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 the 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 a 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 variances; relationships (Nei's minimum distances or Fst indexes); and assumed heterozygosities. 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. Simulation results confirmed our derivations. Slight biases were observed, depending on the non-additive mechanism and the reference allele. Our derivations are useful to understand and forecast the possibility of prediction across populations and the similarity of GWAS effects.

INTRODUCTION

One of the aims of quantitative genetics is to provide methods for prediction, for instance genomic prediction (prediction of livestock breeding values or of crop performance) or polygenic risk score (risk of a disease in humans). These predictions would ideally work across a range of populations (different breeds, future generations). Ideally, the prediction goes through a process of identifying causal genes, estimating their effects in some population, and transposing these effects to newly genotyped individuals (Lande and Thompson 1990; Meuwissen *et al.* 2001). These "gene effects" are substitution effects – the regression of the own phenotype (for polygenic risk scores) or expected progeny phenotypes (for estimated breeding values) on gene content at the locus. Being able to use substitution effects at causal genes across populations and generations is a goal of genomic prediction, QTL detection and also of causal mutation finding (Grisart *et al.* 2002).

There are several obstacles for these aims. Finding and validating causal genes and understanding their functional mechanism is extremely difficult (Grobet *et al.* 1997; Bonifati *et al.* 2003; Rupp *et al.* 2015). In practice, predictions are done using SNP markers using statistical genetics techniques. In livestock, use of markers results in very good predictions within populations, but mediocre (at best) predictions across populations, even with very sophisticated techniques (Hayes *et al.* 2009; Karoui *et al.* 2012; Porto-Neto *et al.* 2015; MacLeod *et al.* 2016). Indeed, livestock and human genetics empirical results show decreasing predictive ability with increasing genetic distance across distinct populations or generations (Liu *et al.* 2016; Martin *et al.* 2019). In humans, there is strong correlation between GWAS effect estimates across human populations, with typical values around 0.82

70 (Marigorta and Navarro 2013; Shi *et al.* 2021). The lack of perfect linkage disequilibrium (LD)
71 across markers and genes has been claimed to be a reason for this decrease in accuracy.
72 Adding extra information (more dense maps, biological prior information) should result in a
73 better choice of markers close to causal genes, and therefore in a boost in predictive abilities
74 across populations (Roos *et al.* 2009; MacLeod *et al.* 2016). However, in practice, the
75 increase in predictive ability across populations is small at best (MacLeod *et al.* 2016;
76 Moghaddar *et al.* 2019). This is against results from simulations (Roos *et al.* 2009; MacLeod
77 *et al.* 2016) – but a problem in simulations is that typically gene effects are assumed to be
78 biologically additive and therefore constant across populations.
79
80 We argue that, although imperfect LD is a likely cause for not being able to predict across
81 populations, it is not the only one. In fact, substitution (statistical) additive gene action is not
82 homogeneous across populations, even for exactly the same causal mutation. Examples in
83 livestock genetics include myostatin gene (Aiello *et al.* 2018) or DGAT1 (Gautier *et al.* 2007).
84 For instance, in the latter, the "K" allele had rather different substitution effects across
85 breeds: -611, -142 and -351 kg of milk for the respective breeds Montbéliarde, Normande
86 and Holstein, for a trait with a genetic standard deviation of ~600 kg. Although part of these
87 differences may be due to genotype by environment interactions, it is also plausible that this
88 is due to epistasis or dominance; for instance this is the case in DGAT1 (Streit *et al.* 2011).
89 For instance, in the 5-loci epistatic network in Carlborg et al. (2006) some substitution
90 effects at genes switch signs depending on genetic background. In Drosophila, estimated
91 substitution effects of P-element insertions switched signs depending on the genetic
92 background (Magwire *et al.* 2010; Mackay 2015).
93

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 does imply 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, assuming linkage equilibrium (LE), 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 human studies, the correlation is estimated through the meta-analysis of GWAS statistics (Marigorta and Navarro 2013; Shi *et al.* 2021). However, in addition to empirical results, some theory to describe the resemblance of substitution effects across different populations would be helpful to (1) better understand and quantify the change in *true* substitution effects (*i.e.* if true genes instead of markers were being used), (2) give upper bounds for genomic prediction across populations/generations, and (3) allow *a priori* planning of genomic predictions *i.e.* to include or not different subpopulations.

The aim of this work is to develop a theory to understand and predict, under a neutral scenario, the extent of change of substitution effects across space (breeds, lines) and across time (generations), without invoking or assuming specific modes of biological gene action. As a result, we obtain explicit estimators that are functions of additive, dominance and additive by additive variances, genetic distances across populations, and distributions of allele frequencies. We check and illustrate our theory using published results and simulations considering dominance and epistasis (additive by additive and complementary) from 5-loci interactions. Main factors affecting correlation of substitution effects across populations are their genetic distance and the extent of additive by additive variation, which is rarely large.

MATERIALS AND METHODS

Theory

General Theory. Here, we model difference of substitution effects across two populations as Taylor expansions around one of them, the "focal" population. Using Kojima's method (Kojima 1959, 1961), we put additive substitution effects as a function of differences in allele frequencies across populations, "focal" additive substitution effects, and second order (dominance, additive by additive epistasis) statistical effects. From here, we show that the correlation of substitution effects across populations is (approximately) a function of their differentiation (or genetic distance), the additive, dominant and additive by additive genetic variances, the average heterozygosities and average squared heterozygosities. All these parameters can be estimated in real populations. In the following, we try to stick to Mäki-Tanila and Hill (2014) notation. Many details are given as Appendix in the Supplementary Material. Main notation is presented in Table 1.

Table 1. Notation.

$\alpha_i^b, \alpha_i^{b'}$	Substitution effect of locus i in population b, in population b'	$\sigma^2_{\alpha,b}$	Variance of substitution effects in population b
d_i^{*b}	Dominance deviation at locus i in population b	$\sigma^2_{d,b}, \mu_{d,b}$	Variance and mean of dominance deviations in population b
$(\alpha\alpha)_{ij}^b$	Additive by additive effect of loci i, j at population b	$\sigma^2_{(\alpha\alpha,b)}$	Variance of additive by additive effects in

			population b
$(\boldsymbol{\alpha\alpha})_i^b$	Vector of additive by additive effects of locus i with all other loci in population b		
$p_i^b, p_i^{b'}$	Allele frequency of locus i in population b, in population b'		
$\boldsymbol{p}^b, \boldsymbol{p}^{b'}$	Vectors of allele frequencies in population b, in population b'		
$\overline{H_b}$	Average heterozygosity at population b, $\overline{H}_b = E\left(2p_i^b(1-p_i^b)\right)$		
$\overline{H_b^2}$	Average squared heterozygosity at population b, $\overline{H_b^2} = E\left(\left(2p_i^b(1-p_i^b)\right)^2\right)$		
$\overline{HH_b}$	Average product of heterozygosities at population b, $\overline{HH_b} = E_{i>j}\left(4p_i^b(1-p_i^b)p_j^b(1-p_j^b)\right)$		
ϵ_i	Difference in allele frequencies at locus i, $\epsilon_i = p_i^{b'} - p_i^b$	σ_ϵ^2	Variance of the difference in allele frequencies across all loci
$D_{b,b'}$	Nei's minimum genetic distance		
$\sigma_A^2, \sigma_D^2, \sigma_{AA}^2$	Additive, Dominance and Additive by Additive variances (in population b)		

Note that our procedure is general – it does not invoke any particular mechanism for epistasis or dominance, nor knowledge of individual QTL effects and locations. We assume that the population mean is a (possibly complex) function of QTL allele frequencies $\boldsymbol{p}$ and QTL functional or biological effects. The latter, albeit unknown, are assumed to be constant across populations - we therefore do not consider genotype by environment interactions.

Consider the correlation $r(\alpha_i^b, \alpha_i^{b'})$ of substitution effects α^b and $\alpha^{b'}$ across respective populations (breeds, heterotic groups, lines, or generations) b and b'. For simplicity, the allele to which α refers is random – it can be either the wild or the mutant allele. In this manner the average value of α is 0, even in presence of deleterious mutations. It is known that, in presence of dominant and epistatic biological interactions, the value of α depends on the allele frequencies and biological effects; for instance, $\alpha_1 = a_1 + (q_1 - p_1)d_1 + (p_2 - q_2)[aa]_{12}$ for 2-locus epistasis (Fuerst *et al.* 1997), for $p_1 = 1 - q_1$ and $p_2 = 1 - q_2$ frequencies at loci 1 and 2, respectively. Inspired by this, we want to write the correlation $r(\alpha_i^b, \alpha_i^{b'})$ for a locus i as a function of vectors of respective allele frequencies, $\boldsymbol{p}^b$ and $\boldsymbol{p}^{b'}$, but in a general manner, without defining a particular functional or biological gene action. We use a first order Taylor series expansion to approximate additive substitution effects in a population b', as a function of effects in another "focal" population b and their distance. By doing this, it can be shown that the difference of substitution effects between two populations is (approximately) a function of the genetic distance of the two populations, and the magnitude of dominance and second order epistatic variances in the focal population b.

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 statistical effects implying locus *i* (i.e. dominance deviations and epistatic interactions) can be represented by higher order partial derivatives of μ or equivalently 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. Note that interaction of order k implies k-ith order derivative with scaling factor $1/2^k$.

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, $\boldsymbol{p}^b$, so that $\boldsymbol{p}^{b'} = \boldsymbol{p}^b + \boldsymbol{\epsilon}$, such that from values of α in b and changes in allele frequencies $\boldsymbol{\epsilon} = \boldsymbol{p}^{b'} - \boldsymbol{p}^b$ we create a function

$\alpha_i^{b'} \approx f(\alpha_i^b, \boldsymbol{\epsilon})$. In the Appendix (section 1.1), we show that the Taylor linear approximation of the substitution effects of population b': $\alpha_i^{b'}$, from effects from populations b, α_i^b, d_i^{*b}, $(\boldsymbol{\alpha\alpha})_i^b$ is:

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

where we use differences in allele frequencies $\boldsymbol{\epsilon}$, d_i^{*b} is the statistical dominance deviation at the locus i and $(\boldsymbol{\alpha\alpha})_i^b$ is a vector containing epistatic substitution effects of locus i with the rest of loci. By convention we assign $(\alpha\alpha)_{ii}^b = 0$.

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:

$$Cov(\alpha_i^b, \alpha_i^{b'}) \approx Cov(\alpha_i^b, \alpha_i^b + 2\epsilon_i(-d_i^*) + 2\boldsymbol{\epsilon}'(\boldsymbol{\alpha\alpha})_i^b) = Var(\alpha_i^b) \qquad (2)$$

The equality holds because terms $Cov\left(\alpha_i^b, 2\epsilon_i(-d_i^{*b})\right)$ and $Cov(\alpha_i^b, 2\boldsymbol{\epsilon}'(\boldsymbol{\alpha\alpha})_i^b)$ are null, given that the different statistical effects (α_i^b, d_i^{*b} and $(\boldsymbol{\alpha\alpha})_i^b$) are mutually orthogonal by construction.

Thus $r(\alpha_i^b, \alpha_i^{b'}) = \sqrt{\frac{Var(\alpha_i^b)}{Var(\alpha_i^b)Var(\alpha_i^{b'})}}$, and now we need the variance of $\alpha_i^{b'}$ as a function of effects in population b, this is (see the Appendix, section 1.2 for details):

$$Var(\alpha_i^{b'}) \approx Var(\alpha_i^b + 2\epsilon_i(-d_i^{*b}) + 2\boldsymbol{\epsilon}'(\boldsymbol{\alpha\alpha})_i^b) \quad (3)$$

$$= Var(\alpha_i^b) + 4Var(\epsilon_i)\left(Var(d_i^{*b}) + E^2(d_i^{*b})\right)$$

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

where $Var(\boldsymbol{\epsilon}')$ and $Var((\boldsymbol{\alpha\alpha})_i^b)$ are matrices:

$$Var(\boldsymbol{\epsilon}') = \begin{pmatrix} Var(\epsilon_1) & Cov(\epsilon_1,\epsilon_2) & & Cov(\epsilon_1,\epsilon_n) \\ Cov(\epsilon_2,\epsilon_1) & Var(\epsilon_2) & & Cov(\epsilon_2,\epsilon_n) \\ & & \dots & \\ Cov(\epsilon_n,\epsilon_1) & & & Var(\epsilon_n) \end{pmatrix}$$

which represents how loci differ. A locus that may diverge a lot has high $Var(\epsilon)$; two loci in strong linkage will show non-zero $Cov(\epsilon_1,\epsilon_2)$. Also

$$Var((\boldsymbol{\alpha\alpha})_i^b) = \begin{pmatrix} Var((\alpha\alpha)_1^b) & Cov((\alpha\alpha)_1^b,(\alpha\alpha)_2^b) & & Cov((\alpha\alpha)_1^b,(\alpha\alpha)_n^b) \\ Cov((\alpha\alpha)_2^b,(\alpha\alpha)_1^b) & Var((\alpha\alpha)_2^b) & & Cov((\alpha\alpha)_2^b,(\alpha\alpha)_n^b) \\ & & \dots & \\ Cov((\alpha\alpha)_n^b,(\alpha\alpha)_1^b) & & & Var((\alpha\alpha)_n^b) \end{pmatrix}$$

where $Var((\alpha\alpha)_1^b)$ contains the variance of the epistatic effect of marker i with marker 1, and so on. We assume $Var((\boldsymbol{\alpha\alpha})_i^b) \quad = \boldsymbol{I}\,\sigma^2_{(\alpha\alpha,b)}$, i.e. epistatic terms $(\boldsymbol{\alpha\alpha})_i^b$ have a variance $\sigma^2_{(\alpha\alpha,b)}$ and are a priori uncorrelated to each other. In the previous product $tr\left(Var(\boldsymbol{\epsilon}')Var((\boldsymbol{\alpha\alpha})_i^b)\right)$, the use of the trace implies that off-diagonal elements of $Var(\boldsymbol{\epsilon}')$ are ignored. Thus, assuming that all diagonal elements of $Var(\boldsymbol{\epsilon}')$ have the same common variance σ^2_ϵ, this results in $4tr\left(Var(\boldsymbol{\epsilon}')Var((\boldsymbol{\alpha\alpha})_i^b)\right) = 4n\sigma^2_\epsilon\sigma^2_{(\alpha\alpha,b)}$

The expression $Var(\epsilon_i d_i^{*b}) = Var(\epsilon_i)\left(Var(d_i^{*b}) + E^2(d_i^{*b})\right)$ is detailed in the Appendix (section 1.2), and it shows that both the variability of dominance deviations $Var(d_i^{*b}) = \sigma_{d,b}^2$ and its mean $E(d_i^{*b}) = \mu_{d,b}$ which, if not zero, can be understood as the basis of inbreeding depression, enter into the expression. Also, we assume $\boldsymbol{\epsilon}$ (change in allele frequencies, but not the allele frequencies themselves) and the different non-additive effects d_i^{*b} and $(\boldsymbol{\alpha\alpha})_i^b$ to be independent (uncorrelated). This makes sense as loci may be selected for additive effects but not for non-additive effects.

In expression (3) we need the variances of the different genetic effects and the variance of changes in allele frequencies. We address these two terms in turn.

We factorize the variance of statistical additive, dominant and additive by additive effects as follows (Mäki-Tanila and Hill 2014; Vitezica *et al.* 2017):

$$Var(\alpha_i^b) = \sigma_{\alpha,b}^2 = \frac{\sigma_A^2}{n\overline{H_b}}$$

$$Var(d_i^{*b}) + E^2(d_i^{*b}) = \sigma_{d,b}^2 + \mu_{d,b}^2 = \frac{\sigma_D^2}{n\overline{H_b^2}}$$

(the latter is shown in the Appendix, section 1.4) and

$$Var\left((\alpha\alpha)_{i,j,j>i}^b\right) = \sigma_{(\alpha\alpha,b)}^2 = \frac{\sigma_{AA}^2}{n(n-1)\overline{H_b}\,\overline{H_b}} \approx 2\frac{\sigma_{AA}^2}{n^2\overline{H_b}\,\overline{H_b}}$$

for n the number of QTL loci and using functions of heterozygosities (more details in the Appendix, section 1.5):

$$\overline{H_b} = E\left(2p_i^b(1-p_i^b)\right)$$

$$\overline{H_b^2} = E\left(4p_i^b(1-p_i^b)p_i^b(1-p_i^b)\right)$$

$$\overline{HH_b} = E_{i>j}\left(2p_i^b(1-p_i^b)2p_j^b(1-p_j^b)\right) \approx \frac{1}{2}\overline{H_b}\,\overline{H_b}$$

Here we have assumed independence of QTL allele frequencies and QTL effects. All variances and effects, as well as heterozygosities H_b, refer to the focal population b with allele frequencies p^b and effects α^b. Note that we assume HWE and LE within both b and b'. 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.

Consider the scalar σ_ϵ^2, the variance of differences of allele frequency. In fact, $\sigma_\epsilon^2 = Var(p^{b'} - p^b) = E\left((p^{b'} - p^b)^2\right) - E(p^{b'} - p^b)^2 = E\left((p^{b'} - p^b)^2\right)$ because $E(p^{b'} - p^b) = 0$ when averaged across loci. Therefore, $\sigma_\epsilon^2 = E\left((p^{b'} - p^b)^2\right)$ which corresponds to Nei's "minimum genetic distance" (Nei 1987; Caballero and Toro 2002), and, accordingly, will be called $D_{b,b'} = \sigma_\epsilon^2$. The value of $D_{b,b'}$ can be estimated from marker data as $\widehat{D}_{b,b'} = \frac{1}{n}\sum(p_i^{b'} - p_i^b)^2$, although it is sensitive to the spectra of polymorphisms (e.g. SNP chips vs sequencing). In addition, $D_{b,b'}$ is also the numerator of the F_{ST} fixation index, e.g. Hudson's(1992) $F_{ST} = \frac{E\left((p_i^{b'} - p_i^b)^2\right)}{E\left(p_i^{b'}(1-p_i^b) + p_i^b(1-p_i^{b'})\right)}$ with an estimator $\widehat{F}_{ST} = \frac{\frac{1}{n}\sum_i\left((p_i^{b'} - p_i^b)^2\right)}{\frac{1}{n}\sum_i\left(p_i^{b'}(1-p_i^b) + p_i^b(1-p_i^{b'})\right)}$ (Hudson *et al.* 1992; Bhatia *et al.* 2013). In principle, $D_{b,b'}$ and F_{ST} can be estimated from markers, but also from evolutionary distances and effective population size, or both (Weir and Hill 2002; Bonhomme *et al.* 2010). Now we can rewrite the last term of equation (3) as

$$4tr\left(Var(\boldsymbol{\epsilon}')Var\left((\boldsymbol{\alpha\alpha})_i^b\right)\right) = 4n\left(\sigma_\epsilon^2\sigma_{(\alpha\alpha,b)}^2\right) = 4nD_{b,b'}\frac{2\sigma_{AA}^2}{n^2\,\overline{H}_b\,\overline{H}_b}$$

Combining expressions above, equation (3) becomes

$$Var\left(\alpha_i^{b'}\right) = \tag{4}$$

$$\frac{\sigma_A^2}{n\overline{H}_b} + 4D_{b,b'}\frac{\sigma_D^2}{n\overline{H}_b^2} + 4nD_{b,b'}2\frac{\sigma_{AA}^2}{n^2\overline{H}_b\,\overline{H}_b} =$$

$$\frac{1}{n}\left(\frac{\sigma_A^2}{\overline{H}_b} + 4D_{b,b'}\frac{\sigma_D^2}{\overline{H}_b^2} + 8D_{b,b'}\frac{\sigma_{AA}^2}{\overline{H}_b\,\overline{H}_b}\right)$$

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

$$r\left(\alpha_i^b, \alpha_i^{b'}\right) \approx \frac{Cov\left(\alpha_i^b, \alpha_i^{b'}\right)}{\sqrt{\left(Var\left(\alpha_i^b\right)\right)\left(Var\left(\alpha_i^{b'}\right)\right)}} = \tag{5}$$

$$\frac{Var\left(\alpha_i^b\right)}{\sqrt{\left(Var\left(\alpha_i^b\right)\right)\left(Var\left(\alpha_i^{b'}\right)\right)}} = \frac{\sqrt{Var\left(\alpha_i^b\right)}}{\sqrt{Var\left(\alpha_i^{b'}\right)}} =$$

$$\frac{\sqrt{\frac{\sigma_A^2}{\overline{H}_b}}}{\sqrt{\frac{\sigma_A^2}{\overline{H}_b} + 4D_{b,b'}\frac{\sigma_D^2}{\overline{H}_b^2} + 8D_{b,b'}\frac{\sigma_{AA}^2}{\overline{H}_b\,\overline{H}_b}}}$$

Factorizing out $\overline{H_b}$ we arrive to the slightly clearer expression

$$r(\alpha_i^b, \alpha_i^{b'}) \approx \frac{\sqrt{\sigma_A^2}}{\sqrt{\sigma_A^2 + D_{b,b'}\left(4\frac{\bar{H}_b}{\overline{H_b^2}}\,\sigma_D^2 + 8\frac{1}{\bar{H}_b}\sigma_{AA}^2\right)}} \tag{6}$$

which shows well that the correlation is a function of distance across populations $(D_{b,b'})$ and weights of additive *vs.* non-additive variances. Note that $D_{b,b'}$ is always positive, which implies that $0 < r(\alpha_i^b, \alpha_i^{b'}) < 1$ as expected.

The quantities involved in equation (6) are (1) Nei's "minimum genetic distance" $D_{b,b'}$, which describes the similarity of populations b and b', (2) the variance of statistical additive, dominant and additive by additive effects at the individual level σ_A^2, σ_D^2, σ_{AA}^2 in population b (3) first and second moments of heterozygosities $\bar{H}_b$, $\overline{H_b^2}$. All these values can be, in principle, estimated from data or "prudently guessed". For the particular case of heterozygosities, SNP markers are a poor choice and it is likely better to use a guess based on sequence or evolutionary processes (coalescence).

From the definition of F_{ST} and assuming $\bar{H}_b \approx \bar{H}_{b'}$, $D_{b,b'} \approx \frac{F_{ST}}{1-F_{ST}}\bar{H}_b$ (detailed in the Appendix, section 1.6), leading to

$$r(\alpha_i^b, \alpha_i^{b'}) \approx \frac{\sqrt{\sigma_A^2}}{\sqrt{\sigma_A^2 + \frac{F_{ST}}{1-F_{ST}}\left(4\frac{(\bar{H}_b)^2}{\overline{H_b^2}}\,\sigma_D^2 + 8\sigma_{AA}^2\right)}} \tag{7}$$

The advantage of the F_{ST} is that it is more robust to the spectra of allele frequencies used to estimate it (Bhatia *et al.* 2013). If we further assume that dominance variance σ_D^2 is negligible, we can write a neat expression of the correlation in terms of F_{ST}:

$$r(\alpha_i^b, \alpha_i^{b'}) \approx \frac{\sqrt{\sigma_A^2}}{\sqrt{\sigma_A^2 + \frac{8F_{ST}}{1-F_{ST}}\sigma_{AA}^2}} \approx \sqrt{1 - \frac{8F_{ST}}{1-F_{ST}}\frac{\sigma_{AA}^2}{\sigma_A^2}} \quad (8)$$

The second approximation involving $\frac{1}{1+x} \approx 1 - x$. This expression (8) plainly tells that the squared correlation of gene substitution effects across two population is (to a few degrees of approximation) a linear function of the similarity of populations and the additive by additive variance. To our knowledge, these results showing the importance of the different factors on the difference between substitution effects had been shown before only through simulations.

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

1. Estimate in population b
    a. additive, dominance, and additive by additive variances
    b. average heterozygosity $\overline{H_b}$ and average squared heterozygosity $\overline{H_b^2}$
2. Estimate Nei's distance $D_{b,b'}$ and/or F_{ST} of the two populations
3. Apply equation (6) or (7).

Consideration of directionality of substitution effects

In Evolutionary Genetics it is reasonable to think that most mutations are deleterious, thus with a negative effect of the mutant allele. In Medical Genetics reports of estimated substitution effects are also often done in terms of "susceptible" alleles. In both cases $E(\alpha) \neq 0$ or even $\alpha > 0$ for all loci. In both cases α is "oriented" and has no zero mean. In the theory above we have considered that α is the effect of a randomly drawn allele, which leads to $E(\alpha) = 0$ and enormously simplifies the algebra. In order to consider oriented α, we propose to transform the estimate of $r(\alpha_i^b, \alpha_i^{b'})$ into $r(|\alpha_i^b|, |\alpha_i^{b'}|)$, the correlation of the absolute values. Assuming that α follows a normal distribution, $|\alpha|$ follows a so-called folded normal distribution. From here, $r(|\alpha_i^b|, |\alpha_i^{b'}|)$ is obtained from $r(\alpha_i^b, \alpha_i^{b'})$ using expressions (not detailed here) in (Kan and Robotti 2017), conveniently programmed in the R package MomTrunc (Galarza *et al.* 2020). The specific R function is in the Appendix, section 1.7, and we will call it r2rabs().

Covariance across generations within one population. 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 (many) generations to values as low as 0.6 (Tsuruta *et al.* 2004; Haile-Mariam and Pryce 2015). Part of this is likely due to genotype by environment interactions. Anyway, part of the across-generation genetic correlation could be due to changes in the allele frequency due to drift, and therefore it can be accounted for by our model, based on the evolution of average coancestry in the breed. We develop this next.

We will talk about "generations" but ideas apply to pedigrees with overlapping generations as well. Consider that what we previously called populations b and b' are animals born at time t_1 and t_2, with $t_2 > t_1$. Equation (7) becomes $r(\alpha_i^{t_1}, \alpha_i^{t_2}) \approx \frac{\sqrt{\sigma_A^2}}{\sqrt{\sigma_A^2 + \frac{F_{ST}}{1-F_{ST}}\left(4\frac{(\bar{H}_{t_1})^2}{\overline{H_{t_1}^2}}\sigma_D^2 + 8\sigma_{AA}^2\right)}}$.

In pedigree-based context, the F_{ST} is simply half the increase in average relationship from t_1 to t_2 (Powell *et al.* 2010) which is approximately equal to the increase in inbreeding from $t_1 + 1$ to $t_2 + 1$, which in turn is approximately $(t_2 - t_1)\Delta F$ for small values of ΔF and steady increase of inbreeding. Thus, $F_{ST} \approx \frac{(t_2 - t_1)\Delta F}{2}$ to obtain

$$r(\alpha_i^{t_1}, \alpha_i^{t_2}) \approx \frac{\sqrt{\sigma_A^2}}{\sqrt{\sigma_A^2 + \frac{\frac{(t_2 - t_1)\Delta F}{2}}{1 - \frac{(t_2 - t_1)\Delta F}{2}}\left(4\frac{(\bar{H}_{t_1})^2}{\overline{H_{t_1}^2}}\sigma_D^2 + 8\sigma_{AA}^2\right)}}$$

Assuming like in (8) small values of dominance variance and of F_{ST}, we obtain

$$r(\alpha_i^b, \alpha_i^{b'}) \approx \sqrt{1 - 4(t_2 - t_1)\Delta F \frac{\sigma_{AA}^2}{\sigma_A^2}}$$

Thus, the correlation of substitution effects decreases when there is large drift, reflected in high values of ΔF. 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.

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; Fernández *et al.* 2017; Rodríguez-Ramilo *et al.* 2019), although this is potentially changing with genomic selection.

Examples

Estimates of correlation across populations using literature values. From values of the literature, the previous expressions, equations (6) 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. The differentiation F_{ST} Jersey and Holstein breeds of cattle obtained from (VanRaden *et al.* 2011) is 0.06 and among Landrace and Yorkshire is 0.16 (Xiang *et al.* 2017).

Table 2 includes literature estimates for additive, dominance and epistatic variation. We used estimates for milk yield in Simmental 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. These estimates (as most estimates of non-additive variances) are inaccurate and are presented here only as examples. In the particular case of cattle, the estimated value of σ^2_{AA} is higher than usually expected *a priori* (Hill *et al.* 2008), so we take it as an example of an extreme, but not impossible, case.

377 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

378 Estimates for milk yield in cattle (Fuerst and Sölkner 1994) and litter size in pigs (Vitezica *et*
379 *al.* 2018).
380
381 Then, we assumed three distributions for QTL allele frequencies, following Hill *et al.* (2008):
382 a uniform distribution ("Uniform"), and a U-shaped distributions with $f(p) \propto p^{-1}(1-p)^{-1}$
383 with effective population size of $Ne = 50$ ("Hill"), bounded at $\left[\frac{1}{2Ne}, 1 - \frac{1}{2Ne}\right]$. We also
384 assumed a Beta(0.04, 0.04) distribution, which is an extreme U-shaped distribution
385 ("Extreme") in which roughly 80% of the QTLs have minor allele frequency lower than 0.01.
386 This results in respective values of $\overline{H_b} = (0.333, 0.107, 0.037)$ and
387 $\overline{H_b^2} = (0.133, 0.018, 0.013)$. Using the values described above, we estimated the correlation
388 of QTL substitution effects across breeds, $r\left(\alpha_i^b, \alpha_i^{b'}\right)$, using equation (7).
389
390 Similarly, the correlation of QTL substitution effects a certain number of generations away,
391 $r\left(\alpha_i^{t_2}, \alpha_i^{t_1}\right)$, was obtained applying Equation (7). An increase $\Delta F = 0.01$ per generation was
392 assumed, for the same assumed variances and allele frequency distributions.
393
394 **Simulations.** The objective of the simulation is to verify our results considering different
395 kinds of non-additive biological gene action, and not to infer the values of the correlation of

396 substitution effects across populations as this requires realistic presumptions of the forms
397 and magnitudes of epistatic interactions, something that is largely unknown.
398
399 We used *macs* (Chen *et al.* 2009) to simulate a "cattle" scenario of two domestic cattle
400 breeds which diverged from a common ancestral population, in the lines of (Pérez-Enciso
401 2014; Pérez-Enciso *et al.* 2015), where a large population had a 10-fold reduction bottleneck
402 (domestication) 2000 generations ago and a split into two populations *t* generations ago,
403 where we made *t* oscillate between 0 and 100 by steps of 10. Parameters in the simulation
404 were tailored (Pérez-Enciso 2014; Pérez-Enciso *et al.* 2015) to mimic observed levels of
405 diversity in cattle (Gibbs *et al.* 2009) and lead to F_{ST} values between 0 and 0.15. We
406 considered 100 DNA stretches of 300 Kb each. We simulated 200 individuals per population,
407 from which we obtained allele frequencies per population. Details are provided in the
408 Appendix, section 1.8. This provided ~400,000 segregating loci with realistic allele
409 frequencies (L-shaped distribution of allele frequencies).
410
411
412 To simulate non-additive biological gene action, we considered three scenarios. All of them
413 involved 5000 single marker polymorphisms drawn at random from the simulated ones with
414 no particular restriction. First scenario was Complete Dominance (within locus); second one
415 was 1000 networks of 5 loci in Complementary Epistasis, and the third one 1000 networks of
416 5 loci with Multiplicative Epistasis (additive by additive). We also simulated networks of 2
417 and of 10 loci with similar results (not shown). These scenarios are similar to Hill *et al.* (2008)
418 but instead of considering 2 loci we consider 5. We use these forms of epistasis, as the first
419 two ones may be interpreted as biologically meaningful gene actions, and the third one is an

extreme case for the change in substitution effect across populations. In addition, all three scenarios are analytically tractable. Complete dominance has the genotypic value of the heterozygote equal to one of the homozygotes, e.g. the presence of a single copy of the "good" allele is enough to, say, avoid the disease. Complementary epistasis can be seen as a multi-loci dominance, e.g. disease happens when there is a recessive deleterious genotype at *any* of the loci. Additive by additive epistasis can be understood as a pure multiplication of gene contents and has no good biological interpretation. Networks contribute additively to the total phenotype. Note that, by definition, functional gene action is the same across all populations. From the description of the non-additive gene action and some algebra in the Appendix, we are able to derive analytically the values of α, d^* and $(\alpha\alpha)$ in each population and variances σ_A^2, σ_D^2 and σ_{AA}^2. Details are given in the Appendix, section 1.9. For each value of t, we made 10 replicates and averaged the results.

From the true substitution effects α^b and $\alpha^{b'}$ derived above, we obtained the true value $r(\alpha^b, \alpha^{b'})$. Note that because of the coalescent simulation, all reference alleles (i.e. with frequency p) are "mutant" ones, and due to assumed dominance and epistatic actions, α are negative by construction (i.e. the mutant allele is deleterious). To conciliate this fact with our derivations, that assume that α refers to a random allele and has null means, we did two things: (1) compute a "random allele" version of $\alpha^b, \alpha^{b'}$ in which α was changed sign for "odd" loci, and (2) estimate $r(\alpha^b, \alpha^{b'})$ using the transformation of normal distribution into folded normal (Kan and Robotti 2017), i.e. the r2rabs() function mentioned above. Thus, we have two estimands, the correlation for the "mutant allele" effect $r_{mutant}(\alpha^b, \alpha^{b'})$ (which corresponds *e.g.* to typical use in Evolutionary and Medical Genetics) and the correlation for

the "random allele" effect $r_{random}(\alpha^b,\alpha^{b'})$ (which corresponds, *e.g.* to genomic selection and some GWAS). By construction, $r_{mutant}(\alpha^b,\alpha^{b'}) < r_{random}(\alpha^b,\alpha^{b'})$. For instance, in one of the scenarios $r_{mutant}(\alpha^b,\alpha^{b'}) = 0.67$ and $r_{random}(\alpha^b,\alpha^{b'}) = 0.85$.

Now we describe the estimators. We considered either the use of "all" polymorphisms, or of a "SNP" selection in which MAF>0.01 across both populations simultaneously (roughly 40,000 polymorphisms), as this corresponds to typical SNP panels in genetic improvement. Then we used the equation (6) either for "all" or for "SNP", using their frequencies to obtain $D_{b,b'}$, $\bar{H}_b$ and $\overline{H_b^2}$. In "all", the spectra of allele frequencies of polymorphisms and of QTL is the same, but not in "SNP". As "SNP" tends to be biased, we also considered the equation (7) using F_{ST} estimated from SNPs and $\bar{H}_b = 0.107$, and $\overline{H_b^2} = 0.018$ from the "Hill" U-shaped distribution mentioned above. Note that F_{ST} is more robust to the spectra of allele frequencies used to estimate it. Thus, we obtained three estimators for "random allele" $r_{random}(\alpha^b,\alpha^{b'})$: $\hat{r}_{all}$, $\hat{r}_{SNP}$ and $\hat{r}_{SNPFst}$, and also the three corresponding estimators for "mutant allele" $r_{mutant}(\alpha^b,\alpha^{b'})$ applying function r2rabs() to $\hat{r}_{all}$, $\hat{r}_{SNP}$ and $\hat{r}_{SNPFst}$.

Data Availability Statement. Appendix associated with this manuscript is in https://figshare.com/articles/online_resource/Appendix_final_v4_pdf/14509941 . Code and files for simulations can also be found in https://figshare.com/articles/software/scripts_zip/14509956 . All other results can be reproduced using equations and figures from the text.

466 RESULTS

467

468 **Estimates of correlation across populations using literature values.** The results for
469 reasonable guesses of the distribution of allele frequencies at QTL are in Table 3. The
470 correlation of QTL substitution effects across breeds was roughly 0.85 across both species
471 (pigs and cattle) and all assumed distributions of gene frequencies. These numbers are
472 higher than scarce estimates of genetic correlations in the literature: 0.4 to 0.8 for
473 production traits (Karoui *et al.* 2012), 0.3 to 0.5 for milk yield (Legarra *et al.* 2014) and 0.6 for
474 body condition (Porto-Neto *et al.* 2015), and we consider that this is because across-breeds
475 genetic correlations also include genotype by environment interactions. It may be argued
476 that these studies used SNPs instead of true QTLs; however, Wientjes et al. (2017) in a
477 simulation study obtained unbiased estimates of correlation across QTLs using SNP markers
478 for a true correlation of 0.5, and (Cuyabano *et al.* 2018) showed analytically that when SNP
479 markers yield correct estimates of true relationships at QTL, estimates of genetic parameters
480 are unbiased.

481

482 Table 3: Estimates of correlations of QTL effects across breeds based on values from Table 2
483 and Equation 7, for different distributions of QTL frequencies

Species	Uniform	Hill[1]	Extreme[2]
Cattle	0.83	0.82	0.84
Pigs	0.87	0.85	0.88

484 [1] U-shaped distribution with effective population size of 50

485 [2] Beta(0.04, 0.04) distribution

486

487 Concerning decrease of correlations across generations, results (Table 4) show a slow
488 decrease, and this decrease is, again, not sensitive to the assumed distribution of allelic
489 frequencies. Values in Table 4, quite close to 1, somehow disagree with few existing
490 estimates of across-generation genetic correlations, clearly lower than 1 (Tsuruta *et al.* 2004;
491 Haile-Mariam and Pryce 2015). Again, part of the correlation lower than 1 is probably due to
492 non-additive gene action, and part due to genotype by environment interactions.
493
494
495 Table 4: Correlation of QTL effects within breed across time, based on values from Table 1
496 and Equation 7, for different distributions of QTL frequencies

Species	Distance in generations	Uniform	Hill[1]	Extreme[2]
Cattle	1	0.98	0.97	0.98
	2	0.97	0.94	0.97
	5	0.93	0.87	0.93
	10	0.86	0.78	0.87
Pigs	1	1.00	0.99	1.00
	2	0.99	0.98	0.99
	5	0.98	0.96	0.97
	10	0.96	0.93	0.93

497 [1] U-shaped distribution with effective population size of 50
498 [2] Beta(0.04, 0.04) distribution
499

Simulations. Simulated variance components (expressed as ratios from total genetic variance σ^2) with marker genotypes and with QTL genotypes are shown in Table 5. All scenarios yield high additive genetic variances, as expected.

Table 5. Additive, dominance and additive by additive variances in the simulated population 1

	additive	dominance	additive by additive
Complete Dominance	252	113	0
Complementary Epistasis	130	53	12
Multiplicative	219	0	122

True simulated values of correlation across substitution effects $r_{random}(\alpha^b, \alpha^{b'})$ and their estimates $\hat{r}_{all}$, $\hat{r}_{SNP}$ and $\hat{r}_{SNPFst}$, are presented in Figure 1 (for α defined for random alleles) and Figure 2 (for α defined for mutant alleles). Generally speaking, Figure 1 applies to non-fitness related traits (for "random") and Figure 2 to fitness-related traits (for "mutant"). As predicted by our derivations, there is a clear and almost linear decrease of $r(\alpha^b, \alpha^{b'})$ with increasing values of F_{ST}.

In Figure 1 (for α defined for random alleles) it can be observed that our expressions (6-7) tend to slightly over-estimate $r_{random}(\alpha^b, \alpha^{b'})$ for Complete Dominance, and Complementary Epistasis. The estimates are quite good for Multiplicative Epistasis. Estimators using SNP markers are slightly less accurate than using the complete

polymorphism spectra, and using F_{ST} from SNP markers, coupled with a guess of heterozygosities, is similar to using SNP alone to infer both distances and heterozygosities.

Figure 2 (for α defined for mutant alleles) presents values of $r_{mutant}(\alpha^b, \alpha^{b'})$ and estimates $\hat{r}_{all}$, $\hat{r}_{SNP}$ and $\hat{r}_{SNPFst}$. Estimating $r_{mutant}(\alpha^b, \alpha^{b'})$ is more difficult than estimating $r_{random}(\alpha^b, \alpha^{b'})$: depending on the scenario there is more over- or under-estimation that for $r_{random}(\alpha^b, \alpha^{b'})$. In this case the effect of using all polymorphisms or SNP is more marked. However, overall, we find that our estimators explain well the decay in $r(\alpha^b, \alpha^{b'})$. The imperfect disagreement (both in Figures 1 and 2) is probably due to several wrong assumptions; we comment some of them in the Discussion.

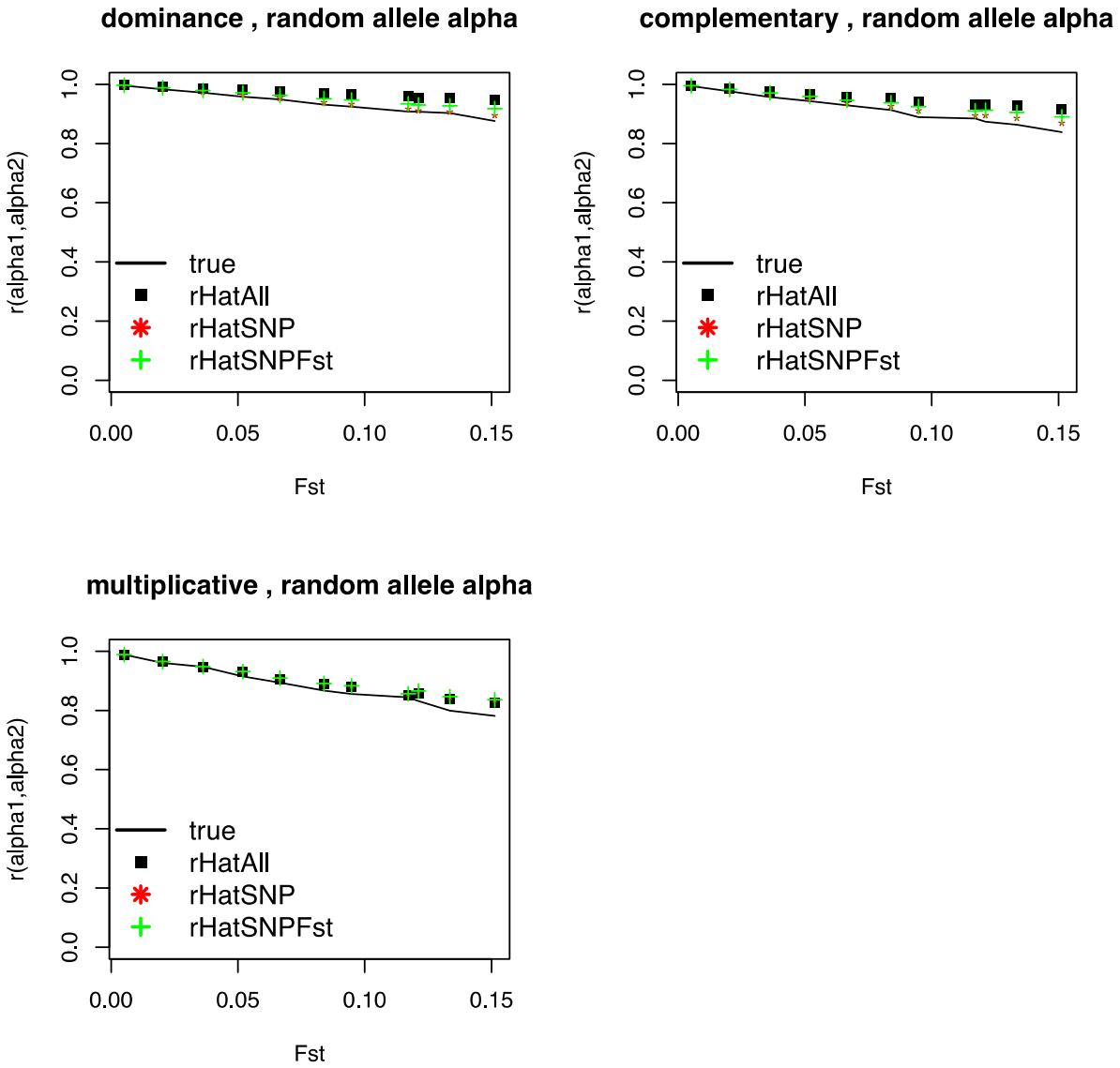


Figure 1. Simulated (straight black line) and estimated (points) correlation across QTL substitution effects of random alleles across two populations as a function of their F_{ST} differentiation coefficient. rHatAll: estimates with all polymorphism. rHatSNP: estimates using SNP-like loci. rHatSNPFst: estimates using SNP-like loci with a correction for heterozygosity. Results of 10 replicates per point with s.e.<0.01.

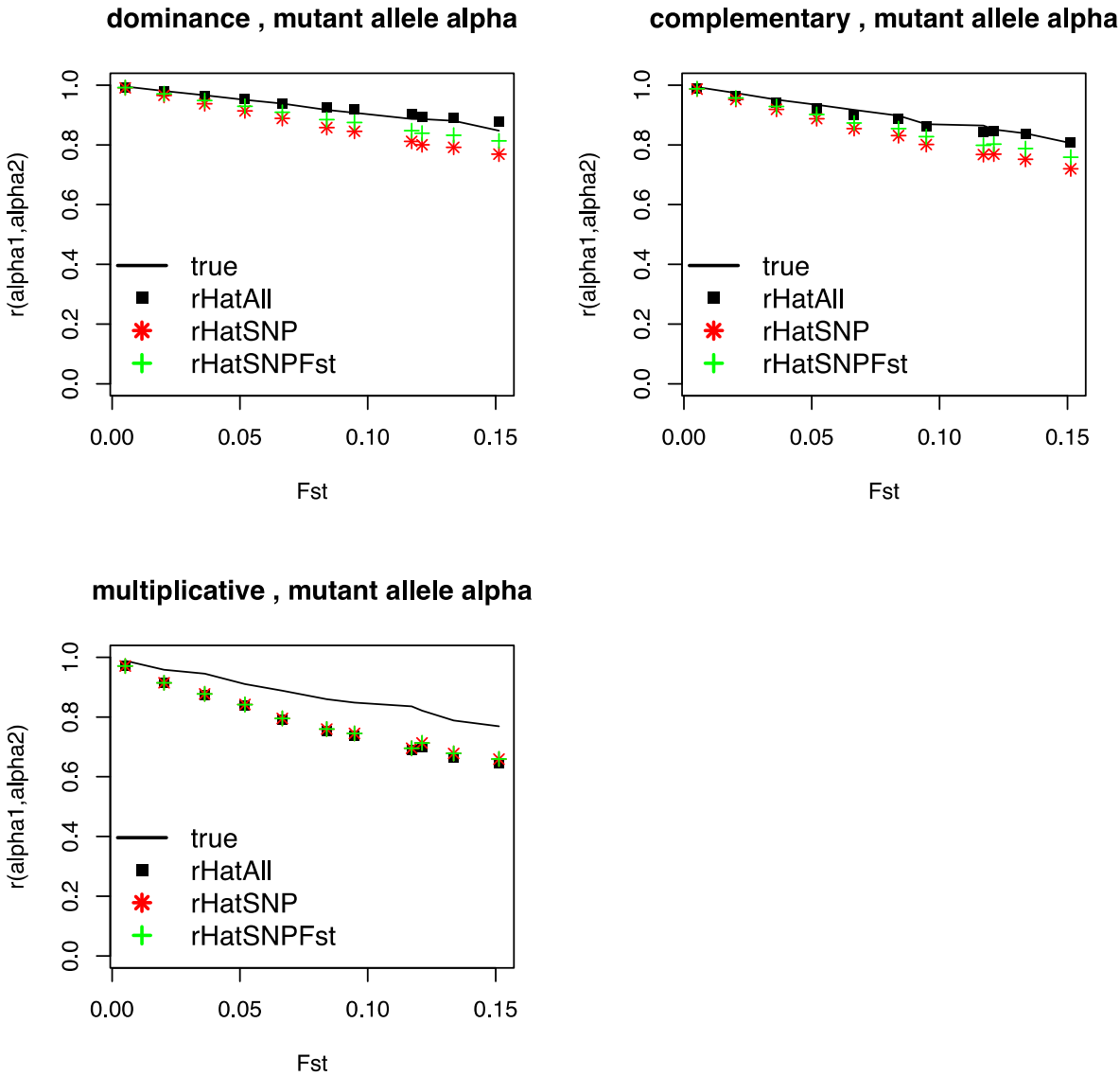


Figure 2. Simulated (straight black line) and estimated (points) correlation across QTL substitution effects of mutant alleles across two populations as a function of their F_{ST} differentiation coefficient. rHatAll: estimates with all polymorphism. rHatSNP: estimates using SNP-like loci. rHatSNPFst: estimates using SNP-like loci with a correction for heterozygosity. Results of 10 replicates per point with s.e.<0.01.

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 breeding values or disease risks across (distinct) populations, either in animal or human genetics. Only recently has non-additive biological gene action been identified as playing a role in determining the degree to which prediction across populations is successful (Dai *et al.* 2020; Duenk *et al.* 2020). 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. Having a theory allows to derive more general expressions and understanding factors influencing across-population differences in substitution effects.

In this work, we present a general, formal framework that does not depend on specific hypothesis regarding 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 also with values from literature. However, our simulations may not be particularly realistic, something that would require considerable thinking on how to simulate biologically meaningful epistasis mechanisms for a variety of traits. We see them as building blocks of non-additive architecture. At any rate, the three

scenarios generated sizeable non-additive variation which is a challenging case for our expressions.

Three main factors influence the correlation of substitution effects between populations $r(\alpha^b, \alpha^{b'})$: the genetic similarity of the two populations, the magnitudes of additive, dominance and additive by additive variances, and the distribution of allele frequencies at QTL. We consider that showing explicitly these three factors is an achievement, as their role is implicit, yet not explicitly shown, in previous works 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 summarized by Nei's minimum genetic distances $D_{b,b'}$ or F_{st} indexes. Under pure drift scenarios, these depend on divergence times and effective population sizes (Weir and Hill 2002; Bonhomme *et al.* 2010; Walsh and Lynch 2018).

The factors $\frac{\overline{H_b}}{\overline{H_b^2}}$ and $\frac{1}{\overline{H_b}}$ are weighting factors on dominance and additive by additive variances. If the allele frequencies are modelled using symmetric $Beta(a, a)$ distributions (see Appendix, section 1.10), these become $\frac{\overline{H_b}}{\overline{H_b^2}} = \frac{1}{2\frac{a+1}{a}\left(1+\frac{a+2}{2a+2}\left(\frac{a+3}{2a+3}-2\right)\right)}$ and $\frac{1}{\overline{H_b}} = 2 + \frac{1}{a}$. The first is bounded between 3 (for U-shaped distributions) and 2 (for peaked distributions), so that dominance variation does not play a big role in the difference between substitution effects across populations unless dominance variation is much larger that additive variation, something that seems unlikely based on theory and estimates. However, the second weight, due to epistasis, is not bounded and is large for small values of a (*e.g.* 27 for Beta(0.04,0.04),

592 and in this case functional epistasis plays a strong role in additive variation for U-shaped
593 distributions (Hill *et al.* 2008). The spectra of allele frequencies of causal mutations is
594 subject to large debate but there seems to be a consensus that low frequency mutations
595 make non-negligible contributions to genetic variance (Eyre-Walker 2010; Gibson 2012). It is
596 unclear if this leads to extreme values of $\frac{1}{H_b}$.
597
598 This result also reinforces the findings of Duenk *et al.* (2020) and Dai *et al.* (2020) that it is
599 mainly biological epistasis that generates changes in additive substitution effects across
600 populations. For instance, when one population drifts, the additive by additive statistical
601 variation enters into additive variation (Hill *et al.* 2006; Walsh and Lynch 2018).
602
603 Thus, $r(\alpha_i^b, \alpha_i^{b'})$ will be very low only if (1) populations are distinct, (2) there is *statistical*
604 additive by additive variation and (3) QTL allele frequency distributions are U- or L- shaped.
605
606 As for the magnitudes of non-additive variance, in most conceivable cases σ_{AA}^2 and σ_D^2 are of
607 magnitudes at most like additive variances, but more often less than half (Fuerst and
608 Sölkner 1994; Palucci *et al.* 2007; Hill *et al.* 2008; Mäki-Tanila and Hill 2014; Vitezica *et al.*
609 2018), so in practice the most limiting factor is distance across breeds. Indeed, there are very
610 few accurate estimates of non-additive variation although more are becoming available. We
611 argue that then, and based also in our results, it will be possible to make theory and data-
612 based choices on the possibilities of using predictions and GWAS results across populations.
613
614 Then, for reasonable assumptions about allele frequency distribution of QTLs, we have
615 shown that the correlation of substitution effects across populations is typically around 0.8

616 or higher, which is higher than scarce estimates of genetic correlations across populations
617 available in the literature, which range from 0.3 to 0.8 (Karoui *et al.* 2012; Legarra *et al.*
618 2014; Porto-Neto *et al.* 2015; Xiang *et al.* 2017) the difference being due, probably, to
619 genotype by environment interaction. Our results seem rather robust to different
620 distributions of allele frequencies. These values are high but not 1, which raises the question
621 of how to conceive optimal strategies for across populations predictions (*e.g.* more data
622 within breed or finer locations of causal variants across breed). This is of practical relevance,
623 *e.g.* for genomic predictions in livestock improvement, but also in human genetics *e.g.* for
624 the use of European-based Polygenic Risk Scores in individuals from other ancestries (Martin
625 *et al.* 2019).
626
627 Similar results apply to the same populations across generations, in which case the
628 correlation of substitution effects across generations goes from 0.99 (next 1-2 generations)
629 to 0.80 (10 generations). This illustrates that, even if genomic predictions would use QTLs or
630 markers in very tight LD with QTLs, there would still be, in the long run, a need for a
631 continuous system of data collecting and re-estimation of effects.
632
633 In the simulations we confirmed that our estimates are reasonably good although not
634 perfect. They are, depending on the scenario and target (random allele or mutant allele),
635 almost unbiased, slightly biased upwards, or biased downwards. There are several reasons
636 for the disagreement. The most obvious one is the inherent approximation of the Taylor
637 series expansion. Secondly, splitting variances such as $Var(\epsilon_i d_i^{*b})$ or $Var(\boldsymbol{\epsilon}'(\boldsymbol{\alpha}\boldsymbol{\alpha})_i^b)$ into
638 basic components implies either a strong assumption of multivariate normality and
639 independence or a less strong one of "expectation and variance-independence" (Bohrnstedt

and Goldberger 1969). For instance, it is assumed that the variance of d_i^{*b} is not related to the magnitude of the difference of allele frequencies ϵ_i, but this is not necessarily true. Third, it is further assumed, in the factorization of genetic variances, that QTL effects and allele frequencies are independent. We have also ignored that the change ϵ is proportional to heterozygosity and therefore small at extremes values of allele frequencies.

Another factor that we did not consider is the empirical evidence of an inverse relationship between heterozygosity and absolute effect at the locus (Park *et al.* 2011). It is unclear how this would affect our findings. A (rather extreme) functional gene action that generates larger α at extreme frequencies is overdominance, which is similar to our "dominance" scenario.

As for our estimators of the correlation across "negative" alleles $r_{mutant}(\alpha^b, \alpha^{b'})$, they are less robust that the estimators for a "random" allele $r_{random}(\alpha^b, \alpha^{b'})$. The reason for this is that obtaining $r_{mutant}(\alpha^b, \alpha^{b'})$ from $r_{random}(\alpha^b, \alpha^{b'})$ involves a further approximation, the normality of α.

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 SelGen metaprograms EpiSel and EpiFun. This project has received funding from the European Unions' Horizon 2020 Research & Innovation programme under grant agreement N°772787 -SMARTER. We are grateful to the Genotoul Bioinformatics Platform Toulouse Midi- Pyrenees (Bioinfo Genotoul) for providing computing and storage resources. We thank editor and reviewers for very detailed feedback, and MA Toro, A Caballero, S Boitard and M Bonhomme for advice.

REFERENCES

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

689 Bhatia G., N. Patterson, S. Sankararaman, and A. L. Price, 2013 Estimating and interpreting
690 FST: The impact of rare variants. Genome Res. 23: 1514–1521.
691 https://doi.org/10.1101/gr.154831.113

692 Bohrnstedt G. W., and A. S. Goldberger, 1969 On the Exact Covariance of Products of
693 Random Variables. J. Am. Stat. Assoc. 64: 1439–1442.
694 https://doi.org/10.1080/01621459.1969.10501069

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

698 Bonifati V., P. Rizzu, M. J. van Baren, O. Schaap, G. J. Breedveld, *et al.*, 2003 Mutations in the
699 DJ-1 Gene Associated with Autosomal Recessive Early-Onset Parkinsonism. Science
700 299: 256–259. https://doi.org/10.1126/science.1077209

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

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

706 Chen G. K., P. Marjoram, and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence
707 data. Genome Res. 19: 136–142. https://doi.org/10.1101/gr.083634.108

708 Cuyabano B. C. D., A. C. Sørensen, and P. Sørensen, 2018 Understanding the potential bias of
709 variance components estimators when using genomic models. Genet. Sel. Evol. 50:
710 41. https://doi.org/10.1186/s12711-018-0411-0

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

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

717 Eyre-Walker A., 2010 Genetic architecture of a complex trait and its implications for fitness
718 and genome-wide association studies. Proc. Natl. Acad. Sci. 107: 1752–1756.
719 https://doi.org/10.1073/pnas.0906182107

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

722 Fernández E., J. Sánchez, R. Martínez, A. Legarra, and M. Baselga, 2017 Role of inbreeding
723 depression, non-inbred dominance deviations and random year-season effect in
724 genetic trends for prolificacy in closed rabbit lines. J. Anim. Breed. Genet. 134: 441–
725 452.

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

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

Galarza C. E., L. A. Matos, D. K. Dey, and V. H. Lachos, 2020 On moments of folded and doubly truncated multivariate extended skew-normal distributions. ArXiv200913488 Math Stat.

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

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

Gibbs R. A., J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversole, *et al.*, 2009 Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Sci. N. Y. NY 324: 528–532.

Gibson G., 2012 Rare and common variants: twenty arguments. Nat. Rev. Genet. 13: 135–145. https://doi.org/10.1038/nrg3118

747 Grisart B., W. Coppieters, F. Farnir, L. Karim, C. Ford, *et al.*, 2002 Positional candidate cloning
748 of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1
749 gene with major effect on milk yield and composition. Genome Res 12: 222–231.

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

753 Haile-Mariam M., and J. E. Pryce, 2015 Variances and correlations of milk production,
754 fertility, longevity, and type traits over time in Australian Holstein cattle. J. Dairy Sci.
755 98: 7364–7379. https://doi.org/10.3168/jds.2015-9537

756 Hayes B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard, 2009 Accuracy
757 of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol
758 41: 51.

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

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

765 Hudson R. R., M. Slatkin, and W. P. Maddison, 1992 Estimation of levels of gene flow from
766 DNA sequence data. Genetics 132: 583–589.

Kan R., and C. Robotti, 2017 On Moments of Folded and Truncated Multivariate Normal Distributions. J. Comput. Graph. Stat. 26: 930–934. https://doi.org/10.1080/10618600.2017.1322092

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.

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

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

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

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

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

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

800 Marigorta U. M., and A. Navarro, 2013 High Trans-ethnic Replicability of GWAS Results
801 Implies Common Causal Variants. PLOS Genet. 9: e1003566.
802 https://doi.org/10.1371/journal.pgen.1003566

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

806 Meuwissen T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value
807 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

Nei M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press.

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

Park J.-H., M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung, *et al.*, 2011 Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc. Natl. Acad. Sci. 108: 18026–18031. https://doi.org/10.1073/pnas.1114759108

Pérez-Enciso M., 2014 Genomic relationships computed from either next-generation sequence or array SNP data. J. Anim. Breed. Genet. Z. Tierzuchtung Zuchtungsbiologie 131: 85–96. https://doi.org/10.1111/jbg.12074

Pérez-Enciso M., J. C. Rincón, and A. Legarra, 2015 Sequence-vs. chip-assisted genomic selection: accurate biological information is advised. Genet. Sel. Evol. 47: 43.

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

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

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

Shi H., S. Gazal, M. Kanai, E. M. Koch, A. P. Schoech, *et al.*, 2021 Population-specific causal disease effect sizes in functionally important regions impacted by selection. Nat. Commun. 12: 1098. https://doi.org/10.1038/s41467-021-21286-1

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. Legarra, M. A. Toro, and L. Varona, 2017 Orthogonal Estimates of Variances for Additive, Dominance and Epistatic Effects in Populations. Genetics genetics.116.199406. https://doi.org/10.1534/genetics.116.199406

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.

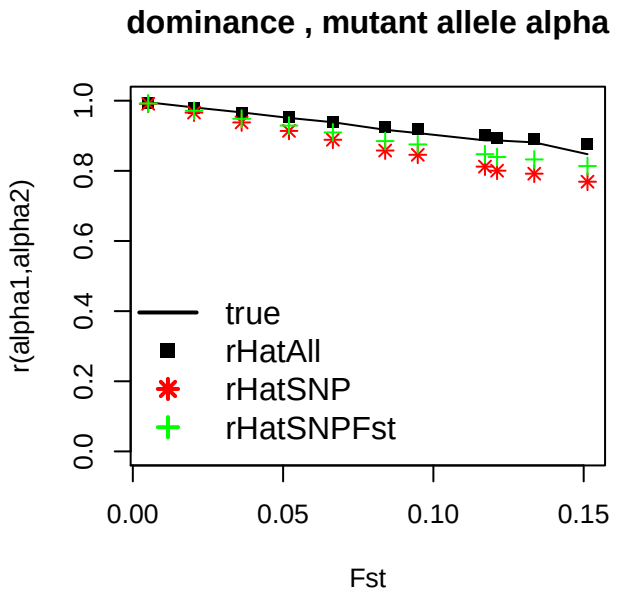
Walsh B., and M. Lynch, 2018 *Evolution and Selection of Quantitative Traits*. Oxford University Press.

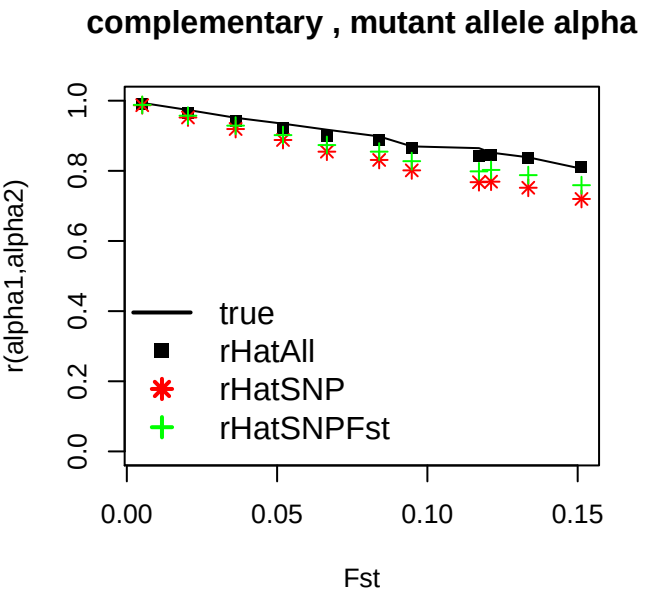
Weir B. S., and W. G. Hill, 2002 Estimating F-statistics. Annu. Rev. Genet. 36: 721–750. https://doi.org/10.1146/annurev.genet.36.050802.093940

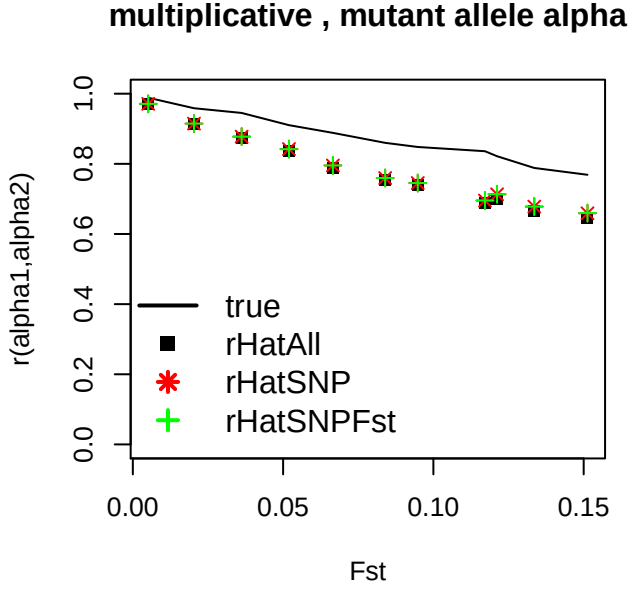
Welsh C. S., T. S. Stewart, C. Schwab, and H. D. Blackburn, 2010 Pedigree analysis of 5 swine breeds in the United States and the implications for genetic conservation. J. Anim. Sci. 88: 1610–1618. https://doi.org/10.2527/jas.2009-2537

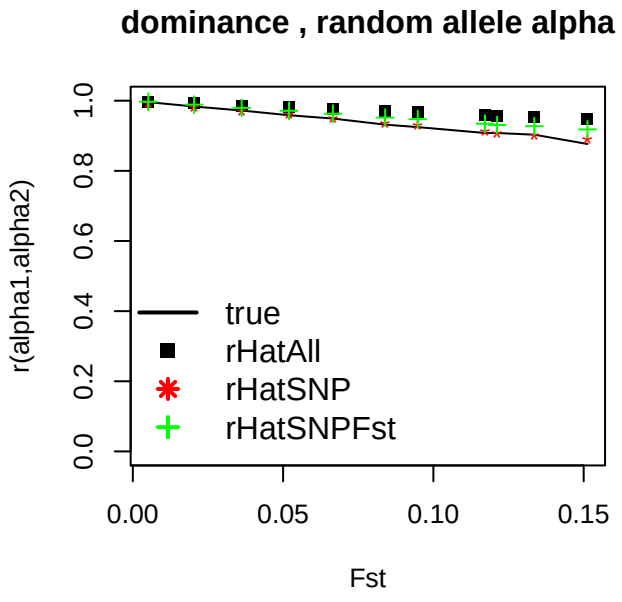
Wientjes Y. C. J., P. Bijma, J. Vandenplas, and M. P. L. Calus, 2017 Multi-population Genomic Relationships for Estimating Current Genetic Variances Within and Genetic Correlations Between Populations. Genetics 207: 503–515. https://doi.org/10.1534/genetics.117.300152

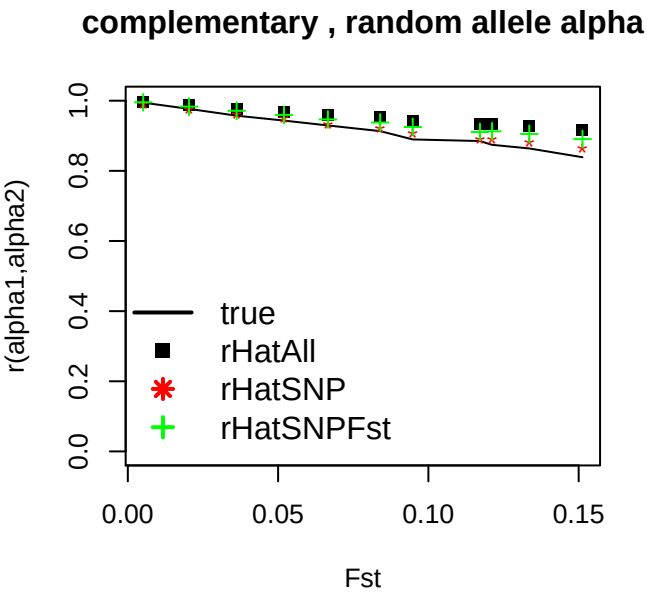
Xiang T., O. F. Christensen, and A. Legarra, 2017 Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. J. Anim. Sci. 95: 1472–1480.

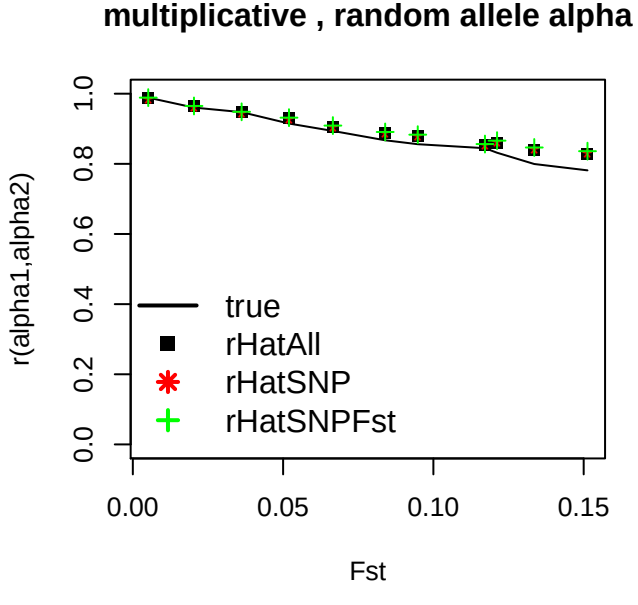

dominance , mutant allele alpha
r(alpha1,alpha2)
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
Fst
true
rHatAll
rHatSNP
rHatSNPFst


complementary , mutant allele alpha
r(alpha1,alpha2)
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
Fst
true
rHatAll
rHatSNP
rHatSNPFst


multiplicative , mutant allele alpha
r(alpha1,alpha2)
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
Fst
true
rHatAll
rHatSNP
rHatSNPFst


dominance , random allele alpha
r(alpha1,alpha2)
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
Fst
true
rHatAll
rHatSNP
rHatSNPFst


complementary , random allele alpha
r(alpha1,alpha2)
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
Fst
true
rHatAll
rHatSNP
rHatSNPFst


multiplicative , random allele alpha
r(alpha1,alpha2)
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
Fst
true
rHatAll
rHatSNP
rHatSNPFst