Appendix to "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.

Here we present detailed derivations of many aspects of this work.

1.1 Taylor series expansion of substitution effects

The function $\alpha_i^{b'} \approx f(\alpha_i^b, \boldsymbol{\epsilon})$, constructed as a Taylor series expansion around α_i^b evaluated at $\boldsymbol{p}^b$, is, e.g. (Walsh & Lynch, 2018, eq. A6.7b):

$$\alpha_i^{b'} \approx \alpha_i^b + \boldsymbol{\nabla}_i' \boldsymbol{\epsilon} + \frac{1}{2}\boldsymbol{\epsilon}' \boldsymbol{\mathcal{H}}_i \boldsymbol{\epsilon} \ldots$$

where $\boldsymbol{\nabla}$ and $\boldsymbol{\mathcal{H}}$ are a vector and a matrix containing, respectively, the gradient and Hessian of α_i with respect to $\boldsymbol{p}$ evaluated at $\boldsymbol{p}^b$ and $\boldsymbol{\epsilon} = \boldsymbol{p}^{b'} - \boldsymbol{p}^b$.

There will be different $\boldsymbol{\nabla}_i$ and $\boldsymbol{H}_i$ per each of the n loci. The elements of $\boldsymbol{\nabla}_i$ are:

$$\boldsymbol{\nabla}_i = \begin{pmatrix} \frac{\partial \alpha_i}{\partial p_1}\big|(p_i = p_i^b) \\ \ldots \\ \frac{\partial \alpha_i}{\partial p_i}\big|(p_i = p_i^b) \\ \ldots \\ \frac{\partial \alpha_i}{\partial p_n}\big|(p_i = p_i^b) \end{pmatrix} = 2\begin{pmatrix} (\alpha\alpha)_{i1}^b \\ \\ -d_i^{*b} \\ \\ (\alpha\alpha)_{in}^b \end{pmatrix}$$

For instance, $(\alpha\alpha)_{i1} = \frac{1}{2}\frac{\partial \alpha_i}{\partial p_1}$ and therefore $\frac{\partial \alpha_i}{\partial p_1}\big|(p_i = p_i^b) = 2(\alpha\alpha)_{i1}^b$. In other words, the vector of first derivatives of α_i can be written as function of statistical effects in a focal ("*b*") population with allele frequencies p_i^b. In the 1st order expansion, these effects are additive by additive $(\alpha\alpha)$ (of a locus with other loci) and dominant d^* (of a locus with itself). The *b* in the notation $(\alpha\alpha)_{i1}^b$ implies that the Taylor series is expanded around a population with $\boldsymbol{p} = \boldsymbol{p}^b$.

The elements of $\boldsymbol{\mathcal{H}}_i$ for a pair of other loci j and k are:

$$\boldsymbol{\mathcal{H}}_{i(jk)} = \left(\frac{\partial \alpha_i}{\partial p_j p_k}\big|(p_i = p_i^b)\right) = 4(\alpha\alpha\alpha)_{ijk}^b$$

For $i \neq j \neq k$. If $i = j = k$, then $\boldsymbol{\mathcal{H}}_{i(jk)}$ is null and if $j = k$, then $\boldsymbol{\mathcal{H}}_{i(jk)}$ is the additive by dominance epistatic interaction $(-\alpha d)_{ij}$.

We are going to ignore these higher-order interactions (2nd order including dominance, and 3-way and higher) for mathematical convenience, but also because such statistical interactions are expected to be small (Mäki-Tanila & Hill, 2014). Therefore, from the Taylor series expansion above, we will consider only the first two terms, i.e. $\alpha_i^{b'} \approx \alpha_i^b + \boldsymbol{\nabla}_i' \boldsymbol{\epsilon}$. Written as function of substitution effects this gives, for locus i,

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

Where $(\boldsymbol{\alpha\alpha})_i^b$ is a vector containing epistatic substitution effects of locus i with the rest of loci. By convention and to simplify notation we set $(\alpha\alpha)_{ii}^b = 0$ (this corresponds to the interaction of a locus with itself, i.e. the dominance effect d_i^{*b}). Note that α_i^b, d_i^{*b} and $(\boldsymbol{\alpha\alpha})_i^b$ do *not* depend on allele frequencies in b' (as they are defined for the focal population b with allele frequencies $\boldsymbol{p}^b$).

The expression above also shows, without invoking any explicit "functional" effect, that the lower level statistical effects in a reference framework (or set of allele frequencies) contain higher order statistical effects in another framework (Álvarez-Castro & Carlborg, 2007).

1.2 Expansion of $Var\left(\alpha_i^b + 2\epsilon_i\left(-d_i^{*b}\right) + 2\boldsymbol{\epsilon}'(\boldsymbol{\alpha\alpha})_i^b\right)$

First,

$$Var\left(\alpha_i^b + 2\epsilon_i\left(-d_i^{*b}\right) + 2\boldsymbol{\epsilon}'(\boldsymbol{\alpha\alpha})_i^b\right) = Var\left(\alpha_i^b\right) + 4Var\left(\epsilon_i d_i^{*b}\right) + 4Var\left(\boldsymbol{\epsilon}'(\boldsymbol{\alpha\alpha})_i^b\right)$$

as cross-products $Cov\left(\alpha^b, \epsilon_i\left(-d_i^{*b}\right)\right)$ and $Cov\left(\alpha^b, \boldsymbol{\epsilon}'(\boldsymbol{\alpha\alpha})_i^b\right)$ are 0 because statistical effects (α_i^b, d_i^{*b} and $(\boldsymbol{\alpha\alpha})_i^b$) are mutually orthogonal. For $Var\left(\epsilon_i d_i^{*b}\right)$, we use (under certain assumptions, roughly speaking of independence of ϵ_i and d_i^{*b}) result [9] in Bohrnstedt and Goldberger (1969) : $Var(xy) = E^2(x)Var(y) + E^2(y)Var(x) + Var(x)Var(y)$, leading to

$$\begin{aligned} Var\left(\epsilon_i d_i^{*b}\right) &= E^2(\epsilon_i)Var\left(d_i^{*b}\right) + E^2(d_i^{*b})Var(\epsilon_i) + Var(\epsilon_i)Var\left(d_i^{*b}\right) \\ &= Var(\epsilon_i)\left(Var\left(d_i^{*b}\right) + E^2\left(d_i^{*b}\right)\right) \end{aligned}$$

as $E(\epsilon_i) = 0$ (as changes are assumed to be random and not directional). In turn, $E^2\left(d_i^{*b}\right) = \mu_{d,b}^2$ from above which yields $Var(\epsilon_i)\left(\sigma_{d,b}^2 + \mu_{d,b}^2\right)$.
The expansion of $4Var\left(\boldsymbol{\epsilon}'(\boldsymbol{\alpha\alpha})_i^b\right)$ involves the variance of a quadratic form as shown below (section 1.3).

1.3 Variance of a quadratic form

The expression for the variance $Var(\boldsymbol{x}_1'\boldsymbol{x}_2)$ of a bilinear form $(\boldsymbol{x}_1'\boldsymbol{x}_2)$ for two vectors, with joint covariance matrix $Var\begin{pmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{pmatrix} = \boldsymbol{C} = \begin{pmatrix}\boldsymbol{C}_{11} & \boldsymbol{C}_{12}\\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22}\end{pmatrix}$ and $E\begin{pmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}$, is equal to $Var(\boldsymbol{x}_1'\boldsymbol{x}_2) = tr(\boldsymbol{C}_{21})^2 + tr(\boldsymbol{C}_{22}\boldsymbol{C}_{11})$ (Searle, 1971, Chapter 2). For our specific case, we are interested in $Var\left(\boldsymbol{\epsilon}'(\boldsymbol{\alpha\alpha})_i^b\right)$ with $E\begin{pmatrix}\boldsymbol{\epsilon}'\\ (\boldsymbol{\alpha\alpha})_i^b\end{pmatrix} = \begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}$ and $Var\begin{pmatrix}\boldsymbol{\epsilon}'\\ (\boldsymbol{\alpha\alpha})_i^b\end{pmatrix} = \begin{pmatrix}Var(\boldsymbol{\epsilon}') & \boldsymbol{0}\\ \boldsymbol{0} & Var\left((\boldsymbol{\alpha\alpha})_i^b\right)\end{pmatrix}$ where we assume $Cov\left(\boldsymbol{\epsilon}', (\boldsymbol{\alpha\alpha})_i^b\right) = \boldsymbol{0}$. Thus, the above expression reduces to

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

1.4 Expansion of substitution effects can also be done around a third (focal, f) population

The expansion of substitution effects can also be done around a third (focal, f) population as follows:

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

Where $\epsilon_i^{(b')} = p_i^{b'} - p_i^f$ and $\epsilon_i^{(b)} = p_i^b - p_i^f$. This leads to

$$Var\left(\alpha_i^{b'}\right) \approx Var\left(\alpha_i^f\right) + 4Var\left(\epsilon_i^{(b')}d_i^{*f}\right) + 4Var\left(\boldsymbol{\epsilon}^{(b')'}(\boldsymbol{\alpha\alpha})_i^f\right)$$
$$Var\left(\alpha_i^b\right) \approx Var\left(\alpha_i^f\right) + 4Var\left(\epsilon_i^{(b)}d_i^{*f}\right) + 4Var\left(\boldsymbol{\epsilon}^{(b)'}(\boldsymbol{\alpha\alpha})_i^f\right)$$

And

$$Cov\left(\alpha_i^{b'}, \alpha_i^b\right) \approx Var\left(\alpha_i^f\right) + Cov\left(\epsilon_i^{(b')}, \epsilon_i^{(b)}\right)\left(Var\left(d_i^{*b}\right) + E^2\left(d_i^{*b}\right)\right)$$
$$+ Cov\left(\boldsymbol{\epsilon}^{(b')'}(\boldsymbol{\alpha\alpha})_i^f, \boldsymbol{\epsilon}^{(b)'}(\boldsymbol{\alpha\alpha})_i^f\right)$$

If the two populations drifted from a common ancestral one, the covariance of deviations from the focal populations, $Cov\left(\epsilon_i^{(b')}, \epsilon_i^{(b)}\right)$, could in principle be put as a function of genealogical kinships across populations:

$$Cov\left(\epsilon_i^{(b)}, \epsilon_i^{(b')}\right) = Cov\left(p^{(b)} - p^f, p^{(b')} - p^f\right)$$
$$= Cov\left(p^{(b)}, p^{(b')}\right) - Cov\left(p^{(b)}, p^f\right) - Cov\left(p^f, p^{(b')}\right) + Cov\left(p^f, p^f\right)$$
$$= \left(\theta_{b,b'} - \theta_{b,f} - \theta_{b',f} + \theta_{f,f}\right)p(1-p)$$

where θ are across- or within-population kinships referring to the ancestral population and $p(1-p)$ is half the heterozygosity at the ancestral population. If the focal population f is the ancestor of both populations and there is only drift (but no migration), then $\theta_{b,b'} = \theta_{b,f} = \theta_{b',f} = \theta_{f,f}$, the term $\left(\theta_{b,b'} - \theta_{b,f} - \theta_{b',f} + \theta_{f,f}\right)$ cancels out and $Cov\left(\epsilon_i^{(b)}, \epsilon_i^{(b')}\right) = 0$. The reason is that, after the split of the populations, both deviations are independent from each other (Weir and Hill, 2002; Bonhomme et al., 2010) and thus $Cov\left(\epsilon_i^{(b)}, \epsilon_i^{(b')}\right) = 0$. This is not true if the focal population is *not* an ancestor of the other two.

However, these genealogical kinships, and allelic frequencies at the ancestral population, are usually unknown. For this reason, and because then the focal population needs not to be an ancestor, we formulate $Cov\left(\epsilon_i^{(b)}, \epsilon_i^{(b')}\right)$ in terms of Nei's genetic distance in the following.

Consider $Var\left(p^{b'} - p^b\right) = E\left(\left(p^{b'} - p^b\right)^2\right) - E\left(p^{b'} - p^b\right)^2 = E\left(\left(p^{b'} - p^b\right)^2\right)$ because $E\left(p^{b'} - p^b\right) = 0$ when averaged across loci. Therefore, $Var\left(p^{b'} - p^b\right) = E\left(\left(p^{b'} - p^b\right)^2\right)$

which corresponds to Nei's $D_{b,b'}$ "minimum genetic distance" (Nei 1987; Caballero and Toro 2002). We massage this expression differently:

$$Var(p^{b'} - p^b) = Cov(p^{b'} - p^b, p^{b'} - p^b) = Var(p^{b'}) + Var(p^b) - 2Cov(p^b, p^{b'})$$

from which

$$Cov(p^b, p^{b'}) = \frac{Var(p^{b'})}{2} + \frac{Var(p^b)}{2} - \frac{D_{b,b'}}{2}$$

Now

$$Cov\left(\epsilon_i^{(b)}, \epsilon_i^{(b')}\right) = Cov(p^b - p^f, p^{b'} - p^f)$$
$$= Cov(p^b, p^{b'}) - Cov(p^b, p^f) - Cov(p^{b'}, p^f) + Var(p^f)$$

Substituting each of those covariances we obtain after cancellations

$$Cov\left(\epsilon_i^{(b)}, \epsilon_i^{(b')}\right)$$
$$= \frac{Var(p^{b'})}{2} + \frac{Var(p^b)}{2} - \frac{D_{b,b'}}{2} - \frac{Var(p^b)}{2} - \frac{Var(p^f)}{2} + \frac{D_{b,f}}{2} - \frac{Var(p^{b'})}{2}$$
$$- \frac{Var(p^f)}{2} + \frac{D_{b',f}}{2} + Var(p^f) = \frac{D_{b,f}}{2} + \frac{D_{b',f}}{2} - \frac{D_{b,b'}}{2}$$

Which is a (strictly positive) measure of similarity of b and b' after discounting similarity of both to f. Note that if $b = f$ (the focal population is b), $D_{b,f} = 0$ and $D_{b,b'} = D_{b',f}$ and thus $Cov\left(\epsilon_i^{(b)}, \epsilon_i^{(b')}\right) = 0$; and similarly, for $b' = f$.

Assuming uncorrelated changes across loci, $Cov\left(\boldsymbol{\epsilon}^{(b)}, \boldsymbol{\epsilon}^{(b')'}\right) = \boldsymbol{I}\left(-\frac{D_{b,b'}}{2} + \frac{D_{b,f}}{2} + \frac{D_{b',f}}{2}\right)$

From here and with a development similar to the main text (and 1.5, 1.6 in this Appendix) we obtain:

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

$$Var(\alpha_i^{b}) \approx \frac{1}{n}\left(\frac{\sigma_A^2}{\overline{H}_f} + 4D_{b,f}\frac{\sigma_D^2}{\overline{H}_f^2} + 8D_{b,f}\frac{\sigma_{AA}^2}{\overline{H}_f\,\overline{H}_f}\right)$$

Where σ_A^2, σ_D^2 and σ_{AA}^2 refer now to population f. Leading to

$$r(\alpha_i^b, \alpha_i^{b'}) \approx \frac{\left(\frac{\sigma_A^2}{\overline{H}_f} + 4\left(\frac{D_{b,f}}{2} + \frac{D_{b',f}}{2} - \frac{D_{b,b'}}{2}\right)\frac{\sigma_D^2}{\overline{H}_f^2} + 8\left(\frac{D_{b,f}}{2} + \frac{D_{b',f}}{2} - \frac{D_{b,b'}}{2}\right)\frac{\sigma_{AA}^2}{\overline{H}_f\overline{H}_f}\right)}{\sqrt{\left(\frac{\sigma_A^2}{\overline{H}_f} + 4D_{b,f}\frac{\sigma_D^2}{\overline{H}_f^2} + 8D_{b,f}\frac{\sigma_{AA}^2}{\overline{H}_f\,\overline{H}_f}\right)\left(\frac{\sigma_A^2}{\overline{H}_f} + 4D_{b',f}\frac{\sigma_D^2}{\overline{H}_f^2} + 8D_{b',f}\frac{\sigma_{AA}^2}{\overline{H}_f\,\overline{H}_f}\right)}}$$

Note that if $f = b$, we obtain expression [5] in the main text as $\left(\frac{D_{b,f}}{2} + \frac{D_{b',f}}{2} - \frac{D_{b,b'}}{2}\right) = 0$ and $D_{f,b} = 0$.

Also note that if f is an "average" population of b and b' (for instance an F2 cross), such that $\frac{D_{b,b'}}{2} = D_{b,f} = D_{b',f}$, then we obtain

$$r\left(\alpha_i^b, \alpha_i^{b'}\right) \approx \frac{\sqrt{\frac{\sigma_A^2}{\overline{H_f}}}}{\sqrt{\left(\frac{\sigma_A^2}{\overline{H_f}} + 4\frac{D_{b',b}}{2}\frac{\sigma_D^2}{\overline{H_f^2}} + 8\frac{D_{b',b}}{2}\frac{\sigma_{AA}^2}{\overline{H_f}\,\overline{H_f}}\right)}}$$

Similar in form to [5], , but not identical, because variances and heterozygosities refer to the F2 population.

1.5 Variance due to dominance effects and inbreeding depression

In a HWE population with dominance deviations, the genetic variance due to dominance deviations is

$$\sigma_D^2 = 4\sum p_i^2 q_i^2 (d_i^*)^2$$

To account for directional dominance, consider $E(d_i^*) = \frac{1}{n}\sum(d_i^*) = \mu_d$, $Var(d_i^*) = \sigma_d^2$. Then we can define "centered" $d_i^{*(c)} = d_i^* - \mu_d$, with $Var\left(d_i^{*(c)}\right) = \sigma_d^2$. If there are n loci, then the effect of inbreeding per unit of homozygosity is $b = -\mu_d n$. Now we can decompose $\sigma_D^2 = 4\sum p_i^2 q_i^2 (d_i^*)^2 = 4\sum p_i^2 q_i^2 \left(d_i^{*(c)}\right)^2 - 4\sum p_i^2 q_i^2 (2d_i^{*c}\mu_D) + \mu_d^2 4\sum p_i^2 q_i^2 = 4\sum p_i^2 q_i^2 \left(d_i^{*(c)}\right)^2 + \mu_d^2 4\sum p_i^2 q_i^2$ because $4\sum p_i^2 q_i^2 (2d_i^{*c}\mu_D) = 0$ across loci. Assuming independence of $p_i^2 q_i^2$ and $d_i^{*(c)}$, then $\sigma_D^2 = 4\sum p_i^2 q_i^2 (\sigma_d^2 + \mu_d^2)$ or in other terms, $\sigma_D^2 = n\overline{H^2}(\sigma_d^2 + \mu_d^2)$ where $\overline{H^2}$ is average squared heterozygosity.

1.6 Variance of statistical effects

We use the fact that the population variances are function of the variances of the different effects (Maki-Tanila and Hill, 2014) and moments of heterozygosities. First, we define functions of heterozygosities for all n loci at the focal population. The average heterozygosity across loci is:

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

The average squared heterozygosity is

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

and the average cross-product of heterozygosities across all pairs of distinct loci is

$$\overline{HH_b} = \frac{1}{0.5n(n-1)}\sum_i\sum_{j>i} 2p_i^b\left(1 - p_i^b\right)2p_j^b\left(1 - p_j^b\right)$$

In fact $n^2(\overline{H_b})^2 = 2n(n-1)\overline{HH_b} + n\overline{H_b^2}$ and from here we make the approximation $\overline{HH_b} \approx \frac{1}{2}\overline{H_b H_b}$.

If we knew these functions of heterozygosities, we could obtain, from estimates of additive, dominant and additive by additive variances, the variance of statistical additive, dominant and additive by additive effects:

$$Var\left(\alpha_i^b\right) = \sigma^2_{\alpha,b} = \frac{\sigma_A^2}{n\overline{H}_b}$$

$$Var\left(d_i^{*b}\right) + E^2\left(d_i^{*b}\right) = \sigma_d^2 + \mu_d^2 = \frac{\sigma_D^2}{n\overline{H_b^2}}$$

$$Var\left((\alpha\alpha)^b_{i,j,j>i}\right) = \sigma^2_{(\alpha\alpha,b)} = \frac{\sigma^2_{AA}}{n(n-1)\overline{HH}_b} \approx 2\frac{\sigma^2_{AA}}{n^2\overline{H}_b\overline{H}_b}$$

$$Var\left((\boldsymbol{\alpha\alpha})^0_i\right) = \boldsymbol{I}\otimes\sigma^2_{(\alpha\alpha,b)} \approx 2\boldsymbol{I}\frac{\sigma^2_{AA}}{n^2\overline{H}_b\,\overline{H}_b}$$

All variances and effects refer to the focal population with allele frequencies p^b and effects α^b. Note that we assume HWE and LE.

1.7 Nei's minimal genetic distance $D_{b,b'}$ as function of F_{ST} and $\overline{H}_b$

The objective is to put $D_{b,b'}$ as a function of assumed known parameters F_{ST} and $\overline{H}_b$ (heterozygosity of population b). In fact, $F_{ST} = 1 - \frac{0.5(\overline{H}_b + \overline{H}_{b'})}{\overline{H}_a} = \frac{D_{b,b'}}{\overline{H}_a}$ where $\overline{H}_a = \frac{1}{n}\sum_i\left(p_i^{b'}\left(1-p_i^b\right) + p_i^b\left(1-p_i^{b'}\right)\right)$. Assuming $\overline{H}_b \approx \overline{H}_{b'}$ and after some manipulation, this yields $D_{b,b'} = \frac{F_{ST}}{1-F_{ST}}\overline{H}_b$. Note that although Nei's F_{ST} and Hudson et al. (1992) definitions of F_{ST} differ (Bhatia et al., 2013), their numerator is identical, and identical to $D_{b,b'}$ by definition.

1.8 Correlation of absolute values of bivariate normal

Consider x and y that are multivariate normal with 0 expectation and $Var\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{pmatrix}$

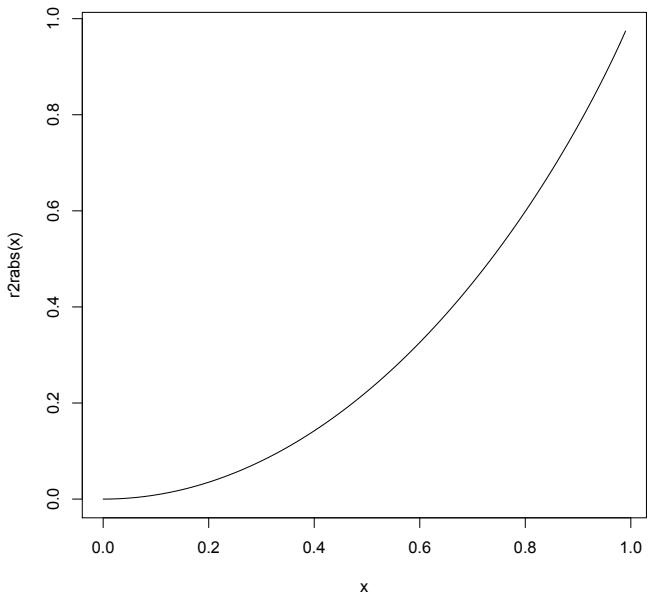
Now we define the transformed $|x|$ and $|y|$. Using Kan and Robotti (2017) one gets expressions for $E(|x|)$, $Var(|x|)$, and also for $|y|$, and $Cov(|x|,|y|)$. From here the values of the correlation can be obtained through a rather long expression.

This has been conveniently programmed in R package MomTrunc (Galarza et al., 2020). The r code to obtain the correlation of absolute values from the regular correlation is function `r2rabs` as follows:

```
require("MomTrunc")
```

```
foo=function(x)
 cov2cor(meanvarFMD(c(0,0),Sigma=matrix(c(1,x,x,1),2),dist='nor
 mal')$varcov)[1,2]
r2rabs = function(r){
 out=c()
 for (i in r){
     out=c(out,foo(i))
 }
 out
}
```

Which translates into the following quadratic:



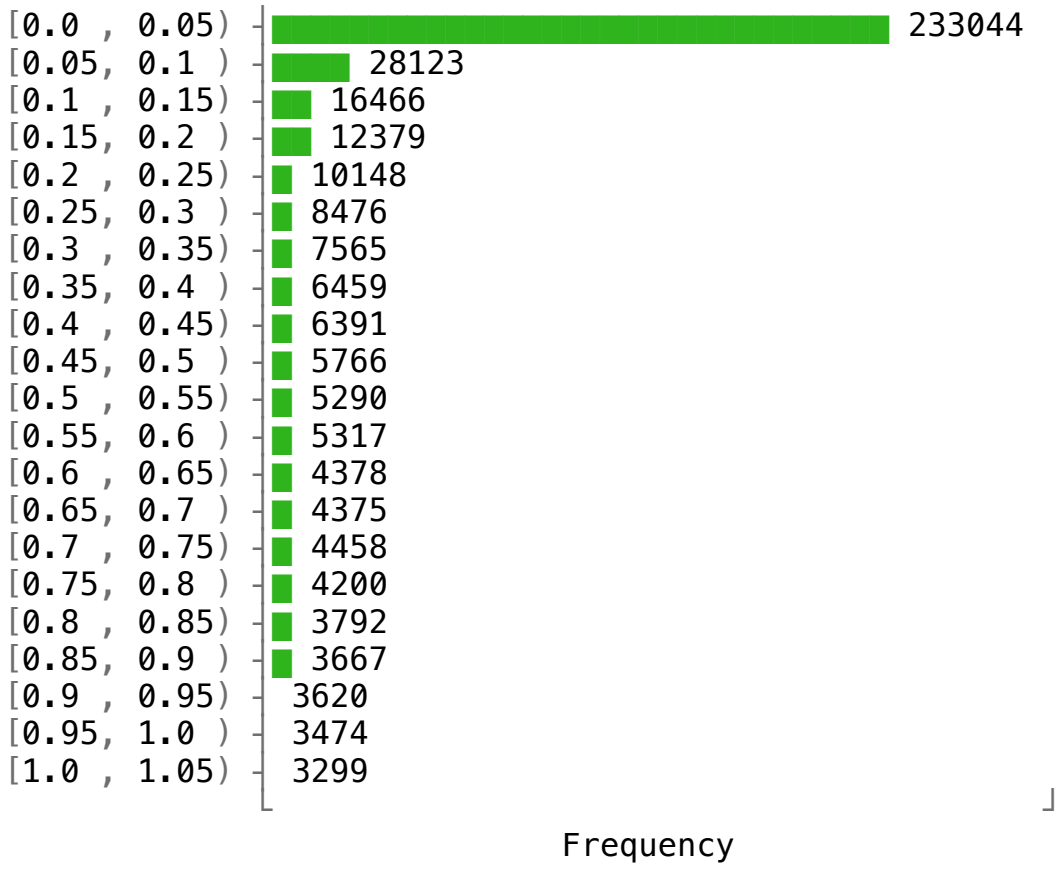
1.9 Simulation

The basic macs command is:

```
macs 400 3e5 -i 100 -t 0.0012 -r 0.001 -I 2 200 200 -ej $t10ver4Ne
2 1 -eN 1.67 10
```

Which translates as: generate 400 sequences of 3e5 bases, 100 times (i.e. 100 DNA stretches). -t specifies the nucleotide diversity per base pair, -r the scaled recombination rate (in 4Ne units), = $\theta = 4Ne\,r$ where $Ne = 300$, the population split into two populations of 200 gametes each (`-I 2 200 200`) at time `$t10ver4Ne` i.e. $t/4Ne$, -eN specifies that the population had a 10-fold population bottleneck 1.67 generations ago in $4Ne$ units.

This results in the following histogram of allele frequencies for population 1:



1.10 Derivation of substitution effects in dominance and epistatic systems

Generally, we use gene content (z-score) coded as {-1,0,1} for the three different genotypes.

Complete dominance. In this case we have:

	cc	Cc	CC
z	-1	0	1
	a	a	0

Here the mean is $a(1-p^2)$ and $\alpha = \frac{1}{2}\frac{\partial E(y)}{\partial p_i} = -ap$ where $p = frequency(C)$. The dominance deviation is $d_i^* = -\frac{1}{2}\frac{\partial \alpha_i}{\partial p_i} = -a/2$.

Complementary Epistasis. We can write the 2-locus complementary epistasis model as follows:

		cc	Cc	CC
	z	-1	0	1
bb	-1	a	a	0
bB	0	a	a	0
BB	1	0	0	0

The n-locus case generalizes to:

- a if genotypes at all n loci are NOT "upper case upper case"
- 0 otherwise

in other words

$$y = [z_1 \neq 1][z_2 \neq 1] \dots [z_n \neq 1]a$$

where z is the genotype, the operator "[ ]" is Iverson brackets notation (1 if true and 0 otherwise). Then, assuming independence across loci (i.e. LE), we generalize to n loci:

$$\begin{aligned} E(y) &= E([z_1 \neq 1][z_2 \neq 1][z_3 \neq 1]a) \\ &= aE([z_1 \neq 1])E([z_2 \neq 1])E([z_3 \neq 1]) \dots E([z_n \neq 1]) \end{aligned}$$

Now we obtain $E([z_1 \neq 1])$. The expression $[z_1 \neq 1]$ can be written as $[z_1 \neq 1] = 1 - \frac{(z_1 + z_1^2)}{2}$ (this is not really needed). Alternatively, the expectation $E([z_1 \neq 1]) = q_1^2 + 2p_1q_1 = 1 - p_1^2$ is obtained looking at the following table:

	z-score	$[z_1 \neq 1]$	frequency	sum
bb	-1	1	q^2	q^2
bB	0	1	$2pq$	$2pq$
BB	1	0	p^2	0

Thus, $E(y) = a(1 - p_1^2)(1 - p_2^2)(1 - p_3^2) \dots (1 - p_n^2)$

Then we can get the different α as *e.g.* for the first locus we have

$$\alpha_1 = \frac{1}{2}\frac{\partial E(y)}{\partial p_1} = -ap_1\Pi_{j>1}\left(1 - p_j^2\right)$$

If we call $K = \Pi\left(1 - p_j^2\right)$ then

$$\alpha_1 = \frac{1}{2}\frac{\partial E(y)}{\partial p_1} = -a\frac{p_1}{(1 - p_1^2)}K$$

and generally

$$\alpha_i = \frac{1}{2}\frac{\partial E(y)}{\partial p_i} = -ap_i\Pi_{j\neq i}\left(1 - p_j^2\right) = -a\frac{p_i}{(1 - p_i^2)}K$$

The dominance deviation is

$$d_i^* = -\frac{1}{2}\frac{\partial \alpha_i}{\partial p_i} = \frac{a}{2}\Pi_{j\neq i}\left(1 - p_j^2\right) = \frac{a}{2}\frac{1}{(1 - p_i^2)}K$$

The i – by – j epistatic effect $(\alpha\alpha)_{ij}$ is

$$(\alpha\alpha)_{ij} = \frac{1}{2}\frac{\partial \alpha_i}{\partial p_j} = \frac{1}{2}\frac{\partial}{\partial p_j}\left(-ap_i\left(1 - p_j^2\right)\Pi_{k\neq i,j}(1 - p_k^2)\right) =$$

$$ap_ip_j\Pi_{k\neq i,j}(1-p_k^2) = a\frac{p_ip_j}{(1-p_i^2)(1-p_j^2)}K$$

Note that the expression is oriented: p_i is the frequency of the "recessive" allele (say B) at locus i.

Multiplicative. Consider, for instance, the following values of the genotypic value for two loci with their respective frequencies:

		cc	Cc	CC
	z	-1	0	1
bb	-1	a	0	$-a$
bB	0	0	0	0
BB	1	$-a$	0	a

The genotypic value can be expressed as the product of z values at each loci, i.e. $G = z_1z_2$ for two loci. The generalization to several loci is immediate as $\prod_{i=1,n} z_i$. The average genotypic value considering *n* loci is, assuming Cov(Z_i, Z_j) = 0 (LE):

$$\mu = E(Z_1Z_2Z_3 \dots Z_n) = E(Z_1)\,E(Z_2)\,E(Z_3) \dots E(Z_n) \quad [1]$$

In general, $E(Z_i) = p_i^2 - q_i^2 = p_i - q_i = 2p_i - 1$ and consequently,

$$\mu = \prod_{i=1}^{n}(2p_i - 1)$$

Following Kojima's definition, the additive substitution effect at locus *i* is the first derivative of μ :

$$\alpha_i = \frac{1}{2}\frac{\partial\mu}{\partial p_i} = \frac{1}{2}\frac{\partial(2p_i-1)}{\partial p_i}\,\Pi_{i\neq j}(2p_j-1) = \prod_{j\neq i}(2p_j-1)$$

Or, calling $K = \Pi(2p_j - 1)$

$$\alpha_i = \frac{1}{2p_i - 1}K$$

The dominance deviation is 0 as expected:

$$d_i^* = -\frac{1}{2}\frac{\partial\alpha_i}{\partial p_i} = -\frac{1}{2}\frac{\partial}{\partial p_i}\prod_{j\neq i}(2p_j-1) = 0$$

The additive by additive effect is:

$$(\alpha\alpha)_{ij} = \frac{1}{2}\frac{\partial\alpha_i}{\partial p_j} = \prod_{k\neq i,j}(2p_k-1) = \frac{1}{(2p_i-1)(2p_j-1)}K$$

1.11 Derivation of the moments of functions of allele frequencies assuming beta distributions

We need the moments

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

and

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

from the parameters of the $Beta(\alpha,\beta)$ distribution of the allele frequencies p. The moments can be obtained using the Moment generating function of the Beta distribution such that $(p^k) = \prod_{r=1}^{k} \frac{\alpha+r-1}{\alpha+\beta+r-1}$. For instance

$$\overline{H} = 2E\big(p(1-p)\big) = 2E(p) - 2E(p^2) = 2\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}$$

Which for $\alpha = \beta = a$ yields

$$\overline{H} = \frac{2a^2}{2a(2a+1)} = \frac{a}{(2a+1)}$$

then

$$\overline{H^2} = E\big(4p(1-p)p(1-p)\big) = 4E(p^2 - 2p^3 + p^4)$$
$$= 4\frac{\alpha}{\alpha+\beta}\frac{\alpha+1}{\alpha+\beta+1}\left(1 + \frac{\alpha+2}{\alpha+\beta+2}\left(\frac{\alpha+3}{\alpha+\beta+3} - 2\right)\right)$$

Which for $\alpha = \beta = a$ yields

$$\overline{H^2} = 2\frac{a+1}{2a+1}\left(1 + \frac{a+2}{2a+2}\left(\frac{a+3}{2a+3} - 2\right)\right)$$

Then, is $\overline{HH} = E_{i>j}\left(2p_i(1-p_i)2p_j(1-p_j)\right)$, that we approximated as $\overline{HH} \approx \frac{1}{2}(\overline{H})^2 = \frac{1}{2}\left(\frac{a}{(2a+1)}\right)^2$

Finally,

$$\frac{\overline{H^2}}{\overline{H}} = 2\frac{a+1}{a}\left(1 + \frac{a+2}{2a+2}\left(\frac{a+3}{2a+3} - 2\right)\right)$$