# Genetic variance components and heritability of multi-allelic heterozygosity under inbreeding

Pirmin Nietlisbach*, Lukas F. Keller, Erik Postma

Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland;
* correspondence: pirmin.nietlisbach@ieu.uzh.ch

## *Supplementary Information*

**Supplementary information 1:** genetic variance components and heritability of heterozygosity as a function of variance in allele frequencies and mean heterozygosity

**Supplementary information 2:** Correction of formula for additive effect and additive genetic variance under inbreeding

**SUPPLEMENTARY INFORMATION 1: GENETIC VARIANCE COMPONENTS AND HERITABILITY OF HETEROZYGOSITY AS A FUNCTION OF VARIANCE IN ALLELE FREQUENCIES AND MEAN HETEROZYGOSITY**

*Variance in allele frequencies*

In the main text we have shown how $\sigma_G^2$, $\sigma_{AR}^2$, $\sigma_{DR}^2$, and $h^2$ for heterozygosity are functions of allele frequencies alone. Here we explore how allele frequencies, and in particular their variance, shape genetic variance components and the heritability of heterozygosity. Using $\mu = \sum_{i=1}^{n} p_i/n = 1/n$ for the mean allele frequency and assuming we know the true allele frequencies $p_i$ (i.e. we are not dealing with sampling variance), the variance in allele frequencies ($Var(p)$) within a population (Crawley 2007, page 52) is calculated as

$$\text{Var}(p) = \frac{\sum_{i=1}^{n}(p_i - \mu)^2}{n}$$

$$= \frac{\sum_{i=1}^{n}\left(p_i - 1/n\right)^2}{n}$$

$$= \frac{\sum_{i=1}^{n}(p_i^2 - 2\,p_i\,n^{-1} + n^{-2})}{n}.$$

Using $\sum_{i=1}^{n} p_i = 1$, $\sum_{i=1}^{n} n^{-2} = nn^{-2} = n^{-1}$, and some rearrangements, we get

$$\text{Var}(p) = \frac{\sum_{i=1}^{n}(p_i^2) - 2\,n^{-1} + n^{-1}}{n}$$

$$= \frac{\sum_{i=1}^{n} p_i^2}{n} - \frac{1}{n^2}.$$

Some further rearrangements show that

$$\sum_{i=1}^{n} p_i^2 = n \operatorname{Var}(p) + \frac{1}{n}. \tag{S1.1}$$

Substituting equation S1.1 into equation 5b yields

$$\sigma_G^2 = \left[n \operatorname{Var}(p) + \frac{1}{n}\right] - \left[n \operatorname{Var}(p) + \frac{1}{n}\right]^2. \tag{S1.2}$$

In equation 5a we have already shown that genetic variance in heterozygosity is a function of the allele frequencies only. If we prefer to use variance in allele frequencies rather than actual allele frequencies, equation S1.2 clarifies that genetic variance in heterozygosity is a quadratic function of only the variance in allele frequencies and the number of alleles. In other words, knowledge of the variance in allele frequencies in a population and the number of alleles is sufficient to predict variance in heterozygosity.

Following the same approach for additive genetic variance and substituting equation S1.1 into equation 2 yields

$$\sigma_{AR}^2 = 2 \left( \sum_{i=1}^{n} p_i \left( \left[n \operatorname{Var}(p) + \frac{1}{n}\right] - p_i \right)^2 \right). \tag{S1.3}$$

Equation S1.3 shows that additive genetic variance is again a quadratic function of the variance in allele frequencies and the number of alleles, but also of actual allele frequencies. Thus, additive genetic variance in heterozygosity is only partly determined by the variance in allele frequencies and the number of alleles. Because the variance in allele frequencies (which, like any summary statistic, loses some information about the underlying data) is similar for certain combinations of quite dissimilar individual allele frequencies, relatively large differences in resulting $\sigma_{AR}^2$ may result from relatively similar variances in allele frequencies. Because $\sigma_{AR}^2$

occurs in the expressions for $\sigma_{DR}^2 = \sigma_G^2 - \sigma_{AR}^2$ and $h^2 = \sigma_{AR}^2/\sigma_G^2$, also $\sigma_{DR}^2$ and $h^2$ are functions of variance in allele frequencies, number of alleles, as well as individual allele frequencies. Thus, it will usually be easier to predict genetic variance components of heterozygosity based only on allele frequencies using equations 2, 5b, 6, and 7.

*Mean heterozygosity*

We now analyse how $\sigma_G^2$, $\sigma_{AR}^2$, $\sigma_{DR}^2$, and $h^2$ for heterozygosity depend on mean heterozygosity $\mu_G$. We rearrange $\mu_G = 1 - \sum_{i=1}^n p_i^2$ to get

$$\sum_{i=1}^n p_i^2 = 1 - \mu_G. \tag{S1.4}$$

Substituting equation S1.4 into equation 5b yields

$$\sigma_G^2 = 1 - \mu_G - (1 - \mu_G)^2$$

$$= \mu_G - \mu_G^2. \tag{S1.5}$$

Thus, variance in heterozygosity is a quadratic function of mean heterozygosity, independent of the number of alleles at the locus. Therefore, by knowing mean heterozygosity we can exactly predict variance in heterozygosity.

Substituting equation S1.4 into equation 2 shows that additive genetic variance is again a quadratic function of mean heterozygosity, but also of actual allele frequencies:

$$\sigma_{AR}^2 = 2 \left( \sum_{i=1}^n p_i [1 - \mu_G - p_i]^2 \right). \tag{S1.6}$$

Hence, similar to when describing the variance components of heterozygosity as functions of variance in allele frequencies, we find that mean heterozygosity only partly describes additive

genetic variance in heterozygosity. Because $\sigma^2_{AR}$ occurs in the expressions for $\sigma^2_{DR} = \sigma^2_G - \sigma^2_{AR}$ and $h^2 = \sigma^2_{AR}/\sigma^2_G$, also $\sigma^2_{DR}$ and $h^2$ are functions of mean heterozygosity, as well as individual allele frequencies. It will therefore usually be easier to predict genetic variance components of heterozygosity based only on allele frequencies using equations 2, 5b, 6, and 7.

# SUPPLEMENTARY INFORMATION 2: CORRECTION OF FORMULA FOR ADDITIVE EFFECT AND ADDITIVE GENETIC VARIANCE UNDER INBREEDING

$\sigma_{AF}^2$ can be derived by decomposing genetic variance of an inbred population into only two components following methods from Kempthorne (1957, page 350). However, the derivation by Kempthorne (1957, page 350) contains a mistake, leading to an erroneous formula that is also repeated in Lynch and Walsh (1998, equation 4.24). The error in this formula lies in its second term in the version of Lynch and Walsh (1998, equation 4.24), which becomes obvious when studying a case of symmetrical overdominance such as heterozygosity. In such a case all homozygotes have the same genotypic value, which leads to subtracting a constant value from all inbreeding-corrected additive effects. Thus, it is possible (and likely) that the additive effects of all alleles become negative, which violates the requirement of $\sum_{i=1}^n p_i \alpha_i = 0$ (Kempthorne 1957, page 350).

The root of the error lies in the derivation of the additive effects $(\alpha_k)_F$ in an inbred population, where Kempthorne uses the genotypic deviations $y_{ij}$ from the mean genotypic value $\mu_R$ of a randomly mating population with the same allele frequencies as the focal inbred population. This assumption is evident from Kempthorne's substitution of $(1 - F) \sum_{j=1}^n p_j y_{ij}$ with the additive effects of allele $k$ in a randomly mating population $(\alpha_k)_R$. However, this is not allowed here, because $y_{ij}$ in the formula for $(\alpha_k)_F$ should be evaluated against the mean $\mu_F$ of the inbred population, thus $y_{ij} = \mu_F - G_{ij}$.

We now derive the relationship between $\mu_R$ and $\mu_F$ and then use this relationship to derive a correct prediction of $(\alpha_k)_F$ that will be used to calculate $\sigma_{AF}^2$. The mean genotypic value $\mu_F$ of an inbred population differs from the value $\mu_R$ of a randomly mating population in the following

way: The genotypic values $G_{ij}$ of heterozygotes lost due to inbreeding have a mean of $F \sum_{i,j=1,i\neq j}^{n} p_i p_j G_{ij}$ (frequency of heterozygotes e.g. from Halliburton 2004, equation 3.22). All heterozygotes lost due to inbreeding are instead counted as homozygotes with a mean genotypic value of $\sum_{i=1}^{n}(F \sum_{j=1,i\neq j}^{n} p_i p_j G_{ii})$, which can also be written as $\sum_{i=1}^{n}(F p_i G_{ii} - F p_i^2 G_{ii})$, or $F \sum_{i=1}^{n} G_{ii} - F \sum_{i=1}^{n} p_i^2 G_{ii}$. $\mu_F$ can be obtained from $\mu_R$ by subtracting the mean genotypic values of the heterozygotes lost and adding the mean of the homozygotes gained due to inbreeding, resulting in

$$\mu_F = \mu_R - F \sum_{i,j=1,i\neq j}^{n} p_i p_j G_{ij} + F \sum_{i=1}^{n} G_{ii} - F \sum_{i=1}^{n} p_i^2 G_{ii}, \tag{S2.1}$$

which reduces to

$$\mu_F = \mu_R - F \sum_{i,j=1}^{n} p_i p_j G_{ij} + F \sum_{i=1}^{n} G_{ii}.$$

Because

$$\sum_{i,j=1}^{n} p_i p_j G_{ij} = \mu_R,$$

we get

$$\mu_F = (1 - F)\mu_R + F \sum_{i=1}^{n} G_{ii}. \tag{S2.2}$$

Having derived $\mu_F$ as a function of $\mu_R$, we show now how a correct formula for $(\alpha_k)_F$ can be obtained and then apply this formula to the case of heterozygosity.

Kempthorne (1957, page 350) provides the following equation for the additive effect of allele $k$ in an inbred population:

$$(1 + F)(\alpha_k)_F = F\, y_{kk} + (1 - F) \sum_{l=1}^{n} p_l y_{kl}.$$

We start from this equation and use $i, j, k, l$ as indicators of the $n$ alleles at a locus (we need these four indicators for the two alleles at a locus to avoid confusion about summation), $y_{ij} = \mu_F - G_{ij}$, and equation S2.2 we get

$$(1 + F)(\alpha_k)_F = F \left( (1 - F)\mu_R + F \sum_{i=1}^{n} G_{ii} - G_{kk} \right)$$

$$+ (1 - F) \sum_{l=1}^{n} p_l \left( (1 - F)\mu_R + F \sum_{i=1}^{n} G_{ii} - G_{kl} \right),$$

which can be expanded to

$$(1 + F)(\alpha_k)_F = F\mu_R - F^2\mu_R + F^2 \sum_{i=1}^{n} G_{ii} - F G_{kk} + (1 - F) \sum_{l=1}^{n} p_l(\mu_R - G_{kl})$$

$$+ (1 - F) \sum_{l=1}^{n} p_l \left( F \sum_{i=1}^{n} G_{ii} - F\,\mu_R \right).$$

Using $(\alpha_k)_R = \sum_{l=1}^{n}(\mu_R - G_{kl})$ from Kempthorne (1957, page 350) as well as some other rearrangements, we get

$$(1 + F)(\alpha_k)_F = (1 - F)(\alpha_k)_R + F\mu_R - F^2\mu_R + F^2 \sum_{i=1}^{n} G_{ii} - FG_{kk}$$

$$+ (1 - F) \sum_{l=1}^{n} p_l \left( F \sum_{i=1}^{n} G_{ii} \right) - F(1 - F)\mu_R \sum_{l=1}^{n} p_l$$

$$= (1 - F)(\alpha_k)_R + F\mu_R - F^2\mu_R + F^2 \sum_{i=1}^{n} G_{ii} - FG_{kk} + (1 - F)F \sum_{i=1}^{n} G_{ii} - F(1 - F)\mu_R$$

$$= (1 - F)(\alpha_k)_R + F\mu_R - F^2\mu_R + F^2 \sum_{i=1}^{n} G_{ii} - FG_{kk} + F \sum_{i=1}^{n} G_{ii} - F^2 \sum_{i=1}^{n} G_{ii} - F\mu_R + F^2\mu_R$$

$$= (1 - F)(\alpha_k)_R + F \left( \sum_{i=1}^{n} G_{ii} - G_{kk} \right).$$

We thus get the general equation that replaces the wrong equation 4.24 in Lynch and Walsh (1998):

$$(\alpha_k)_F = \frac{(1 - F)(\alpha_k)_R + F \left( \sum_{i=1}^{n} G_{ii} - G_{kk} \right)}{1 + F}.$$

*Heterozygosity*

When describing heterozygosity as a trait, the genotypic values of homozygotes $G_{kk}$ or $G_{ii}$ equal 0 and thus,

$$(\alpha_k)_F = \frac{1 - F}{1 + F}(\alpha_k)_R.$$

In inbred populations, $(\alpha_k)_F$ has to be used in equation 4.23c of Lynch and Walsh (1998), so that by modifying our equation 2 (main text), we get the equation for additive genetic variance of heterozygosity under inbreeding,

$$\sigma^2_{AF} = 2\,(1 + F)\left(\sum_{i=1}^{n} p_i \left\{ \frac{1-F}{1+F}\left[\left(\sum_{k=1}^{n} p_k^2\right) - p_i\right]\right\}^2\right). \tag{S2.3}$$

Equation S2.3 is numerically equivalent to equation 1 from Cockerham and Weir (1984), which is equation 9 in the main text. For a biallelic locus, equation S2.3 is equivalent to

$$\sigma^2_{AF} = \frac{(1-F)^2}{1+F}\, 2pq[(q-p)^2]. \tag{S2.4}$$

Equation S2.4 is also given by Falconer (1985) when setting $a = 0$ and $d = 1$, as is appropriate for heterozygosity.