THE CORRELATION OF SUBSTITUTION EFFECTS ACROSS POPULATIONS AND GENERATIONS IN THE PRESENCE OF NON-ADDITIVE FUNCTIONAL GENE ACTION

A. Legarra*, C.A. Garcia-Baccino*[,†], Y.C.J. Wientjes[‡], , Z.G. Vitezica*

* INRAE/INP, UMR 1388 GenPhySE, 31326, Castanet-Tolosan, France.

[†] Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos Aires, C1417DSQ, Buenos Aires, Argentina.

[‡] Wageningen University & Research, Animal Breeding and Genomics, 6700 AH Wageningen, The Netherlands.

RUNNING TITLE: Correlation of QTL substitution effects

Keywords: QTL, substitution effects, epistasis, dominance, genetic distance

Corresponding author:

Andres Legarra,

andres.legarra@inrae.fr

+33 561285182

INRAE, UMR1388 GenPhySE

CS 52627

31326 Castanet Tolosan, France

ABSTRACT

Allele substitution effects at quantitative trait loci (QTL) are part of the basis of quantitative genetics theory and applications such as association analysis and genomic prediction. In presence of non-additive functional gene action, substitution effects are not constant across populations. We develop an original approach to model the difference in substitution effects across populations as first order Taylor series expansion from a “focal” population. This expansion involves the difference in allele frequencies and second-order statistical effects (additive by additive and dominance). The change in allele frequencies is a function of relationships (or genetic distances) across populations. As a result, it is possible to estimate the correlation of substitution effects across two populations using three elements: magnitudes of additive, dominance and additive by additive variance; relationships across populations (similar to Fst indexes); and functions of heterozygosities at the markers. Similarly, the theory applies as well to distinct generations in a population, in which case the distance across generations is a function of increase of inbreeding. Using published estimates of the needed parameters, we estimate the correlation between substitution effects to be around 0.60 for distinct breeds and higher than 0.9 for generations closer than 5. Simulation results confirmed our derivations although our estimators tended to underestimate the correlation. Our derivations are useful to understand the difficulty in predicting across populations, to forecast the possibility of prediction across populations, and we suggest that they can be useful to disentangle genotype by environment interaction from genotype by genotype interaction.

50 INTRODUCTION

51

52 One of the aims of quantitative genetics is to provide methods for prediction, for instance
53 genomic prediction (prediction of livestock breeding values or of crop performance) or
54 polygenic risk score (risk of a disease in humans). These predictions would ideally work
55 across a range of populations (different breeds, future generations). Ideally, the prediction
56 goes through a process of identifying causal genes, estimating their effects in some
57 population, and transposing these effects to newly genotyped individuals (Lande and
58 Thompson 1990; Meuwissen *et al.* 2001). These "gene effects" are substitution effects –
59 loosely speaking, the regression of the own phenotype (for polygenic risk scores) or
60 expected progeny phenotypes (for estimated breeding values) on gene content at the locus.
61 Being able to use substitution effects at causal genes across populations and generations is
62 the ultimate goal of genomic prediction, QTL detection and to some extent of molecular
63 biology description of gene effects.

64

65 There are several obstacles for these aims. Finding and validating causal genes and
66 understanding their functional mechanism is extremely difficult (Grobet *et al.* 1997; Bonifati
67 *et al.* 2003; Rupp *et al.* 2015). In practice, predictions are done using SNP markers using
68 statistical genetics techniques. Use of markers results in very good predictions within
69 population but mediocre (at best) predictions across populations, even with very
70 sophisticated techniques (Hayes *et al.* 2009; Karoui *et al.* 2012; Porto-Neto *et al.* 2015;
71 MacLeod *et al.* 2016). Indeed, livestock and human genetics empirical results show
72 decreasing predictive ability with increasing genetic distance across distinct populations or
73 generations (Liu *et al.* 2016; Martin *et al.* 2019). The lack of perfect linkage disequilibrium

across markers and genes has been claimed to be a reason for this decrease in accuracy. Adding extra information (more dense maps, biological prior information) should result in a better choice of markers close to causal genes, and therefore in boosting predictive abilities across populations (Roos *et al.* 2009; MacLeod *et al.* 2016). However, in practice, the increase in predictive ability across populations is small at best (MacLeod *et al.* 2016; Moghaddar *et al.* 2019). This is against results in simulations (Roos *et al.* 2009; MacLeod *et al.* 2016) – but a problem in simulations is that typically gene effects are assumed to be biologically additive and therefore constant across populations.

In fact, substitution (statistical) additive gene action is not homogeneous across populations, even for exactly the same mutation. Examples in livestock genetics include myostatin gene (Aiello *et al.* 2018) or DGAT1 (Gautier *et al.* 2007). Although part of these differences may be due to genotype by environment interactions, it is also plausible that this is due to epistasis or dominance; for instance this is the case in DGAT1 (Streit *et al.* 2011). For instance, in the 5-loci epistatic network in Carlborg et al. (2006) some substitution effects at genes switch signs depending on genetic background. In Drosophila, estimated (GWAS) substitution effects switched signs depended on the genetic background (Magwire *et al.* 2010; Mackay 2015).

There is indeed widespread evidence of biological epistasis (Mackay 2014). Whereas biological epistasis does not impede (on the contrary) large additive variation (Hill *et al.* 2008; Mackay 2014; Mäki-Tanila and Hill 2014), it implies is that substitution effects do vary across genetic backgrounds. Thus, in the presence of functional dominance and epistasis, there is no stability of substitution effects across different genetic backgrounds. Even if in all

of these populations, additive variation accounts for most genetic variation, and additive substitution effects are sizable (Hill *et al.* 2008), substitution effects may differ across populations.

Recent simulations (Dai *et al.* 2020; Duenk *et al.* 2020) showed that the difference in substitution effects across populations may be quite large under non-additive biological gene action and increases with divergence of populations. It is relatively easy to derive algebraic expressions for substitution effects α assuming specific hypothesis of biological gene action. For instance, assuming biological additive and dominance effects only (Falconer and Mackay 1996) results in $\alpha = a + (q - p)d$, whereas assuming additive, dominance and additive by additive biological gene effects results in $\alpha_1 = a_1 + (q_1 + p_1)d_1 + (p_2 - q_2)[aa]_{12}$ (Fuerst *et al.* 1997). Both of the approaches (simulation or analytical) are limited because the hypotheses of specific biological gene actions (*e.g.* 2-loci interaction but not 3-loci interaction) are too restrictive. Classically, these hypotheses are bypassed in quantitative genetics by working on statistical effects (Falconer and Mackay 1996; Lynch and Walsh 1998; Mäki-Tanila and Hill 2014).

The key parameter to describe the resemblance of statistical effects across populations is the correlation of substitution effects across populations. Under certain assumptions (independence of allele frequencies and substitution effects, appropriate coding; (Wientjes *et al.* 2017)), this correlation can be estimated from SNP markers and data of two populations (Karoui *et al.* 2012; Wientjes *et al.* 2017), and can easily be accommodated into genomic prediction models (Karoui *et al.* 2012; Xiang *et al.* 2017). In addition to empirical results, some theory to describe the resemblance of substitution effects across different

122 populations would be helpful to (1) better understand and quantify the change in *true*
123 substitution effects (*i.e.* if true genes instead of markers were being used), (2) give upper
124 bounds for genomic prediction across populations/generations, and (3) allow *a priori*
125 planning of genomic predictions *i.e.* including or not different subpopulations.
126
127 The aim of this work is to develop a theory to predict the extent of change of substitution
128 effects across space (breeds, lines) and across time (generations), without invoking or
129 assuming specific modes of biological gene action. We obtain explicit estimators that are
130 functions of additive, dominance and additive by additive variances, genetic distances across
131 populations, and distributions of allele frequencies. We check and illustrate our theory using
132 published results and simulations considering dominance and epistasis (additive by additive
133 and complementary) from 5-loci interactions.
134
135 MATERIALS AND METHODS
136
137 **Theory**
138
139 **General Theory.** Here, we model difference of substitution effects across populations as
140 Taylor expansions around a "focal" population. Using Kojima's method (Kojima 1959, 1961),
141 we put additive substitution effects as a function of differences in allele frequencies across
142 populations, "focal" additive substitution effects, and higher order (dominance, additive by
143 additive epistasis) statistical effects. From here, we show that the correlation of substitution
144 effects across populations is a function of their relationship (or genetic distance), the
145 additive, dominant and additive by additive genetic variances, the average heterozygosities

146 and average squared heterozygosities. All these parameters can be estimated in real
147 populations. In the following, we try to stick to Mäki-Tanila and Hill (2014) notation. Many
148 details are given as Appendix in the Supplementary Material.
149
150 Note that our procedure is general – it does not invoke neither any particular mechanism for
151 epistasis or dominance, nor knowledge of individual QTL effects and locations. We assume
152 that the population mean is a (possibly complex) function of QTL allele frequencies $\boldsymbol{p}$ and
153 QTL functional or biological effects. The latter, albeit unknown, are assumed to be constant
154 across populations - we therefore assume no genotype by environment interactions.
155
156 Consider the correlation $r\left(\alpha_i^b, \alpha_i^{b'}\right)$ of substitution effects $\alpha^{(b)}$ and $\alpha^{(b')}$ across respective
157 populations (breeds, heterotic groups, lines, or generations) b and b'. It is known that, in
158 presence of dominant and epistatic biological interactions, the value of α depends on the
159 allele frequencies and biological effects; for instance, $\alpha_1 = a_1 + (q_1 + p_1)d_1 +$
160 $(p_2 - q_2)[aa]_{12}$ for 2-locus epistasis (Fuerst *et al.* 1997), for $p_1 = 1 - q_1$ and $p_2 = 1 - q_2$
161 frequencies at loci 1 and 2, respectively. Inspired by this, we want to write the correlation
162 $r\left(\alpha_i^b, \alpha_i^{b'}\right)$ for a locus i as a function of respective allele frequencies, $\boldsymbol{p}^b$ and $\boldsymbol{p}^{b'}$, but in a
163 general manner, without defining a particular functional or biological gene action. We use a
164 first order Taylor series expansion to approximate additive substitution effects in a
165 population b and in population b', as a function of effects in another "focal" population (that
166 we call "0") and the differences in allelic frequencies across the three populations. By doing
167 this, it can be shown that the difference of substitution effects between two populations is
168 (approximately) a function of the genetic distance of the two populations to the focal

population, and the magnitude of dominance and second order epistatic variances in the focal population.

To derive additive substitution effects α as function of allelic frequencies $\boldsymbol{p}$, we use Kojima's definition of statistical effects as first, second… derivatives of the mean of the population (μ) as a function of p . We don't invoke any explicit function – we just presume that there is one, in other words, change in allele frequencies of the population implies change in the total average genotypic value. Using Kojima's method, the additive substitution effect at the *i*-th locus is the first derivative (Kojima 1959, 1961):

$$\alpha_i = \frac{1}{2}\frac{\partial \mu}{\partial p_i}$$

Higher order effects implying locus *i* are higher order derivatives of μ but we remark that they can equally be obtained as derivatives of α_i. The dominance deviation at the *i*-th locus (that we denote as d^* to distinguish from the biological or functional effect d (Falconer and Mackay 1996)) is:

$$d_i^* = -\frac{1}{4}\frac{\partial^2 \mu}{\partial p_i^2} = -\frac{1}{2}\frac{\partial \alpha_i}{\partial p_i}$$

The negative sign comes because the dominance deviation is usually understood as a feature of *heterozygosity,* in other words, it is of opposite sign than the increase of homozygosity in ∂p_i^2. Last, the epistatic pairwise deviation of locus i with j is

$$(\alpha\alpha)_{ij} = \frac{1}{4}\frac{\partial^2 \mu}{\partial p_i \partial p_j} = \frac{1}{2}\frac{\partial \alpha_i}{\partial p_j}$$

This is positive because it is the effect of increasing both p_i and p_j.

Kojima's method shows, therefore, that higher order effects of one locus are derivatives of lower order effects. With these elements we can make a Taylor order expansion of α_i around frequencies in the focal population, $\boldsymbol{p_0}$, so that $\boldsymbol{p} = \boldsymbol{p_0} + \boldsymbol{\epsilon}$, such that we create a function $\alpha_i(\boldsymbol{p})$. A meaningful value for $\boldsymbol{p_0}$ is a "focal" population for which there are estimates of additive, dominance and second order epistatic variances. In the Appendix, we show that the Taylor linear approximation of the substitution effects α_i^b and $\alpha_i^{b'}$ for two populations b and b' is:

$$\alpha_i^b \approx \alpha_i^0 + 2\epsilon_i^b(-d_i^{*0}) + 2\boldsymbol{\epsilon}^{(\boldsymbol{b})'}(\boldsymbol{\alpha\alpha})_i^0 \qquad (1)$$

$$\alpha_i^{b'} \approx \alpha_i^0 + 2\epsilon_i^{b'}(-d_i^{*0}) + 2\boldsymbol{\epsilon}^{(\boldsymbol{b'})'}(\boldsymbol{\alpha\alpha})_i^0$$

where we use differences in allele frequencies from the focal population, $\boldsymbol{\epsilon}^{(b)} = \boldsymbol{p}^{(b)} - \boldsymbol{p_0}$ and $\boldsymbol{\epsilon}^{(b')} = \boldsymbol{p}^{(b')} - \boldsymbol{p_0}$, d_i^{*0} is the statistical dominance deviation at the locus i and $(\boldsymbol{\alpha\alpha})_i^0$ is a vector containing epistatic substitution effects of locus i with the rest of loci.

From equation 1, the covariance across two populations b and b' of the two substitution effects α_i^b and $\alpha_i^{b'}$ of the locus i is a function of the covariance of $\boldsymbol{\epsilon}^{(b)}$ and $\boldsymbol{\epsilon}^{(b')}$ and is:

$$Cov\left(\alpha_i^{(b)}, \alpha_i^{(b')}\right) \tag{2}$$

$$= Cov\left(\alpha_i^0 + 2\epsilon_i^{(b)}(-d_i^{*0}) \ + 2\boldsymbol{\epsilon}^{(b)'}(\boldsymbol{\alpha\alpha})_i^0, \alpha_i^0\right.$$

$$\left. + 2\epsilon_i^{(b')}(-d_i^{*0}) \ + 2\boldsymbol{\epsilon}^{(b')'}(\boldsymbol{\alpha\alpha})_i^0\right)$$

$$= Var(\alpha_i^0) + 4Cov\left(\epsilon_i^{(b)}, \epsilon_i^{(b')}\right) Var(d_i^{*0})$$

$$+ 4tr\left(Cov\left(\boldsymbol{\epsilon}^{(b)}, \boldsymbol{\epsilon}^{(b')'}\right) Var((\boldsymbol{\alpha\alpha})_i^0)\right)$$

Where $Cov\left(\boldsymbol{\epsilon}^{(b)}, \boldsymbol{\epsilon}^{(b')'}\right)$ and $Var((\boldsymbol{\alpha\alpha})_i^0)$ are matrices. Note that we can drop terms including e.g. $Cov(\alpha^0, (\boldsymbol{\alpha\alpha})_i^0)$ as the different statistical effects (additive dominant, epistatic) are mutually orthogonal. Also, we assume $\boldsymbol{\epsilon}^{(b)}$ (change in allele frequencies) and the different effects d_i^{*0} and $(\boldsymbol{\alpha\alpha})_i^0$ to be independent. The equation 2 has different terms that are briefly presented here; more details are in the Appendix.

The variance of statistical additive, dominant and additive by additive effects are (Mäki-Tanila and Hill 2014):

$$Var(\alpha_i^0) = \sigma_\alpha^{(0)2} \ = \frac{\sigma_A^2}{n\overline{H}}$$

$$Var(d^{*0}) = \sigma_d^{(0)2} = \frac{\sigma_D^2}{n\overline{H^2}}$$

$$Var\left((\alpha\alpha)_{i,j,j>i}^0\right) = \sigma_{(\alpha\alpha)}^{(0)2} = \frac{\sigma_{AA}^2}{0.5n(n-1)\overline{HH}} \approx 4\frac{\sigma_{AA}^2}{n^2\overline{H}\overline{H}}$$

$$Var((\boldsymbol{\alpha\alpha})_i^0) \ \ = \boldsymbol{I} \otimes \sigma_{(\alpha\alpha)}^{(0)2} \approx 4\boldsymbol{I}\frac{\sigma_{AA}^2}{n^2\overline{H}\overline{H}}$$

for n the number of QTL loci and using functions of heterozygosities:

$$\overline{H} = E\left(2p_i(1-p_i)\right)$$

$$\overline{H^2} = E\big(4p_i(1-p_i)p_i(1-p_i)\big)$$

$$\overline{HH} = E_{i>j}\left(2p_i(1-p_i)2p_j\big(1-p_j\big)\right) \approx \frac{1}{2}\overline{H}\overline{H}$$

All variances and effects refer to a focal population (with sub/superindex "0") with allele frequencies p^0 and effects α^0. Note that we assume HWE and LE. Of course, we don't know allele frequencies or even numbers of true causal genes, but the first can be prudently guessed and the later cancel out in the following.

A relationship coefficient across populations b and b' is $\gamma_{b,b'} = 8Cov\left(p_i^{(b)}, p_i^{(b')}\right)$; also, $\gamma_{b,b'} = -4D^2$ where D^2 is the average square of the Euclidean distance (Robertson 1975; Bonhomme *et al.* 2010; Garcia-Baccino *et al.* 2017). Similarly, we introduce a measure of similarity $\gamma_{b,b'}^{\epsilon}$ (more details in the Appendix) such that

$$Cov\left(\epsilon_i^{(b)}, \epsilon_i^{(b')}\right) = \frac{1}{8}\gamma_{b,b'}^{\epsilon} = \frac{1}{8}\big(\gamma_{b,b'} - \gamma_{b,0} - \gamma_{b',0} + \gamma_{0,0}\big)$$

This is the remaining similarity of b and b' once similarity of both to "0" has been discounted If the focal ("0") population is population b, then $Cov\left(\epsilon_i^{(b')}, \epsilon_i^{(b')}\right) = Var\left(\epsilon_i^{(b')}\right) = \frac{1}{8}\big(\gamma_{b'} + \gamma_b - 2\gamma_{b,b'}\big)$. In fact, in this case $Cov\left(\epsilon_i^{(b')}, \epsilon_i^{(b')}\right) = \frac{1}{8}\big(\gamma_{b'} + \gamma_b - 2\gamma_{b,b'}\big)$ is proportional to the F_{st} differentiation index across b and b' such that the pairwise $F_{st} = \frac{\frac{1}{8}\left(\gamma_{b'}+\gamma_b-2\gamma_{b,b'}\right)}{1-\frac{1}{8}\left(\gamma_{b'}+\gamma_b+2\gamma_{b,b'}\right)}$. This is obtained from the definition of F_{st} from average relationships within and across populations (Caballero and Toro 2002; Caballero 2020) as detailed in the Supplementary Material. The main difference between the F_{st} and our relationship coefficients γ resides in the fact that F_{st} describes divergence, and is relative to the set of populations being considered, whereas γ coefficients describe similarity, and are

not relative to any particular set of populations; for instance, they can be defined and obtained for pairs of populations, to be later combined. Across all loci, $Cov\left(\boldsymbol{\epsilon}^{(b)}, \boldsymbol{\epsilon}^{(b')'}\right) = \boldsymbol{I}\frac{1}{8}\gamma^{\epsilon}_{b,b'}$. Thus, the covariance of changes in allele frequencies is a function of relatedness across populations or, equivalently, of their differentiation.

The term including $Cov\left(\boldsymbol{\epsilon}^{(b)}(\boldsymbol{\alpha\alpha})^0_i, \boldsymbol{\epsilon}^{(b')}(\boldsymbol{\alpha\alpha})^0_i\right)$ uses the covariance of quadratic forms (Searle 1971) (eq. 58) and is detailed in the Appendix. It assumes uncorrelated $\boldsymbol{\epsilon}^{(b)}$ and $(\boldsymbol{\alpha\alpha})^{05}_i$, and normality for both. It is customary in population genetics to assume normality for small changes in allele frequencies (Lewontin and Krakauer 1973; Bonhomme *et al.* 2010).

Combining expressions above, equation (2) becomes

$$
\begin{aligned}
Cov\left(\alpha^b_i, \alpha^{b'}_i\right) &= Var(\alpha^0_i) + 4Cov\left(\epsilon^{(b)}_i, \epsilon^{(b')}_i\right) Var(d^{*0}_i) \\
&\quad + 4tr\left(Cov\left(\boldsymbol{\epsilon}^{(b)}, \boldsymbol{\epsilon}^{(b')'}\right) Var((\boldsymbol{\alpha\alpha})^0_i)\right) \\
&= \frac{\sigma^2_A}{n\bar{H}} + 4\frac{\gamma^{\epsilon}_{b,b'}}{8}\frac{\sigma^2_D}{n\overline{H^2}} + 4n\frac{\gamma^{\epsilon}_{b,b'}}{8}4\frac{\sigma^2_{AA}}{n^2\bar{H}\bar{H}} \\
&= \frac{\sigma^2_A}{n\bar{H}} + \frac{1}{2}\gamma^{\epsilon}_{b,b'}\frac{\sigma^2_D}{n\overline{H^2}} + 2\gamma^{\epsilon}_{b,b'}\frac{\sigma^2_{AA}}{n\bar{H}\bar{H}} \\
&= \frac{1}{n}\left(\frac{\sigma^2_A}{\bar{H}} + \frac{1}{2}\gamma^{\epsilon}_{b,b'}\frac{\sigma^2_D}{\overline{H^2}} + 2\gamma^{\epsilon}_{b,b'}\frac{\sigma^2_{AA}}{\bar{H}\bar{H}}\right)
\end{aligned}
\tag{3}
$$

The quantities involved in Equation 3 are (1) the covariance in change of allelic frequencies $\boldsymbol{\epsilon}$ in b and b' (called $\gamma^{\epsilon}_{b,b'}$), which describes the similarity of populations b and b' using the focal population as a reference, and (2) the variance of statistical additive, dominant and

additive by additive effects. We have used $4tr\left(Cov\left(\boldsymbol{\epsilon}^{(b)}, \boldsymbol{\epsilon}^{(b')'}\right) Var((\boldsymbol{\alpha\alpha})_i^0)\right) =$

$4tr\left(\left(\boldsymbol{I}\frac{1}{8}\gamma_{b,b'}^{\epsilon}\right)\left(4\boldsymbol{I}\frac{\sigma_{AA}^2}{n^2\bar{H}\bar{H}}\right)\right) = 4n\left(\left(\frac{1}{8}\gamma_{b,b'}^{\epsilon}\right)\left(4\frac{\sigma_{AA}^2}{n^2\bar{H}\bar{H}}\right)\right) = 2\gamma_{b,b'}^{\epsilon}\frac{\sigma_{AA}^2}{n\bar{H}\bar{H}}$

Whereas

$$Var(\alpha_i^b) = \frac{1}{n}\left(\frac{\sigma_A^2}{\bar{H}} + \frac{1}{2}\gamma_b^{\epsilon}\frac{\sigma_D^2}{\overline{H^2}} + 2\gamma_b^{\epsilon}\frac{\sigma_{AA}^2}{\bar{H}\bar{H}}\right)$$

From here the correlation of α across populations is (factorizing out $\frac{1}{n}$)

$$r(\alpha_i^b, \alpha_i^{b'}) = \tag{4}$$

$$\frac{\frac{\sigma_A^2}{\bar{H}} + \frac{1}{2}\gamma_{b,b'}^{\epsilon}\frac{\sigma_D^2}{\overline{H^2}} + 2\gamma_{b,b'}^{\epsilon}\frac{\sigma_{AA}^2}{\bar{H}\bar{H}}}{\sqrt{\left(\frac{\sigma_A^2}{\bar{H}} + \frac{1}{2}\gamma_b^{\epsilon}\frac{\sigma_D^2}{\overline{H^2}} + 2\gamma_b^{\epsilon}\frac{\sigma_{AA}^2}{\bar{H}\bar{H}}\right)\left(\frac{\sigma_A^2}{\bar{H}} + \frac{1}{2}\gamma_{b'}^{\epsilon}\frac{\sigma_D^2}{\overline{H^2}} + 2\gamma_{b'}^{\epsilon}\frac{\sigma_{AA}^2}{\bar{H}\bar{H}}\right)}}$$

It can already be seen that $\frac{\sigma_A^2}{\bar{H}}$ is both on the numerator and denominator. If the focal population is population b, then $\gamma_b^{\epsilon} = \gamma_{b,b'}^{\epsilon} = 0$, but $\gamma_{b'}^{\epsilon} = (\gamma_{b'} + \gamma_b - 2\gamma_{b,b'})$ (as an aside, this is positive, which results in the correlation $r(\alpha_i^b, \alpha_i^{b'})$ being positive), and Equation 4 becomes

$$r(\alpha_i^b, \alpha_i^{b'}) = \frac{\frac{\sigma_A^2}{\bar{H}}}{\sqrt{\frac{\sigma_A^2}{\bar{H}}\left(\frac{\sigma_A^2}{\bar{H}} + \frac{1}{2}\gamma_{b'}^{\epsilon}\frac{\sigma_D^2}{\overline{H^2}} + 2\gamma_{b'}^{\epsilon}\frac{\sigma_{AA}^2}{\bar{H}\bar{H}}\right)}} = \frac{\sqrt{\frac{\sigma_A^2}{\bar{H}}}}{\sqrt{\frac{\sigma_A^2}{\bar{H}} + \frac{1}{2}\gamma_{b'}^{\epsilon}\frac{\sigma_D^2}{\overline{H^2}} + 2\gamma_{b'}^{\epsilon}\frac{\sigma_{AA}^2}{\bar{H}\bar{H}}}}$$

Factorizing out $\bar{H}$ we arrive to the slightly clearer expression

$$r\left(\alpha_i^b, \alpha_i^{b'}\right) = \frac{\sqrt{\sigma_A^2}}{\sqrt{\sigma_A^2 + \gamma_{b'}^{\epsilon}\left(\frac{1}{2}\frac{\bar{H}}{\overline{H^2}}\,\sigma_D^2 + 2\frac{\sigma_{AA}^2}{\bar{H}}\right)}} \tag{5}$$

$$= \frac{\sqrt{\sigma_A^2}}{\sqrt{\sigma_A^2 + \left(\gamma_{b'} + \gamma_b - 2\gamma_{b,b'}\right)\left(\frac{1}{2}\frac{\bar{H}}{\overline{H^2}}\,\sigma_D^2 + 2\frac{\sigma_{AA}^2}{\bar{H}}\right)}}$$

which shows well that the correlation is a function of relationships across populations $\left(\gamma_{b'} + \gamma_b - 2\gamma_{b,b'}\right)$ and weights of additive *vs.* non-additive variances.

Finally, our expression for $r\left(\alpha_i^b, \alpha_i^{b'}\right)$ includes values for the three main genetic variances (which can be estimated from data), measures of genetic distances (which can be estimated from data) and "good guesses" of the distribution of QTL effects – the latter can come either via markers (sequence or chip SNP) or guesses (*i.e.*, U-shaped, flat, etc.).

Thus, the algorithm to estimate *a priori* the correlation of α across populations b and b', $r\left(\alpha_i^b, \alpha_i^{b'}\right)$, is:

1. Define a focal population with existing estimates of statistical additive, dominance, and additive by additive variances (this focal population can actually be population b or b')
2. Compute average heterozygosity $\bar{H}$ and average squared heterozygosity $\overline{H^2}$ in the focal population, either from markers or from guesses.
3. Using markers, compute relationships γ within and across the three populations

4. Apply equation 4. If the focal population is population b or b', apply the equivalent Equation 5.

Covariance across generations within one breed. The definition above of two populations is general enough that we can consider any two populations, *e.g.* two breeds (Angus and Hereford), two strains (New Zealand Holstein and US Holstein) or two generations or time frames (*e.g.* animals born in 2000 vs. animals born in 2005, or animals born in 2000 vs. their descendants). There is evidence that across-generations genetic correlation decreases with time to values as low as 0.6 (Tsuruta *et al.* 2004; Haile-Mariam and Pryce 2015). As explained before, the across-generation genetic correlation does not depend only on the across-breed correlation of substitution effects, first because allele frequencies and substitution effects might be correlated due to selection, second because of genotype by environment interactions. Anyway, part of the across-generation genetic correlation can be accounted for by our model, based on the evolution of average coancestry in the breed.

We will talk about "generations" but ideas apply to pedigrees with overlapping generations as well. Consider that we called populations b and $b'$are animals born at time t_1 and t_2, with $t_2 > t_1$. Allele frequencies at generation t are (for a single locus) $p_t = \frac{1}{2n_t}\mathbf{1}'\boldsymbol{z}$ where $\boldsymbol{z}$ has gene content in the form {0,1,2}. The variance of p_t is $Var(p_t) = \frac{1}{4}\bar{A}_t 2pq = \frac{pq}{2}\bar{A}_t$ (because $Var(\boldsymbol{z}) = \boldsymbol{A}_t 2pq$), where $\boldsymbol{A}$ is the pedigree-based relationship matrix and $\boldsymbol{A}_t$ the submatrix corresponding to animals born at time t. Considering $\boldsymbol{A}^{(\gamma)}$ build with metafounders (Legarra *et al.* 2015) then $Var(p_t) = \bar{A}_t^{(\gamma)}/8$. Note that if there is a single metafounder then $\boldsymbol{A}^{(\gamma)} = \boldsymbol{A}\left(1 - \frac{\gamma}{2}\right) + \mathbf{1}\mathbf{1}'\gamma$. The covariance between two generations is, similarly,

$Cov(p_{t_1}, p_{t_2}) = \bar{\boldsymbol{A}}^{(\gamma)}_{t_1,t_2}/8$, so here $\bar{\boldsymbol{A}}^{(\gamma)}_{t_1,t_2}$ takes the role of the $\gamma_{b,b'}$ before. Leading to, if we take t_1 as the focal population,

$$Cov\left(\epsilon_i^{(t_2)}, \epsilon_i^{(t_2)}\right) = \frac{1}{8}\left(\bar{A}^{(\gamma)}_{t_2} + \bar{A}^{(\gamma)}_{t_1} - 2\bar{A}^{(\gamma)}_{t_1,t_2}\right)$$

This involves average relationships within and across generations. From here, the correlation of substitution effects across t_1 and t_2 is

$$r\left(\alpha_i^{t_2}, \alpha_i^{t_1}\right) = \frac{\sqrt{\sigma_A^2}}{\sqrt{\sigma_A^2 + \left(\bar{\boldsymbol{A}}_{(\gamma)t_2} + \bar{\boldsymbol{A}}_{(\gamma)t_1} - 2\bar{\boldsymbol{A}}_{(\gamma)t_1,t_2}\right)\left(\frac{1}{2}\frac{\bar{H}}{\overline{H^2}}\sigma_D^2 + 2\frac{\sigma_{AA}^2}{\bar{H}}\right)}} \quad (6)$$

Thus, the correlation of substitution effects decreases when there is large drift: high average relationships within each generation combined with low relationship across. This is the case for instance if parents of the next generation are very highly selected without restrictions in future inbreeding, resulting in a considerable change in allele frequencies over the generations.

Values of $\left(\bar{\boldsymbol{A}}_{(\gamma)t_2} + \bar{\boldsymbol{A}}_{(\gamma)t_1} - 2\bar{\boldsymbol{A}}_{(\gamma)t_1,t_2}\right)$ can be easily computed, even for large pedigrees, using Colleau (2002). This can be approximated (see Appendix) as $\bar{\boldsymbol{A}}_{(\gamma)t_2} + \bar{\boldsymbol{A}}_{(\gamma)t_1} - 2\bar{\boldsymbol{A}}_{(\gamma)t_1,t_2} \approx (t_2 - t_1)\Delta F_{(\gamma)}$ for small values of $\Delta F_{(\gamma)}$ and steady increase of inbreeding. Note that $\Delta F_{(\gamma)} = \left(1 - \frac{\gamma}{2}\right)\Delta F$ as $\boldsymbol{A}_{(\gamma)} = \boldsymbol{A}\left(1 - \frac{\gamma}{2}\right) + \mathbf{1}\mathbf{1}'\gamma$. This agrees with classical theory (Falconer and Mackay 1996): the variance in change of allele frequencies from one generation to the next is simply ΔF. Typical values of ΔF in livestock (rabbits, pigs, cattle or sheep) are at most 0.01 per generation (Welsh *et al.* 2010; García-Ruiz *et al.* 2016; E. N.

Fernandez, J.P. Sanchez, R. Martinez, A. Legarra, M. Baselga 2017; Rodríguez-Ramilo *et al.* 2019), although this is potentially changing with genomic selection. Substituting in Equation 6 we obtain the correlation of substitution effects as a function of increase in inbreeding:

$$r\left(\alpha_i^{t_2}, \alpha_i^{t_1}\right) = \frac{\sqrt{\sigma_A^2}}{\sqrt{\sigma_A^2 + \Delta F_{(\gamma)}\left(\frac{1}{2}\frac{\bar{H}}{\overline{H^2}}\sigma_D^2 + 2\frac{\sigma_{AA}^2}{\bar{H}}\right)}} \tag{7}$$

Examples

Estimates of correlation across populations using literature values. From values of the literature, the previous expressions, equations (5) and (7) were applied to obtain guesses of the correlation of substitution effects across different populations and generations within population. Estimates of statistical non-additive variation are scarce, as are estimates of distances across populations in terms of covariance of allele frequencies. Table 1 shows two pairs of populations in pigs (Xiang *et al.* 2016) and cattle (VanRaden *et al.* 2011; Legarra *et al.* 2015), and the corresponding relationships expressed as $\gamma_{b,b'} = 8Cov(p^b, p^{b'})$. In all cases, the reference allele has been randomized so $E(p) = 0.5$.

Table 1. Relationships within and across two pairs of populations

		Holstein	Jersey
Cattle	Holstein	0.55	0.48
	Jersey	0.48	0.77
		Landrace	Yorkshire
Pigs	Landrace	0.76	0.26
	Yorkshire	0.26	0.73

Estimates from cattle (VanRaden *et al.* 2011; Legarra *et al.* 2015) and pigs (Xiang *et al.* 2016).

Table 2 includes estimates for additive, dominance and epistatic variation iWe used estimates for milk yield in Simmenthal cattle (Fuerst and Sölkner 1994) as if it was Holstein,

and for litter size in a commercial pig line (Vitezica *et al.* 2018) as if it was Landrace, because we could not find estimates in the same populations.

Table 2. Variance component estimates, as ratio from phenotypic variance, from literature

Species	σ_A^2	σ_D^2	σ_{AA}^2
Cattle	0.20	0.09	0.15
Pigs	0.092	0.020	0.016

Estimates for milk yield in cattle (Fuerst and Sölkner 1994) and litter size in pigs (Vitezica *et al.* 2018).

Then, we assumed three distributions for allele frequencies, following Hill et al. (2008): a uniform distribution ("Uniform"), a U-shaped distributions with $f(p) \propto p^{-1}(1-p)^{-1}$ with effective population size of 50 ("Hill"). We also assumed a Beta(0.04, 0.04) distribution, which is an extreme U-shaped distribution ("Extreme") in which roughly 80% of the QTLs have minor allele frequency lower than 0.01. This results in respective values of $\overline{H} = (0.333, 0.107, 0.037)$ and $\overline{H^2} = (0.133, 0.018, 0.013)$. Using the values described above, we estimated the correlation of QTL effects across breeds, $r\left(\alpha_i^b, \alpha_i^{b'}\right)$, using equation (5).

Similarly, the correlation of QTL effects a certain number of generations away, $r\left(\alpha_i^{t_2}, \alpha_i^{t_1}\right)$, was obtained applying Equation (7). An increase $\Delta F = 0.01$ resulting in $\Delta F_\gamma = 0.01(1 - \gamma/2)$ per generation was assumed, for the same assumed variances and allele frequency distributions.

361 **Simulations.** The objective of the simulation is to verify our results considering different
362 kinds of non-additive biological gene action, and not to infer the values of the correlation of
363 substitution effects across populations as this requires realistic presumptions of the forms
364 and magnitudes of epistatic interactions, something that is largely unknown.
365
366 We simulated two divergent populations using QMSim (Sargolzaei and Schenkel 2009). First,
367 2000 generations of the historical population were generated with an effective population
368 size of 1500 individuals, followed by a gradual contraction during the following 500
369 generations to an effective population size of 100 individuals. It followed a gradual
370 expansion to a population size of 1500 during the last 500 generations. From this common
371 historical population, two divergent populations were selected in opposite directions for 5
372 generations of 1000 individuals each. Selection was for a trait with narrow-sense heritability
373 of 1, as a device to speed up divergence; however, for the purposes of this work we
374 simulated another, uncorrelated, trait (with a broad-sense heritability of 1). There were 30
375 chromosomes of 100 cM each and 66,000 segregating loci (60,000 SNPs and 6,000 QTLs) in
376 the first generation of the historical population. The number of segregating QTLs and SNPs in
377 the last generation of the historical population was 750 and 13,555, respectively.
378
379 To simulate non-additive biological gene action, we considered three scenarios: Complete
380 Dominance (within locus), 5 locus complementary epistasis and 5 locus additive by additive
381 epistasis. Complete dominance has the genotypic value of the heterozygote equal to one of
382 the homozygotes. Additive by additive epistasis can be understood as a pure multiplication
383 of gene contents. Complementary epistasis implies that a simultaneous homozygote
384 genotype at all loci results in a genotypic value, and any other combination in another

genotypic value. Explicit functional additive effects were not simulated. These scenarios are similar to Hill *et al.* (2008) but instead of considering 2 loci we consider 5 (details are given in the Appendix).

For epistasis, networks are internally epistatic but contribute additively to the total phenotype. Loci within a network were picked up at random within chromosomes, resulting in very low LD across loci within network. We considered 5 networks per chromosome yielding a total of 150 networks. We derived substitution effects α algebraically for the three different cases (details are given in the Appendix). We used those expressions to evaluate the true substitution effects, and to obtain the true value $r(\alpha^b, \alpha^{b'})$. At the end, we had two populations of 5000 individuals each (1000 individuals per generation). Using markers (alternatively, QTL loci) and simulated phenotypes, we estimated (1) variance components σ_A^2, σ_D^2 and σ_{AA}^2 using G-REML (Vitezica *et al.* 2018) in population 1 (all 5 generations totaling 5000 animals). We then computed, using markers (alternatively, QTL loci), relationships γ within and across populations and the functions of heterozygosities $\bar{H}$ and $\overline{H^2}$. Then, using Equation 4 we predicted $r(\alpha^b, \alpha^{b'})$ across populations 1 and 2 within generation 5, and we predicted $r(\alpha^{t_1}, \alpha^{t_2})$ within population 1 for $t_1 = 1$ and $t_2 = 2 \ldots 5$. The purpose of using, alternatively to genotypes at markers, genotypes at QTL was to verify or discard that results were due to omperfect Linkage Disequilibrium.

Data Availability Statement. Appendix associated with this manuscript is in Figshare. Code and files for simulations can be found in http://genoweb.toulouse.inra.fr/~alegarra/simulation_alpha.tar.gz . All other results can be reproduced using equations and figures from the text.

409

410 RESULTS

411

412 **Estimates of correlation across populations using literature values.** The results for
413 reasonable guesses of the distribution of QTL substitution effects are in Table 3. The
414 correlation of QTL substitution effects across breeds were roughly 0.6 for the "Uniform"
415 distribution, whereas the "Extreme" distribution resulted in numbers close to 0.4. These
416 numbers agree reasonably well with scarce estimates of genetic correlation (although,
417 strictly speaking, not the same thing) in the literature: 0.4 to 0.8 for production traits (Karoui
418 *et al.* 2012), 0.3 to 0.5 for milk yield (Legarra *et al.* 2014) and 0.6 for body condition (Porto-
419 Neto *et al.* 2015). It may be argued that these studies used SNPs instead of true QTLs;
420 however, Wientjes et al. (Wientjes *et al.* 2017) in a simulation study obtained unbiased
421 estimates of correlation across QTLs using SNP markers for a true correlation of 0.5.

422

423 Table 3: Estimates of correlations of QTL effects across breeds based on values from Tables 1
424 and 2 and Equation 5, for different distributions of QTL frequencies

Species	Uniform	Hill[1]	Extreme[2]
Cattle	0.60	0.39	0.25
Pigs	0.66	0.46	0.31
Maize	0.57	0.36	0.22

425 [1] U-shaped distribution with effective population size of 50

426 [2] Beta(0.04, 0.04) distribution

427

428 Concerning decrease of correlations across generations, results (Table 4) show a mild
429 decrease, and this decrease is more sensitive to the assumed distribution of allelic
430 frequencies than the previous ones concerning correlation across populations. There are no
431 explicit estimates of genomic correlation across generations, but drops in the accuracy of
432 genomic prediction have been reported (Wolc *et al.* 2011; Baloche *et al.* 2014; Liu *et al.*
433 2016).

434

435
436 Table 4: Correlation of QTL effects across time, based on values from Table 1 and 2 and
437 Equation 7, for different distributions of QTL frequencies

Species	Distance in generations	Uniform	Hill[1]	Extreme[2]
Cattle	1	0.98	0.95	0.88
	2	0.97	0.90	0.79
	5	0.92	0.80	0.63
	10	0.85	0.67	0.50
Pigs	1	1.00	0.99	0.97
	2	0.99	0.98	0.94
	5	0.98	0.94	0.88
	10	0.96	0.89	0.79

438 [1] U-shaped distribution with effective population size of 50

439 [2] Beta(0.04, 0.04) distribution

440

441 **Simulations.** Estimated variance components (expressed as ratios from total phenotypic
442 variance σ^2) with marker genotypes and with QTL genotypes are shown in Table 5. The part
443 of non-additive variance increases with the complexity of the system, although this is in part
444 because we chose on purpose very artificial examples. In general, the two estimates agree,
445 although with markers we estimate non-zero epistatic additive by additive variance in the
446 case of complete dominance, where it is zero. This is due to imperfect LD between QTL and
447 markers and LD across QTLs and also across markers (Campos *et al.* 2019). Indeed, when we
448 estimated variance components with QTLs, this epistatic variance disappeared as shown in
449 the bottom three rows of Table 5.
450

Table 5. Estimated Ratios of additive, dominance and additive by additive variances σ_A^2/σ^2 , σ_D^2/σ^2 and σ_{AA}^2/σ^2 (standard errors) in population 1, estimated with markers (three top rows) and QTLs (three bottom rows)

	σ_A^2/σ^2	σ_D^2/σ^2	σ_{AA}^2/σ^2
Marker-based			
Complete Dominance	0.709 (0.017)	0.130 (0.011)	0.144 (0.031)
Complementary Epistasis	0.440 (0.025)	0.144 (0.016)	0.309 (0.047)
Additive by additive	0.232 (0.025)	0.040 (0.015)	0.587(0.053)
QTL-based			
Complete Dominance	0.676 (0.009)	0.324 (0.009)	0.00001(0.004)
Complementary Epistasis	0.512 (0.019)	0.145 (0.011)	0.248 (0.022)
Additive by additive	0.265 (0.020)	0.009 (0.006)	0.520(0.036)

Estimates of correlation across substitution effects are presented in Table 6 (across population) and 7 (within population). Our estimates tend to agree with true values, but they underestimate the true values. Still, they clearly tell apart one scenario with low correlation (i.e. additive by additive) from one with high correlation (i.e. complete dominance). Using QTL genotypes for estimation instead of markers yield similar results, so incomplete linkage disequilibrium does not seem to be the reason.

Table 6. True and estimated $r(\alpha^b, \alpha^{b'})$ across populations 1 and 2, generation 5

	True $r(\alpha^b, \alpha^{b'})$	Estimated $r(\alpha^b, \alpha^{b'})$ with markers	Estimated $r(\alpha^b, \alpha^{b'})$ with QTL
Complete Dominance	0.981	0.945	0.970
Complementary Epistasis	0.933	0.842	0.854
Additive by additive	0.76	0.60	0.532

Table 7. True and estimated $r(\alpha^b, \alpha^{b'})$ across generations 1 and 5, population 1

	True $r(\alpha^b, \alpha^{b'})$	Estimated $r(\alpha^b, \alpha^{b'})$ with markers	Estimated $r(\alpha^b, \alpha^{b'})$ with QTL
Complete Dominance	0.993	0.976	0.989
Complementary Epistasis	0.975	0.926	0.943
Additive by additive	0.899	0.777	0.780

DISCUSSION

The fact that most genetic variation is statistically additive is well known. However, this does not imply equality of statistical additive effects across populations, something that is possibly reflected in the difficulty of predicting across (distinct) populations, either in animal or human genetics. Only recently the role of non-additive biological gene actions in this difficulty has started to be recognized (Dai *et al.* 2020; Duenk *et al.* 2020). Here we confirm and reinforce their findings. These authors presented result of simulations, where the problem is that the results are scenario specific and do not allow a good appraisal of the different factors.

In this work, we present a general, formal framework that does not depend on specific hypothesis on gene action or order of the epistatic interactions. In our derivation, we approached *high-order functional* epistasis by Taylor expansions, leading to expressions that involve only *low-order statistical* additive epistasis (actually pairwise). Including one extra term in our Taylor expansion would include *three-loci statistical* epistasis and so on, but extra terms would lead to more difficult expressions (covariance of allele frequencies across loci and populations), and it is expected that the magnitude of the statistical effects is smaller and smaller with higher orders of interactions (Mäki-Tanila and Hill 2014). We find reasonable agreement with our simulation and values from literature. However, we have indeed not thoroughly studied by simulation the quality of our approximation, which would require considerable thinking on how to simulate epistasis in very general manners.

Three main factors influence the correlation of substitution effects between populations $r(\alpha^b, \alpha^{b'})$: the genetic similarity of the two populations, the extend of additive, dominance and additive by additive variance, and the distribution of allele frequencies at QTL. We consider that showing explicitly these three factors is an achievement, as their role is clear from previous work in simulated and real data, (Martin *et al.* 2019; Dai *et al.* 2020; Duenk *et al.* 2020). Now we discuss these three factors.

The distance across populations is conveniently summarized by relationships γ, which describe covariances of allelic frequencies across populations, and are related to F_{st} indexes. On ideal populations, these γ relationships depend on the divergence time t but also on the effective population size N, as shown by the expression $\frac{\gamma_{b,b'}}{8} = Cov(p^b, p^{b'}) \approx 1 - \exp(-t/2N)$ (Bonhomme *et al.* 2010). In equation (4), to estimate $r(\alpha^b, \alpha^{b'})$, relationship enters into the form $(\gamma_{b'} + \gamma_b - 2\gamma_{b,b'})$ which as explained is proportional to the F_{st} index. If the relationship across populations is very small, $\gamma_{b,b'} = 0$ which results in $(\gamma_{b'} + \gamma_b - 2\gamma_{b,b'})$ bounded between 0 (all allele frequencies are 0.5) and 4 (alleles are fixed). The more diverse each of the populations ($\gamma_{b'}, \gamma_b$), and the more similar one to the other, the higher $r(\alpha^b, \alpha^{b'})$ will be.

The factors $\frac{1}{2}\frac{\bar{H}}{\overline{H^2}}$ and $\frac{2}{\bar{H}}$ are weighting factors on dominance and additive by additive variances. If the allele frequencies are modelled using symmetric $Beta(a, a)$ distributions (see Appendix), these become $\frac{1}{2}\frac{\bar{H}}{\overline{H^2}} = \frac{1}{4\frac{a+1}{a}\left(1+\frac{a+2}{2a+2}\left(\frac{a+3}{2a+3}-2\right)\right)}$ and $\frac{2}{\bar{H}} = 1 + \frac{1}{2a}$. The first is bounded between 1.5 (for U-shaped distributions) and 1 (for peaked distributions), so that

514 dominance variation does not play a big role in the difference between substitution effects
515 across populations unless dominance variation is way much larger that additive variation,
516 something that seems unlikely based on theory and estimates. However, the second weight,
517 due to epistasis, is not bounded and is large for small values of a (*e.g.* 13.5 for
518 Beta(0.04,0.04), which reflects that functional epistasis plays a strong role in additive
519 variation for U-shaped distributions (Hill *et al.* 2008). This result also reinforces the findings
520 of Duenk et al. (2020) and Dai et al. (2020) that it is mainly biological epistasis that generates
521 changes in additive substitution effects across populations. For instance, when one
522 population drifts, the additive by additive statistical variation enters into additive variation
523 (Hill *et al.* 2006; Walsh and Lynch 2018).
524
525 Thus, $r\left(\alpha_i^b, \alpha_i^{b'}\right)$ will be very low only if (1) populations are distinct, (2) there is *statistical*
526 epistatic variation and (3) QTL distributions are U-shaped.
527
528 In most conceivable cases σ_{AA}^2 and σ_D^2 are of magnitudes roughly similar to additive variances
529 (Fuerst and Sölkner 1994; Palucci *et al.* 2007; Hill *et al.* 2008; Mäki-Tanila and Hill 2014;
530 Vitezica *et al.* 2018), and relationships across populations are moderate. Then, for
531 reasonable assumptions about allele distribution of QTLs, we have shown that the
532 correlation of substitution effects across populations is typically around 0.6, which basically
533 agrees with scarce estimates of genetic correlations across populations available in the
534 literature. Our results seem rather robust to different distributions of allele frequencies.
535 These values are low enough so that, for practical purposes, predictions across populations
536 are of little interest, moreover if populations are not closely related as is often the case. This
537 is of practical relevance, *e.g.* for genomic predictions in livestock improvement, but also in

538 human genetics *e.g.* for the use of European-based Polygenic Risk Scores in individuals from
539 other ancestries (Martin *et al.* 2019)
540
541 Similar results apply to the same populations across generations, in which case the
542 correlation of substitution effects across generations goes from 0.99 (next 1-2 generations)
543 to 0.70 (5-10 generations). This illustrates that, even if genomic predictions would use QTLs
544 or markers in thight LD with QTLs, there would still be a need for a continuous system of
545 data collecting and re-estimation of effects.
546
547
548 In the simulations we confirmed that our estimates are reasonably good although slightly
549 biased downwards. Still, we find that, for the challenging forms of non-additive gene action
550 that we simulated, the results are quite good. For instance, an explicit functional value was
551 only included in the case of complete dominance. The additive by additive epistasis is highly
552 extreme (e.g. substitution effects change sign across spectra of allelic frequency). The other
553 form of epistasis, complementary epistasis, has genotypic value depending on every locus
554 being at a particular homozygous genotype and is deemed to be a biological mechanism that
555 contributes to variance stabilization (Mackay 2014). Complementary epistasis generates
556 more additive variation that additive by additive epistasis and, as expected, it leads to higher
557 similarity of substitution effects across populations.
558
559 Another result of our theory is that it allows, at least in principle, to assess if genetic
560 correlation across breeds is due to gene-by-environment or gene-by-gene interaction,
561 without going to a fully parametric analysis. For instance, Legarra *et al*. (2014) estimated a

genetic correlation of 0.5 for Latxa Cara Rubia - Manech Tête Rousse, two breeds that have frequent exchanges and are almost identical (in this case, $\Gamma = \begin{pmatrix} \gamma_b & \gamma_{b,b'} \\ \gamma_{b,b'} & \gamma_{b'} \end{pmatrix} = \begin{pmatrix} 0.49 & 0.47 \\ 0.47 & 0.51 \end{pmatrix}$. This would result in a value of $r(\alpha^b, \alpha^{b'})$ much higher than 0.5. A possibility is that the low value of 0.5 is due to GxE interaction, as the production systems may differ. This opens new perspectives.

CONCLUSIONS

We presented a coherent, approximate theory, that does not invoke any particular mechanism of gene action, to explain and appraise the change in magnitude of (additive) QTL substitution effects across populations and generations. The theory gives good approximate estimates of this correlation, that needs to be otherwise explicitly estimated. More importantly, the theory shows that the main sources for the change of effects are relationships across populations, magnitudes of additive and first-order non-additive variances (dominance and additive by additive), and spectra of allele frequencies. These findings provide better understanding of the properties of genomic prediction methods and of quantitative genetics in general.

ACKNOWLEDGMENTS

Part of this work was while Y. Wientjes was visiting the GenPhySE unit at INRAE, Toulouse, financed by the Netherlands Organisation of Scientific Research (NWO). Authors thank INRA

585 SelGen metaprograms Episel and Epifun. This project has received funding from the
586 European Unions' Horizon 2020 Research & Innovation programme under grant agreement
587 N°772787 -SMARTER. We are grateful to the Genotoul Bioinformatics Platform Toulouse
588 Midi- Pyrenees (Bioinfo Genotoul) for providing computing and storage resources.

589

590

591

592 REFERENCES

593

594 Aiello D., K. Patel, and E. Lasagna, 2018 The myostatin gene: an overview of mechanisms of
595 action and its relevance to livestock animals. Anim. Genet. 49: 505–519.
596 https://doi.org/10.1111/age.12696

597 Baloche G., A. Legarra, G. Sallé, H. Larroque, J. M. Astruc, *et al.*, 2014 Assessment of accuracy
598 of genomic prediction for French Lacaune dairy sheep. J. Dairy Sci. 97: 1107–1116.

599 Bonhomme M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, *et al.*, 2010 Detecting Selection
600 in Population Trees: The Lewontin and Krakauer Test Extended. Genetics 186: 241–
601 262. https://doi.org/10.1534/genetics.110.117275

602 Bonifati V., P. Rizzu, M. J. van Baren, O. Schaap, G. J. Breedveld, *et al.*, 2003 Mutations in the
603 DJ-1 Gene Associated with Autosomal Recessive Early-Onset Parkinsonism. Science
604 299: 256–259. https://doi.org/10.1126/science.1077209

605 Caballero A., and M. A. Toro, 2002 Analysis of genetic diversity for the management of
606 conserved subdivided populations. Conserv. Genet. 3: 289.

Caballero A., 2020 *Quantitative Genetics*. Cambridge University Press.

Campos G. de los, D. A. Sorensen, and M. A. Toro, 2019 Imperfect Linkage Disequilibrium Generates Phantom Epistasis (& Perils of Big Data). G3 Genes Genomes Genet. 9: 1429–1436. https://doi.org/10.1534/g3.119.400101

Carlborg Ö., L. Jacobsson, P. Åhgren, P. Siegel, and L. Andersson, 2006 Epistasis and the release of genetic variation during long-term selection. Nat. Genet. 38: 418–420. https://doi.org/10.1038/ng1761

Dai Z., N. Long, and W. Huang, 2020 Influence of Genetic Interactions on Polygenic Prediction. G3 Genes Genomes Genet. 10: 109–115. https://doi.org/10.1534/g3.119.400812

Duenk P., P. Bijma, M. P. L. Calus, Y. C. J. Wientjes, and J. H. J. van der Werf, 2020 The Impact of Non-additive Effects on the Genetic Correlation Between Populations. G3 Genes Genomes Genet. 10: 783–795. https://doi.org/10.1534/g3.119.400663

Fernandez E. N., J.P. Sanchez, R. Martinez, A. Legarra, M. Baselga, 2017 Role of inbreeding depression, non-inbred dominance deviations and random year-season effect in genetic trends for prolificacy in rabbit closed lines. J. Anim. Breed. Genet. 6: 441–452.

Falconer D. S., and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*. Longman New York.

Fuerst C., and J. Sölkner, 1994 Additive and Nonadditive Genetic Variances for Milk Yield, Fertility, and Lifetime Performance Traits of Dairy Cattle. J. Dairy Sci. 77: 1114–1125. https://doi.org/10.3168/jds.S0022-0302(94)77047-8

628 Fuerst C., J. W. James, J. Sölkner, and A. Essl, 1997 Impact of dominance and epistasis on the
629 genetic make-up of simulated populations under selection: a model development. J.
630 Anim. Breed. Genet. 114: 163–175. https://doi.org/10.1111/j.1439-
631 0388.1997.tb00502.x

632 Garcia-Baccino C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic, *et al.*, 2017
633 Metafounders are related to Fst fixation indices and reduce bias in single-step
634 genomic evaluations. Genet. Sel. Evol. 49: 34. https://doi.org/10.1186/s12711-017-
635 0309-2

636 García-Ruiz A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-López, *et al.*, 2016
637 Changes in genetic selection differentials and generation intervals in US Holstein
638 dairy cattle as a result of genomic selection. Proc. Natl. Acad. Sci. 113: E3995–E4004.

639 Gautier M., A. Capitan, S. Fritz, A. Eggen, D. Boichard, *et al.*, 2007 Characterization of the
640 DGAT1 K232A and Variable Number of Tandem Repeat Polymorphisms in French
641 Dairy Cattle. J. Dairy Sci. 90: 2980–2988. https://doi.org/10.3168/jds.2006-707

642 Grobet L., L. J. Royo Martin, D. Poncelet, D. Pirottin, B. Brouwers, *et al.*, 1997 A deletion in
643 the bovine myostatin gene causes the double–muscled phenotype in cattle. Nat.
644 Genet. 17: 71–74. https://doi.org/10.1038/ng0997-71

645 Haile-Mariam M., and J. E. Pryce, 2015 Variances and correlations of milk production,
646 fertility, longevity, and type traits over time in Australian Holstein cattle. J. Dairy Sci.
647 98: 7364–7379. https://doi.org/10.3168/jds.2015-9537

Hayes B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard, 2009 Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41: 51.

Hill W. G., N. H. Barton, and M. Turelli, 2006 Prediction of effects of genetic drift on variance components under a general model of epistasis. Theor. Popul. Biol. 70: 56–62. https://doi.org/10.1016/j.tpb.2005.10.001

Hill W. G., M. E. Goddard, and P. M. Visscher, 2008 Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. PLOS Genet. 4: e1000008. https://doi.org/10.1371/journal.pgen.1000008

Karoui S., M. J. Carabaño, C. Díaz, and A. Legarra, 2012 Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet Sel Evol 44: 39.

Kojima K., 1959 Role of epistasis and overdominance in stability of equilibria with selection. Proc. Natl. Acad. Sci. U. S. A. 45: 984–989.

Kojima K.-I., 1961 Effects of dominance and size of population on response to mass selection*. Genet. Res. 2: 177–188. https://doi.org/10.1017/S0016672300000689

Lande R., and R. Thompson, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124: 743–756.

Legarra A., G. Baloche, F. Barillet, J.-M. Astruc, C. Soulas, *et al.*, 2014 Within-and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J. Dairy Sci. 97: 3200–3212.

Legarra A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal, 2015 Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. Genetics 200: 455–468. https://doi.org/10.1534/genetics.115.177014

Lewontin R. C., and J. Krakauer, 1973 Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms. Genetics 74: 175–195.

Liu Z., H. Alkhoder, F. Reinhardt, and R. Reents, 2016 Accuracy and bias of genomic prediction for second-generation candidates. Interbull Bull. 50: 17-23.

Lynch M., and B. Walsh, 1998 *Genetics and analysis of quantitative traits*. Sinauer associates.

Mackay T. F. C., 2014 Epistasis and Quantitative Traits: Using Model Organisms to Study Gene-Gene Interactions. Nat. Rev. Genet. 15: 22–33. https://doi.org/10.1038/nrg3627

Mackay T. F. C., 2015 Epistasis for Quantitative Traits in Drosophila, pp. 47–70 in *Epistasis: Methods and Protocols*, Methods in Molecular Biology. edited by Moore J. H., Williams S. M. Springer, New York, NY.

MacLeod I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, *et al.*, 2016 Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17: 144. https://doi.org/10.1186/s12864-016-2443-6

Magwire M. M., A. Yamamoto, M. A. Carbone, N. V. Roshina, A. V. Symonenko, *et al.*, 2010 Quantitative and Molecular Genetic Analyses of Mutations Increasing Drosophila Life Span. PLOS Genet. 6: e1001037. https://doi.org/10.1371/journal.pgen.1001037

Mäki-Tanila A., and W. G. Hill, 2014 Influence of Gene Interaction on Complex Trait Variation with Multilocus Models. Genetics 198: 355–367. https://doi.org/10.1534/genetics.114.165282

Martin A. R., M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, *et al.*, 2019 Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. 51: 584–591. https://doi.org/10.1038/s41588-019-0379-x

Meuwissen T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Moghaddar N., M. Khansefid, J. H. J. van der Werf, S. Bolormaa, N. Duijvesteijn, *et al.*, 2019 Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. Genet. Sel. Evol. 51: 72. https://doi.org/10.1186/s12711-019-0514-2

Palucci V., L. R. Schaeffer, F. Miglior, and V. Osborne, 2007 Non-additive genetic effects for fertility traits in Canadian Holstein cattle (Open Access publication). Genet. Sel. Evol. 39: 181. https://doi.org/10.1186/1297-9686-39-2-181

Porto-Neto L. R., W. Barendse, J. M. Henshall, S. M. McWilliam, S. A. Lehnert, *et al.*, 2015 Genomic correlation: harnessing the benefit of combining two unrelated populations

for genomic selection. Genet. Sel. Evol. 47: 84. https://doi.org/10.1186/s12711-015-0162-0

Robertson A., 1975 Gene Frequency Distributions as a Test of Selective Neutrality. Genetics 81: 775–785.

Rodríguez-Ramilo S., J. M. Elsen, and A. Legarra, 2019 Inbreeding and effective population size in French dairy sheep: Comparison between genomic and pedigree estimates. J. Dairy Sci. 102: 4227–4237.

Roos A. P. W. de, B. J. Hayes, and M. E. Goddard, 2009 Reliability of Genomic Predictions Across Multiple Populations. Genetics 183: 1545–1553. https://doi.org/10.1534/genetics.109.104935

Rupp R., P. Senin, J. Sarry, C. Allain, C. Tasca, *et al.*, 2015 A Point Mutation in Suppressor of Cytokine Signalling 2 (Socs2) Increases the Susceptibility to Inflammation of the Mammary Gland while Associated with Higher Body Weight and Size and Higher Milk Production in a Sheep Model. PLOS Genet. 11: e1005629. https://doi.org/10.1371/journal.pgen.1005629

Searle S., 1971 *Linear models*. New York, NY (EUA). John Wiley.

Streit M., N. Neugebauer, T. H. E. Meuwissen, and J. Bennewitz, 2011 Short communication: evidence for a major gene by polygene interaction for milk production traits in German Holstein dairy cattle. J. Dairy Sci. 94: 1597–1600. https://doi.org/10.3168/jds.2010-3834

Tsuruta S., I. Misztal, T. J. Lawlor, and L. Klei, 2004 Modeling final scores in US Holsteins as a function of year of classification using a random regression model. Livest. Prod. Sci. 91: 199–207. https://doi.org/10.1016/j.livprodsci.2003.09.016

VanRaden P., K. Olson, G. Wiggans, J. Cole, and M. Tooker, 2011 Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 94: 5673–5682.

Vitezica Z. G., A. Reverter, W. Herring, and A. Legarra, 2018 Dominance and epistatic genetic variances for litter size in pigs using genomic models. Genet. Sel. Evol. 50: 1–8.

Walsh B., and M. Lynch, 2018 *Evolution and Selection of Quantitative Traits*. Oxford University Press.

Welsh C. S., T. S. Stewart, C. Schwab, and H. D. Blackburn, 2010 Pedigree analysis of 5 swine breeds in the United States and the implications for genetic conservation. J. Anim. Sci. 88: 1610–1618. https://doi.org/10.2527/jas.2009-2537

Wientjes Y. C. J., P. Bijma, J. Vandenplas, and M. P. L. Calus, 2017 Multi-population Genomic Relationships for Estimating Current Genetic Variances Within and Genetic Correlations Between Populations. Genetics 207: 503–515. https://doi.org/10.1534/genetics.117.300152

Wolc A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, *et al.*, 2011 Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet Sel Evol 43: 23.

Xiang T., B. Nielsen, G. Su, A. Legarra, and O. F. Christensen, 2016 Application of single-step genomic evaluation for crossbred performance in pig. J. Anim. Sci. 94: 936–948. https://doi.org/10.2527/jas.2015-9930

Xiang T., O. F. Christensen, and A. Legarra, 2017 Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. J. Anim. Sci. 95: 1472–1480.