LJMU Research Online

Stoffel, MA, Esser, M, Kardos, M, Humble, E, Nichols, HJ, David, P and Hoffman, JI

inbreedR: An R package for the analysis of inbreeding based on genetic markers

http://researchonline.ljmu.ac.uk/id/eprint/3593/

Article

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

Stoffel, MA, Esser, M, Kardos, M, Humble, E, Nichols, HJ, David, P and Hoffman, JI (2016) inbreedR: An R package for the analysis of inbreeding based on genetic markers. Methods in Ecology and Evolution, 7 (11). pp. 1331-1339. ISSN 2041-210X

Supporting Information

Appendix S1

g_2 statistics

Consider a sufficiently large population of individuals in inbreeding equilibrium. Sample N individuals out of this population, each sequenced at a set of loci $\{1, \dots, L\}$. Strong inbreeding increases the dependence of loci being homozygous within an individual. Thus, inbreeding is assumed to have an effect on the joint distribution of homozygous loci within an individual compared to the marginal assortment of single-locus ones. Based on David *et al.* (2007) we recall three different estimators for the second-order heterozygosity disequilibrium g_2, which reflects the excess of joint heterozygous loci relative to their expectation under random assortment. For a better comparison to the implementation in R, see the following table:

Estimator	Dataset	Equation	R function
$\widehat{\mathrm{g}}_2$	small datasets without missing values	eqn 3	
$\widehat{\mathrm{g}}'_{\underline{2}}$	small datasets that include missing values	eqn 5	`g2_microsats`
$\widehat{\widehat{\mathrm{g}}}_{\underline{2}}$	large datasets; missing values do not differ too much across loci (see last paragraph)	eqn 7	`g2_snps`

Notations and Mathematical background

Let h_i denote the true heterozygosity at locus i in the population. Due to scoring artefacts, the true value might differ from the apparent one, which will be denoted here by H_i. For our sample, define the indicator function H_{ik}, $i = 1, \dots, L$, $k = 1, \dots, N$ as follows: Set $H_{ik} = 1$, if locus i is heterozygous in individual k and set $H_{ik} = 0$ if locus i is found homozygous.

Following Weir & Cockerham (1973), the second-order heterozygosity disequilibrium $g_2(i, j)$ between loci i and j might be quantified through the identity disequilibrium $\mathbb{E}[h_i h_j] = \mathbb{E}[h_i]\,\mathbb{E}[h_j](1 + g_2(i, j))$. If one assumes independent scoring artefacts across the set of loci, then this identity also holds for the apparent heterozygosity, i.e. $\mathbb{E}[H_i H_j] = \mathbb{E}[H_i]\,\mathbb{E}[H_j](1 + g_2(i, j))$. Commonly, g_2 is assumed to be constant for every pair of loci and defined via

$$\mathbb{E}[H_i H_j] = \mathbb{E}[H_i]\,\mathbb{E}[H_j](1 + g_2). \qquad \text{(eqn 1)}$$

In practice, tightly linked pairs of loci probably have a higher g_2, which is why in the following we will give an alternative to eqn 1 by averaging out over all locus pairs (i, j). To avoid confusions, let us denote the 'averaged' g_2 by $\bar{g}_2$, which will serve for a more robust estimator and goes back to David *et al.* (2007). The well known decomposition for the variance $\mathrm{Var}\left[\sum_i H_i\right] = \sum_i \mathrm{Var}\left[H_i\right] + \sum_i \sum_{j \neq i} \mathrm{Cov}\left[H_i, H_j\right]$, together with

$$\mathrm{Cov}\left[H_i, H_j\right] = \mathbb{E}[H_i H_j] - \mathbb{E}[H_i]\,\mathbb{E}[H_j] = \mathbb{E}[H_i]\,\mathbb{E}[H_j]\; g_2$$

leads to an expression of $\bar{g}_2$ of the form

$$\bar{g}_2 = \frac{\sum_{i=1}^{L} \sum_{j \neq i} \mathbb{E}[H_i\, H_j]}{\sum_{i=1}^{L} \sum_{j \neq i} \mathbb{E}[H_i]\,\mathbb{E}[H_j]} - 1. \qquad \text{(eqn 2)}$$

For the averaged quantity $\bar{g}_2$ in eqn 2 one can find an estimator given in eqn (8) in David *et al.* (2007) (corrected for typographical errors) given by

$$\widehat{g}_2 = \frac{\sum_{i=1}^{L} \sum_{j \neq i} \left(\sum_{k=1}^{N} H_{ik} H_{jk} \right)}{\frac{1}{N-1} \sum_{i=1}^{L} \sum_{j \neq i} \left(\sum_{k_1=1}^{N} \sum_{k_2 \neq k_1} H_{ik_1} H_{jk_2} \right)} - 1, \qquad \text{(eqn 3)}$$

with little bias of order $1/N$ (see Appendix S1 in (David *et al.*, 2007)). Problematically, real data sets, especially *microsatellites* ones, do have missing values, such that the apparent value H_{ik} might be unknown for some pairs (i, k). In this case define

$$\widetilde{H}_{ik} = \begin{cases} 1, & \text{if locus } i \text{ is heterozygous in individual } k \\ 0, & \text{if locus } i \text{ is either homozygous in individual } k, \text{ or unknown,} \end{cases}$$

as well as $M_{ik} = 1$ if the datum is missing at locus i in individual k and $M_{ik} = 0$ otherwise.

The expected values of H_i and $\widetilde{H}_i$ are correlated via $\mathbb{E}[\widetilde{H}_i] = (1 - m_i)\,\mathbb{E}[H_i]$, where $m_i := \frac{1}{N} \sum_{k=1}^{N} M_{ik}$ is the proportion of individuals with missing data at locus i. Equivalently, for the joint distribution one finds the identity $\mathbb{E}[\widetilde{H}_i\, \widetilde{H}_j] = (1 - m_i - m_j + m_{ij})\,\mathbb{E}[H_i\, H_j]$, with m_{ij} being the proportion of individuals with missing values both at loci i and j. Note, that $1 - m_i - m_j + m_{ij}$ is the exact proportion of individuals with non-missing values at both, loci i and j. The analogue of eqn 1 now reads

$$\mathbb{E}[\widetilde{H}_i \widetilde{H}_j] = \frac{(1 - m_i - m_j + m_{ij})}{(1 - m_i)(1 - m_j)}\, \mathbb{E}[\widetilde{H}_i]\,\mathbb{E}[\widetilde{H}_j](1 + g_2),$$

which, with the same procedure as above leads to the more robust averaged parameter

$$\bar{g}_{\underline{2}} = \frac{\sum_{i=1}^{L} \sum_{j\neq i} \mathbb{E}[\widetilde{H}_{ij}\, \widetilde{H}_j]}{\sum_{i=1}^{L} \sum_{j\neq i} \frac{(1-m_i-m_j+m_{ij})}{(1-m_i)(1-m_j)}\, \mathbb{E}[\widetilde{H}_i]\, \mathbb{E}[\widetilde{H}_j]} - 1,$$

where the underline stands for presence of missing data and the overbar again indicates averaging over all locus pairs. Eqn S1 in the Supplementary Informations in Hoffman *et al.* (2014) provide an estimator for $\bar{g}_{\underline{2}}$, which can be rewritten as

$$\widehat{\mathrm{g}}_{\underline{2}} = \frac{\sum_{i=1}^{L} \sum_{j\neq i} \frac{1}{N-M_i-M_j+M_{ij}} \left(\sum_{k=1}^{N} \widetilde{H}_{ik} \widetilde{H}_{jk} \right)}{\sum_{i=1}^{L} \sum_{j\neq i} \frac{1}{(N-1)(N-M_i-M_j)+M_i M_j - M_{ij}} \left(\sum_{k_1=1}^{N} \sum_{k_2\neq k_1} \widetilde{H}_{ik_1} \widetilde{H}_{ik_2} \right)} - 1, \qquad \text{(eqn 4)}$$

where the M_i, $M_{ij} \in \{1, \ldots, N\}$, compared to the m_i, m_{ij}, now relate to the absolute values of missing data at some loci in contrast to the relative ones. Eventhough, eqn S1 in Hoffman *et al.* (2014) differs from (eqn 4) at first glance, it is straightforward to verify their equality.

Unfortunately, eqn 4 entails a weighting problem for loci with many missing data. To see this, image a pair of loci (i, j) with a considerable high fraction of missing values in the sample. In this case, any individual k which is heterozygous at both loci is downweighted, whereas in case of no missing data, it will be fully taken into account. Therefore one can update the formula to

$$\widehat{\mathrm{g}}'_{\underline{2}} = \frac{\left[\sum_{i=1}^{L} \sum_{j\neq i} \left((N-1)(N-M_i-M_j) + M_i M_j - M_{ij}\right) \right] \sum_{i=1}^{L} \sum_{j\neq i} \left(\sum_{k=1}^{N} \widetilde{H}_{ik} \widetilde{H}_{jk} \right)}{\left[\sum_{i=1}^{L} \sum_{j\neq i} (N - M_i - M_j + M_{ij}) \right] \left(\sum_{i=1}^{L} \sum_{j\neq i} \sum_{k_1=1}^{N} \sum_{k_2\neq k_1} \widetilde{H}_{ik_1} \widetilde{H}_{ik_2} \right)} - 1, \qquad \text{(eqn 5)}$$

which now weights with the appropriate number of double non-missing data over all locus pairs (Hardy, 2015). It is indeed eqn 5 (not eqn 4), which is implemented in the RMES software by David *et al.* (2007). The R package `g2_microsats` is also based on eqn 5.

All estimators we have seen so far, require double summation over all loci and are thus unfeasible for large datasets. To fasten the algorithm, the Supplementary Information in Hoffman *et al.* (2014) provides another estimator based on the general decomposition of a double sum of the form $\sum_i \sum_{j\neq i} a_i\, a_j$ into $\sum_{i=1}^{n} \sum_{j\neq i} a_i\, a_j = \left(\sum_{i=1} a_i \right)^2 - \sum_{i=1} a_i^2$, where $a_i \in \mathbb{R}$ is some parameter and $n \in \mathbb{N}$. The estimator reads

$$\widehat{\widehat{\mathrm{g}}}_{\underline{2}} = \frac{1 + (\widehat{\mathrm{B}} - \widehat{\mathrm{C}})/\widehat{\mathrm{A}}}{1 + \hat{\hat{a}}} - 1, \qquad \text{(eqn 6)}$$

where $\widehat{\mathrm{A}}$, $\widehat{\mathrm{B}}$, $\widehat{\mathrm{C}}$, $\hat{\tilde{a}}$ are encapsulated weighted sums of $\widetilde{H}_{ik}$:

$$\widehat{\mathrm{A}} = \frac{N}{N-1}\left[\left(\sum_{i=1}^{L}\hat{\tilde{\mu}}_i\right)^2 - \sum_{i=1}^{L}\hat{\tilde{\mu}}_i^2\right] - \frac{\widehat{\mathrm{J}}}{N-1}, \qquad \widehat{\mathrm{J}} = \frac{1}{N}\sum_{k=1}^{N}\tilde{h}_k^2 - \sum_{i=1}^{L}\hat{\tilde{\mu}}_i,$$

$$\widehat{\mathrm{B}} = \frac{1}{N-1}\left[\sum_{k=1}^{N}\tilde{h}_k^2 - \frac{1}{N}\left(\sum_{k=1}^{N}\tilde{h}_k\right)^2\right], \qquad \widehat{\mathrm{C}} = \frac{N}{N-1}\left[\sum_{i=1}^{L}\hat{\tilde{\mu}}_i - \sum_{i=1}^{L}\hat{\tilde{\mu}}_i^2\right],$$

$$\hat{\tilde{a}} = \frac{\sum\limits_{k=1}^{N}\widehat{\mathrm{M}}_k - N\left(\sum\limits_{i=1}^{L}\hat{\tilde{\mu}}_i\ \frac{m_i}{1-m_i}\right)^2 + N\sum\limits_{i=1}^{L}\left(\hat{\tilde{\mu}}_i\ \frac{m_i}{1-m_i}\right)^2}{(N-1)\widehat{\mathrm{A}} + \widehat{\mathrm{J}}}, \qquad \widehat{\mathrm{M}}_k = \left(\sum_{i=1}^{L}\frac{\hat{\tilde{\mu}}_i\ x_{ik}}{1-m_i}\right)^2 - \sum_{i=1}^{L}\left(\frac{\hat{\tilde{\mu}}_i\ x_{ik}}{1-m_i}\right)^2,$$

with $\hat{\tilde{\mu}}_i := \frac{1}{N}\sum_{i=1}^{N}\widetilde{H}_{ik}$, $x_{ik} := M_{ik}$ and $h_k := \sum_{k=1}^{L}\widetilde{H}_{ik}$. A straightforward calculation leads to a simplification of eqn 6 that uses less normalisation steps, represented as

$$\hat{\hat{\mathrm{g}}}_{\underline{2}} = \frac{\widehat{\mathrm{D}}}{1\hat{+}\mathrm{E}} - 1, \qquad \text{(eqn 7)}$$

with

$$\widehat{\mathrm{D}} := (N-1)\,\frac{\sum\limits_{k=1}^{N}\left(\widetilde{H}_{.k}\right)^2 - \widetilde{H}_{..}}{\left(\widetilde{H}_{..}\right)^2 - \sum\limits_{i=1}^{L}\left(\widetilde{H}_{i.}\right)^2 - \sum\limits_{k=1}^{N}\left(\widetilde{H}_{.k}\right)^2 + \widetilde{H}_{..}}$$

and

$$\widehat{\mathrm{E}} := \frac{\frac{1}{N}\sum\limits_{k=1}^{N}\left[\left(\sum\limits_{i=1}^{L}\widetilde{H}_{i.}\ \frac{M_{ik}}{1-m_i}\right)^2 - \sum\limits_{i=1}^{L}\left(\widetilde{H}_{i.}\ \frac{M_{ik}}{1-m_i}\right)^2\right] + \sum\limits_{i=1}^{L}\left(\widetilde{H}_{i.}\ \frac{m_i}{1-m_i}\right)^2 - \left(\sum\limits_{i=1}^{L}\widetilde{H}_{i.}\ \frac{m_i}{1-m_i}\right)^2}{\left(\widetilde{H}_{..}\right)^2 - \sum\limits_{i=1}^{L}\left(\widetilde{H}_{i.}\right)^2},$$

where we used the common notation for the marginalisation over all individuals or respectively over all loci; more precisely

$$\widetilde{H}_{.k} = \sum_{i=1}^{L}\widetilde{H}_{ik}, \qquad \widetilde{H}_{i.} = \sum_{k=1}^{N}\widetilde{H}_{ik}, \qquad \widetilde{H}_{..} = \sum_{i=1}^{L}\sum_{k=1}^{N}\widetilde{H}_{ik}\ .$$

As mentioned in Hoffman *et al.* (2014), the estimator in eqn 6 and thus also the one in eqn 7 are subject to the assumption, that for each pair (i, j), the term $\frac{m_{ij} - m_i \cdot m_j}{(1-m_i)(1-m_j)}$ can be approximated by the average over all pairs of loci. Thus, $\hat{\hat{\mathrm{g}}}_{\underline{2}}$ only serves as an estimator for g_2, if in the underlying data set, the missing values between pairs do not differ greatly in frequency (which could potentially occur if data quality is very poor for certain individuals).

References

David, P., Pujol, B., Viard, F., Castella, V. & Goudet, J. (2007) Reliable selfing rate estimates from imperfect population genetic data. *Molecular Ecology*, **16**, 2474–2487.

Hardy, O.J. (2015) Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Molecular ecology resources*.

Hoffman, J.I., Simpson, F., David, P., Rijks, J.M., Kuiken, T., Thorne, M.A.S., Lacy, R.C. & Dasmahapatra, K.K. (2014) High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Sciences*, **111**, 3775–3780.

Weir, B. & Cockerham, C.C. (1973) Mixed self and random mating at two loci. *Genetical research*, **21**, 247–262.