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/

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


LJMU has developed LJMU Research Online for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk

http://researchonline.ljmu.ac.uk/
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 $\{1, \ldots, L\}$. 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:

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Dataset</th>
<th>Equation</th>
<th>R function</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{g}_2$</td>
<td>small datasets without missing values</td>
<td>eqn 3</td>
<td></td>
</tr>
<tr>
<td>$\hat{g}_2'$</td>
<td>small datasets that include missing values</td>
<td>eqn 5</td>
<td>g2_microsats</td>
</tr>
<tr>
<td>$\hat{g}_2^\ast$</td>
<td>large datasets; missing values do not differ too much across loci (see last paragraph)</td>
<td>eqn 7</td>
<td>g2_snps</td>
</tr>
</tbody>
</table>

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 $H_{ik}$, $i = 1, \ldots, L$, $k = 1, \ldots, N$ 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 $E[h_i h_j] = E[h_i] E[h_j](1 + g_2(i, j))$. If one assumes independent scoring artefacts across the set of loci, then this identity also holds for the apparent heterozygosity, i.e. $E[H_i H_j] = E[H_i] E[H_j](1 + g_2(i, j))$. Commonly, $g_2$ is assumed to be constant for every pair of loci and defined via

$$E[H_i H_j] = E[H_i] E[H_j](1 + g_2).$$  \hspace{1cm} (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 $\text{Var} \left( \sum_i H_i \right) = \sum_i \text{Var} \left[ H_i \right] + \sum_i \sum_{j \neq i} \text{Cov} \left[ H_i, H_j \right]$, together with

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

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

$$
\tilde{g}_2 = \frac{\sum_{i=1}^{L} \sum_{j \neq i} \mathbb{E}[H_i H_j]}{\sum_{i=1}^{L} \sum_{j \neq i} \mathbb{E}[H_i] \mathbb{E}[H_j]} - 1. \tag{eqn 2}
$$

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

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

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

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

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 $\tilde{H}_i$ are correlated via $\mathbb{E}[\tilde{H}_i] = (1 - m_i) \mathbb{E}[H_i]$, 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 $\mathbb{E}[\tilde{H}_i \tilde{H}_j] = (1 - m_i - m_j + m_{ij}) \mathbb{E}[H_i H_j]$, 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

$$
\mathbb{E}[\tilde{H}_i \tilde{H}_j] = \frac{(1 - m_i - m_j + m_{ij})}{(1 - m_i)(1 - m_j)} \mathbb{E}[\tilde{H}_i] \mathbb{E}[\tilde{H}_j](1 + g_2),
$$
which, with the same procedure as above leads to the more robust averaged parameter

\[
\hat{g}_2 = \frac{\sum_{i=1}^{L} \sum_{j \neq i} \mathbb{E}[\tilde{H}_{ij} \tilde{H}_j]}{\sum_{i=1}^{L} \sum_{j \neq i} \frac{1}{(1-M_i-M_j+M_{ij})} \mathbb{E}[\tilde{H}_{ij}]} - 1,
\]

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 \(\hat{g}_2\), which can be rewritten as

\[
\hat{g}_2 = \frac{\sum_{i=1}^{L} \sum_{j \neq i} (N-M_i-M_j+M_{ij}) \left( \sum_{k=1}^{N} \tilde{H}_{ik} \tilde{H}_{jk} \right)}{\sum_{i=1}^{L} \sum_{j \neq i} (N-1)(N-M_i-M_j+M_{ij}) \left( \sum_{k_1=1}^{N} \sum_{k_2 \neq k_1} \tilde{H}_{ik_1} \tilde{H}_{ik_2} \right)} - 1, \tag{eqn 4}
\]

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

\[
\hat{g}_2' = \frac{\sum_{i=1}^{L} \sum_{j \neq i} (N-1)(N-M_i-M_j+M_{ij}) \left( \sum_{k=1}^{N} \tilde{H}_{ik} \tilde{H}_{jk} \right)}{\sum_{i=1}^{L} \sum_{j \neq i} (N-M_i-M_j+M_{ij}) \left( \sum_{k_1=1}^{N} \sum_{k_2 \neq k_1} \tilde{H}_{ik_1} \tilde{H}_{ik_2} \right)} - 1, \tag{eqn 5}
\]

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 \texttt{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 \(\sum_{i=1}^{n} \sum_{j \neq i} a_i a_j = \left( \sum_{i=1}^{n} a_i \right)^2 - \sum_{i=1}^{n} a_i^2\), where \(a_i \in \mathbb{R}\) is some parameter and \(n \in \mathbb{N}\). The estimator reads

\[
\hat{g}_2 = \frac{1 + (\tilde{B} - \tilde{C})/\tilde{A}}{1 + \tilde{A}} - 1, \tag{eqn 6}
\]
where $\hat{A}$, $\hat{B}$, $\hat{C}$, $\hat{a}$ are encapsulated weighted sums of $\tilde{H}_{ik}$:

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

$$
\hat{C} = \frac{N}{N-1} \left[ \sum_{i=1}^{L} \hat{\mu}_i - \sum_{i=1}^{L} \hat{\mu}_i^2 \right],
$$

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

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

$$
\widehat{\alpha^2} = \frac{\hat{D}}{1+E} - 1, \quad \text{(eqn 7)}
$$

with

$$
\hat{D} := (N-1) \frac{\sum_{k=1}^{N} \left( \tilde{H}_k \right)^2 - \tilde{H}_i}{\left( \tilde{H}_i \right)^2 - \sum_{i=1}^{L} \left( \tilde{H}_i \right)^2 - \sum_{k=1}^{N} \left( \tilde{H}_k \right)^2 + \tilde{H}_i}
$$

and

$$
\widehat{E} := \frac{1}{N} \sum_{k=1}^{N} \left[ \left( \sum_{i=1}^{L} \tilde{H}_i \frac{M_{ik}}{1-m_i} \right)^2 - \sum_{i=1}^{L} \left( \tilde{H}_i \frac{M_{ik}}{1-m_i} \right)^2 \right] + \frac{L}{\sum_{i=1}^{L} \left( \tilde{H}_i \frac{M_{ik}}{1-m_i} \right)^2} - \left( \sum_{i=1}^{L} \tilde{H}_i \frac{M_{ik}}{1-m_i} \right)^2
$$

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

$$
\tilde{H}_{ik} = \sum_{i=1}^{L} \tilde{H}_{ik}, \quad \tilde{H}_{i} = \sum_{k=1}^{N} \tilde{H}_{ik}, \quad \tilde{H} = \sum_{i=1}^{L} \sum_{k=1}^{N} \tilde{H}_{ik}
$$

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 $\frac{m_i - m_j}{(1-m_i)(1-m_j)}$ can be approximated by the average over all pairs of loci. Thus, $\hat{\alpha^2}$ 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


