{3j} + \alpha_{4j})\right)\right]$$

$$= \sum_j E(h_{ij} h_{ij}) \alpha_{1j} (0.25(\alpha_{3j} + \alpha_{4j}))$$

$$= 0.25 \sum_j 2p_{1j}(1 - p_{1j}) (\alpha_{1j}\alpha_{3j} + \alpha_{1j}\alpha_{4j})$$

$$= 0.25(\sigma_{1,1(3)} + \sigma_{1,1(4)})$$

Hence, the  $r_{pc}$  for line 1 under genetic model D is equal

to:

$$r_{pc}^D = \frac{0.25(\sigma_{1,1(3)} + \sigma_{1,1(4)})}{\sigma_1 \sqrt{0.25^2(\sigma_{1(3)}^2 + \sigma_{1(4)}^2 + 2\sigma_{1(3),1(4)})}} = \frac{\sigma_{1,1(3)} + \sigma_{1,1(4)}}{\sigma_1 \sqrt{(\sigma_{1(3)}^2 + \sigma_{1(4)}^2 + 2\sigma_{1(3),1(4)})}}$$



This shows that, under genetic model D, the  $r_{pc}$  for four-way CB performance is similar to the  $r_{pc}$  for three-way CB performance for a sire line.

Under genetic model  $E_{AA}$ , the average effect at locus F for four-way CB performance for line 1 is:

$$\alpha_{1(C)}^{AA} = a - (1 - 2p_C^G)\epsilon$$

where  $p_C^G$  is the allele frequency of allele  $G$  for cross (12)(34). Given the expressions for average effects for PB performance in lines 1, 2, 3, and 4, and using  $p_C^G = 0.25p_1^G + 0.25p_2^G + 0.25p_3^G + 0.25p_4^G$ , the average effect for four-way CB performance for line 1 under genetic model  $E_{AA}$  can be written as the average of average effects for PB performance:

$$\alpha_{1(C)}^{AA} = 0.25\alpha_1 + 0.25\alpha_2 + 0.25\alpha_3 + 0.25\alpha_4$$

The additive genetic variance for four-way CB performance for line 1 can be written as:

$$\begin{aligned} \sigma_{1(C)}^2 &= \text{var}(\mathbf{h}_i\alpha_{1(C)}) \\ &= \text{var}\left(\sum_j h_{ij}(0.25\alpha_1 + 0.25\alpha_2 + 0.25\alpha_3 + 0.25\alpha_4)\right) \\ &= E\left[\left(\sum_j h_{ij}(0.25\alpha_1 + 0.25\alpha_2 + 0.25\alpha_3 + 0.25\alpha_4)\right)^2\right] \\ &= \sum_j E(h_{ij}h_{ij})(0.25\alpha_1 + 0.25\alpha_2 + 0.25\alpha_3 + 0.25\alpha_4)^2 \\ &= 0.25^2 \sum_j 2p_{1j}(1 - p_{1j}) \left(\sum_{l=1}^4 \alpha_{lj}^2 + \sum_{l=1}^4 \sum_{k \neq l}^4 \alpha_{lj}\alpha_{kj}\right) \\ &= 0.25^2 \left(\sum_{l=1}^4 \sigma_{1(l)}^2 + \sum_{l=1}^4 \sum_{k=1, k \neq l}^4 \sigma_{1(l),1(k)}\right) \end{aligned}$$

while defining that  $\sigma_{1(1)}^2 = \sigma_1^2$ , and that  $\sigma_{1(1),1(2)} = \sigma_{1,1(2)}$ . The above notation is the sum of all variances and covariances for individuals from line 1 for the traits expressed in the four parental lines.

The additive genetic covariance between PB and four-way CB performance for line 1 is:

$$\begin{aligned} \sigma_{1,1(C)} &= \text{cov}(\mathbf{h}_i\alpha_1, \mathbf{h}_i\alpha_{1(C)}) \\ &= \text{cov}\left(\sum_j h_{ij}\alpha_{1j}, \sum_j h_{ij}(0.25\alpha_1 + 0.25\alpha_2 + 0.25\alpha_3 + 0.25\alpha_4)\right) \\ &= E\left[\left(\sum_j h_{ij}\alpha_{1j}\right)\left(\sum_j h_{ij}(0.25\alpha_1 + 0.25\alpha_2 + 0.25\alpha_3 + 0.25\alpha_4)\right)\right] \\ &= 0.25 \sum_j E(h_{ij}h_{ij})\alpha_{1j}(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) \\ &= 0.25 \sum_j 2p_{1j}(1 - p_{1j}) \sum_{l=1}^4 \alpha_{1j}\alpha_{lj} = 0.25 \sum_{l=1}^4 \sigma_{1,1(l)} \end{aligned}$$

while defining that  $\sigma_{1,1(1)} = \sigma_1^2$ .

Hence, the  $r_{pc}$  for line 1 with genetic model  $E_{AA}$  is equal to:

$$\begin{aligned} r_{pc}^{AA} &= \frac{0.25 \sum_{l=1}^4 \sigma_{1,1(l)}}{\sigma_1 \sqrt{0.25^2 \left(\sum_{l=1}^4 \sigma_{1(l)}^2 + \sum_{l=1}^4 \sum_{k=1, k \neq l}^4 \sigma_{1(l),1(k)}\right)}} \\ &= \frac{\sum_{l=1}^4 \sigma_{1,1(l)}}{\sigma_1 \sqrt{\sum_{l=1}^4 \sigma_{1(l)}^2 + \sum_{l=1}^4 \sum_{k=1, k \neq l}^4 \sigma_{1(l),1(k)}}} \end{aligned}$$

**Additional file 1: Table S1.** Minimum, mean, and maximum values of realized  $r_{pc}$  for crosses between lines that were divergently selected for 10 (P10-N10), 25 (P25-N25), or 50 generations of selection (P50-N50).

Received: 17 June 2020 Accepted: 8 January 2021  
Published online: 04 February 2021

**References**

1. Smith C. The use of specialised sire and dam lines in selection for meat production. *Anim Sci.* 1964;6:337–44.
2. Dickerson GE. Inbreeding and heterosis in animals. *J Anim Sci.* 1973;1973:54–77.
3. Wientjes YCJ, Calus MPL. BOARD INVITED REVIEW: The purebred-crossbred correlation in pigs: A review of theory, estimates, and implications. *J Anim Sci.* 2017;95:3467–78.
4. Wei M, van der Werf JH. Genetic correlation and heritabilities for purebred and crossbred performance in poultry egg production traits. *J Anim Sci.* 1995;73:2220–6.
5. Mulder HA, Visscher J, Fablet J. Estimating the purebred–crossbred genetic correlation for uniformity of eggshell color in laying hens. *Genet Sel Evol.* 2016;48:39.

6. Lukaszewicz M, Davis R, Bertrand JK, Misztal I, Tsuruta S. Correlations between purebred and crossbred body weight traits in Limousin and Limousin-Angus populations. *J Anim Sci*. 2015;93:1490–3.
7. Newman S, Reverter A, Johnston DJ. Purebred-crossbred performance and genetic evaluation of postweaning growth and carcass traits in *Bos indicus* x *Bos taurus* crosses in Australia. *J Anim Sci*. 2002;80:1801–8.
8. Besbes B, Gibson JP. Genetic variation of egg production traits in purebred and crossbred laying hens. *Anim Sci*. 1999;68:433–9.
9. Duenk P, Calus MPL, Wientjes YCJ, Breen VP, Henshall JM, Hawken R, et al. Estimating the purebred-crossbred genetic correlation of body weight in broiler chickens with pedigree or genomic relationships. *Genet Sel Evol*. 2019;51:6.
10. Lutaaya E, Misztal I, Mabry JW, Short T, Timm HH, Holzbauer R. Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. *J Anim Sci*. 2001;79:3002–7.
11. Wientjes YCJ, Bijma P, Vandenplas J, Calus MPL. Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. *Genetics*. 2017;207:503–15.
12. Zumbach B, Misztal I, Tsuruta S, Holl J, Herring W, Long T. Genetic correlations between two strains of Durocs and crossbreds from differing production environments for slaughter traits. *J Anim Sci*. 2007;85:901–8.
13. Lo LL, Fernando RL, Grossman M. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. *J Anim Sci*. 1997;75:2877–84.
14. Falconer DS. The problem of environment and selection. *Am Nat*. 1952;86:293–8.
15. Wei M, van der Steen HAM, van der Werf JHJ, Brascamp EW. Relationship between purebred and crossbred parameters. *J Anim Breed Genet*. 1991;108:253–61.
16. Baumung R, Solkner J, Essl A. Correlation between purebred and crossbred performance under a two-locus model with additive by additive interaction. *J Anim Breed Genet*. 1997;114:89–98.
17. Duenk P, Bijma P, Calus MPL, Wientjes YCJ, van der Werf JHJ. The impact of non-additive effects on the genetic correlation between populations. *G3 (Bethesda)*. 2020;10:783–95.
18. Wientjes YCJ, Calus MPL, Duenk P, Bijma P. Required properties for markers used to calculate unbiased estimates of the genetic correlation between populations. *Genet Sel Evol*. 2018;50:65.
19. Falconer D, Mackay T. Introduction to quantitative genetics. 4th ed. Harlow: Pearson Education Limited; 1996.
20. Falconer DS. A note on Fisher's 'average effect' and 'average excess'. *Genet Res (Camb)*. 1985;46:337–47.
21. Pirchner F, Mergl R. Overdominance as cause for heterosis in poultry. *Z Tierzücht Züchtungsbio*. 1977;94:151–8.
22. Dekkers JCM. Breeding values for identified quantitative trait loci under selection. *Genet Sel Evol*. 1999;31:421–36.
23. Zeng J, Toosi A, Fernando RL, Dekkers JC, Garrick DJ. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genet Sel Evol*. 2013;45:11.
24. Vitezica ZG, Varona L, Elsen JM, Misztal I, Herring W, Legarra A. Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. *Genet Sel Evol*. 2016;48:6.
25. Wade MJ. A gene's eye view of epistasis, selection and speciation. *J Evol Biol*. 2002;15:337–46.
26. Xiang T, Christensen OF, Vitezica ZG, Legarra A. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. *Genet Sel Evol*. 2016;48:92.
27. Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*. 2008;4:e1000008.
28. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. *Bioinformatics*. 2009;25:680–1.
29. Bennewitz J, Meuwissen THE. The distribution of QTL additive and dominance effects in porcine F2 crosses. *J Anim Breed Genet*. 2010;127:171–9.
30. Sun X, Mumm RH. Method to represent the distribution of QTL additive and dominance effects associated with quantitative traits in computer simulation. *BMC Bioinformatics*. 2016;17:73.
31. Pavlicev M, Le Rouzic A, Cheverud JM, Wagner GP, Hansen TF. Directionality of epistasis in a murine intercross population. *Genetics*. 2010;185:1489–505.
32. Álvarez-Castro JM, Carlborg Ö. A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics*. 2007;176:1151–67.
33. Vitezica ZG, Legarra A, Toro MA, Varona L. Orthogonal estimates of variances for additive, dominance and epistatic effects in populations. *Genetics*. 2017;206:1297–307.
34. Karoui S, Carabaño MJ, Díaz C, Legarra A. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. *Genet Sel Evol*. 2012;44:39.
35. Carillier C, Larroque H, Robert-Granié C. Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. *Genet Sel Evol*. 2014;46:67.
36. Xiang T, Christensen OF, Legarra A. Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. *J Anim Sci*. 2017;95:1472–80.
37. Wientjes YCJ, Veerkamp RF, Bijma P, Bovenhuis H, Schrooten C, Calus MPL. Empirical and deterministic accuracies of across-population genomic prediction. *Genet Sel Evol*. 2015;47:5.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

