APPLICATION

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

Martin A. Stoffel1,2*, Mareike Esser3, Marty Kardos4, Emily Humble1,5, Hazel Nichols2, Patrice David6 and Joseph I. Hoffman1

1Department of Animal Behaviour, Bielefeld University, Box 100131 Bielefeld, Germany; 2Faculty of Science, School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool L3 3AF, UK; 3Faculty of Technology, Bielefeld University, Box 100131 Bielefeld, Germany; 4Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Norbyvägen 18D, Uppsala 75236, Sweden; 5British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 OET, UK; and 6Centre d’Ecologie Fonctionnelle et Evolutive, Centre National de la Recherche Scientifique, 34293 Montpellier, France

Summary

1. Heterozygosity–fitness correlations (HFCs) have been widely used to explore the impact of inbreeding on individual fitness. Initially, most studies used small panels of microsatellites, but more recently with the advent of next-generation sequencing, large SNP datasets are becoming increasingly available and these provide greater power and precision to quantify the impact of inbreeding on fitness.

2. Despite the popularity of HFC studies, effect sizes tend to be rather small. One reason for this may be low variation in inbreeding levels among individuals. Using genetic markers, it is possible to measure variance in inbreeding through the strength of correlation in heterozygosity across marker loci, termed identity disequilibrium (ID).

3. ID can be quantified using the measure $g^2$, which is also a central parameter in HFC theory that can be used within a wider framework to estimate the direct impact of inbreeding on both marker heterozygosity and fitness. However, no software exists to calculate $g^2$ for large SNP datasets nor to implement this framework.

4. inbreedR is an R package that provides functions to calculate $g^2$ based on microsatellite and SNP markers with associated $P$-values and confidence intervals. Within the framework of HFC theory, inbreedR also estimates the impact of inbreeding on marker heterozygosity and fitness. Finally, inbreedR implements user-friendly simulations to explore the precision and magnitude of estimates based on different numbers of genetic markers. We hope this package will facilitate good practice in the analysis of HFCs and help to deepen our understanding of inbreeding effects in natural populations.

Key-words: genetic marker, heterozygosity, heterozygosity-fitness correlation (HFC), identity disequilibrium, inbreeding

Introduction

Offspring of close relatives often show reduced fitness, a phenomenon referred to as inbreeding depression (Charlesworth & Charlesworth 1987; Charlesworth & Willis 2009). This decline in fitness among inbred individuals is a result of the increased proportion of loci in the genome that are identical by descent (IBD). A homozygous locus is IBD or autozygous when it carries two alleles that both originate from a single copy in a common ancestor. An increased proportion of loci in the genome that are identical by descent (IBD$_G$) may lead to the unmasking of deleterious recessive alleles and a reduction in heterozygote advantage by decreasing genomewide heterozygosity (Charlesworth & Charlesworth 1987; Charlesworth & Willis 2009). In populations with unknown pedigrees, many studies have used genetic marker heterozygosity as a measure of IBD$_G$. The result is a large and expanding literature describing heterozygosity-fitness correlations (HFCs) across a range of species and traits (Coltman & Slate 2003; Chapman et al. 2009; Szulkin, Bierne & David 2010).

Despite the large and growing number of HFC studies, effect sizes are usually small (Chapman et al. 2009) and there has been debate over their mechanistic basis (Balloux, Amos & Coulson 2004; Slate et al. 2004; Hansson & Westerberg 2008; Szulkin, Bierne & David 2010). This reflects the fact that under many circumstances multilocus heterozygosity based on the 10–20 microsatellite markers employed by most studies provides little power to estimate...
IBD\(_G\) (Hansson & Westerberg 2002; Balloux, Amos & Coulson 2004; Szulkin, Bierne & David 2010; Hoffman et al. 2014). This is why the pedigree-derived inbreeding coefficient (F\(P\)) has long been the gold standard for estimating IBD\(_G\) (Pemberton, 2004, 2008). F\(P\) is defined as the probability of a given locus in an individual’s genome being autozygous based on its pedigree. However, an individual’s F\(P\) will differ from its IBD\(_G\) as F\(P\) can be imprecise due to linkage among loci and downwardly biased due to incomplete pedigree information (Hill & Weir 2011; Keller, Visscher & Goddard 2011; Kardos, Luikart & Allendorf 2015). Consequently, IBD\(_G\) can vary substantially among individuals with the same F\(P\) (Franklin 1977; Hill & Weir 2011; Forstmeier et al. 2012). In other words, even F\(P\) derived from a perfect pedigree cannot fully capture the variance in genomic autozygosity (\(\sigma^2(\text{IBD}_G)\)) among individuals, as it does not incorporate variation due to linkage.

Recent advances in next-generation sequencing technology (e.g. Baird et al. 2008; Peterson et al. 2012) now allow many tens or even hundreds of thousands of single-nucleotide polymorphisms (SNPs) to be genotyped in virtually any organism. Applied to HFCs, these dense marker panels provide much greater power then a small panel of microsatellites to quantify the impact of inbreeding on fitness (Hoffman et al. 2014). Recent simulation and empirical studies also suggest that inbreeding coefficients based on genomewide SNP data provide more precise measures of IBD\(_G\) and inbreeding depression than F\(P\) (Keller, Visscher & Goddard 2011; Pryce et al. 2014; Kardos, Luikart & Allendorf 2015; Huisman et al. 2016).

**HFC theory**

For marker loci to indicate inbreeding depression, their heterozygosity must be correlated with the heterozygosity of functional loci in the genome (Szulkin, Bierne & David 2010). Such correlations between marker loci and functional loci have been proposed to occur through two possible mechanisms: The ‘general effect hypothesis’ on the one hand assumes that multilocus heterozygosity (MLH) reflects genomewide heterozygosity. This association emerges because variation in inbreeding causes heterozygosity to be correlated across loci, a phenomenon termed identity disequilibrium (ID, Weir & Cockerham 1973). Alternatively, the ‘local effect hypothesis’ states that one or a few of the markers are in linkage disequilibrium (LD) with a trait locus under balancing selection, which creates a pattern whereby heterozygosity at the gene and marker are correlated. However, ID and LD do not necessarily have to be considered as competing hypotheses to explain HFCs as ID is a consequence and LD is a cause of variation in IBD\(_G\) (Bierne, Tsitrone & David 2000; Szulkin, Bierne & David 2010). Both mechanisms can therefore be united under an inbreeding or general effect model (Bierne, Tsitrone & David 2000).

Variance in individual inbreeding levels can be caused by a variety of scenarios other than systematic consanguineous matings (Szulkin, Bierne & David 2010). For example, in small or bottlenecked populations, \(\sigma^2(\text{IBD}_G)\) and therefore ID can occur as a consequence of variation in the relatedness of mating partners. Similarly, immigration and admixture can result in the offspring of parents from different populations being relatively outbred, leading to an increased \(\sigma^2(\text{IBD}_G)\) within a population (Tsitrone, Rousset & David 2001; Szulkin, Bierne & David 2010). In addition, in small randomly mating populations, both genetic drift and immigration generate LD (Hill & Robertson 1968; Sved 1968; Bierne, Tsitrone & David 2000), which in turn leads to ID (Szulkin, Bierne & David 2010). All of these scenarios ultimately increase \(\sigma^2(\text{IBD}_G)\) and lead to ID, which is the fundamental cause of HFCs according to the general effect model.

The general effect model assumes that HFCs arise due to the simultaneous effects of inbreeding on variation among individuals in marker heterozygosity and fitness (David et al. 1995; David 1998; Bierne, Tsitrone & David 2000; Hansson & Westerberg 2002). Specifically, inbreeding affects the genome including the panel of genetic markers by increasing the proportion of loci that are IBD and by causing ID. When the aim of a study is to infer the effects of inbreeding on fitness from a panel of genetic markers, two related questions arise: (i) How well does MLH at genetic markers reflect IBD\(_G^2\)? and (ii) How large is the inbreeding load, that is the correlation between inbreeding and fitness? These questions led to the development of a model to estimate these relationships based on the inbreeding coefficient f defined as individual IBD\(_G\) (Bierne, Tsitrone & David 2000). This model was developed further to estimate how well marker heterozygosity reflects F\(P\), which itself is an imprecise measure of IBD\(_G\), but the best that existed in pre-genomic times (Slate et al. 2004). Within this framework, Szulkin, Bierne & David (2010) used \(g_2\) (David et al. 2007), a point estimate of ID, to measure \(\sigma^2(\text{IBD}_G)\). This allows the derivation of formulae to estimate the correlations between inbreeding, MLH and fitness purely from genetic marker data.

**Quantifying effects of inbreeding on heterozygosity and fitness**

The general effect model assumes that heterozygosity at genetic markers (h, here defined as standardized MLH, Coltman et al. 1999) is correlated with genomic heterozygosity through variation in individual inbreeding levels (f) and that individual fitness (W) declines as a linear function of f, which is expected if deleterious mutations have non-epistatic effects (Bierne, Tsitrone & David 2000). In other words, the correlation between W and h arises through the simultaneous effects of inbreeding level on fitness (r(W, f)) and marker heterozygosity (rh, f) (Bierne, Tsitrone & David 2000; Slate et al. 2004; Szulkin, Bierne & David 2010).

\[
r(W, h) = r(h, f)r(W, f)
\]

**eqn 1**

Although F\(P\) has been used as a measure of f in the above formula (Slate et al. 2004; Szulkin, Bierne & David 2010), here we define the inbreeding coefficient f as a variable that explains all of the variance in genomic heterozygosity (\(\sigma^2(\text{IBD}_G)\)) and
therefore includes both variance depending on an individual's pedigree and the degree of linkage among loci (Bierne, Tsitrone & David 2000). When it is not possible to directly measure an individual's inbreeding level \( f \), we can use ID to characterize the distribution of \( f \) in a population. A measure of ID that can be related to HFC theory is \( g_2 \) (David et al. 2007), which quantifies the extent to which heterozygosities are correlated across pairs of loci (see Appendix S1 for details). Based on \( g_2 \) as an estimate of ID, it is then possible to calculate the expected correlation between \( h \) and inbreeding level \( f \) as follows (Szulkin, Bierne & David 2010):

\[
r^2(h,f) = \frac{g_2}{\sigma^2(h)}
\]

Finally, the expected squared correlation between a fitness trait \( W \) and inbreeding level \( f \) can be derived by rearranging eqn 1 (Szulkin, Bierne & David 2010):

\[
r^2(W,f) = \frac{r^2(W,h)}{r^2(h,f)}
\]

( eqn 3 )

Software is already available for calculating \( g_2 \) from microsatellite datasets (David et al. 2007). However, for larger (e.g. SNP) datasets, the original formula is not computationally practical, as it requires a double summation over all pairs of loci. For example, with 15,000 loci, the double summations take of the order of \( 0.2 \times 10^9 \) computation steps. For this reason, it is necessary to implement a computationally more feasible formula to calculate \( g_2 \), which assumes that the distribution of true heterozygosity is the same in missing data as in non-missing data, i.e. that the frequency of missing values does not vary much between pairs of loci (Hoffman et al. 2014). In turn, the \( g_2 \) parameter builds the foundation for the implementation of the above framework to analyse HFCs, which is recommended to be routinely computed in future HFC studies (Szulkin, Bierne & David 2010; Kardos, Allendorf & Luikart 2014).

The package

\textbf{inbreedR} is an R package (R Core Team 2015) that provides functions for analysing inbreeding and HFCs based on microsatellite and SNP data. The main aims of the package are to (i) calculate \( g_2 \) and its confidence interval and \( P \)-value for both microsatellites and large SNP datasets, (ii) estimate the influence of inbreeding on marker heterozygosity and fitness through the derivation of \( r^2(h,f) \) and \( r^2(W,f) \) and (iii) explore the sensitivity of \( g_2 \) and \( r^2(h,f) \) to marker number through user-friendly simulations. The overall workflow is shown in Fig. 1 and described below. For a more detailed description of the package and the functions, we have supplied a vignette for the package than can be accessed via \texttt{browseVignettes("inbreedR")} once the package is installed.

\textbf{Example datasets}

The functionality of \textbf{inbreedR} is illustrated using genetic and phenotypic data from an inbred captive population of oldfield mice (\textit{Peromyscus polionotus}) (Hoffman et al. 2014). These mice were paired over six laboratory generations to produce offspring with \( F_P \) ranging from 0 to 0.453. Example files are provided containing the genotypes of 36 \textit{P. polionotus} individuals at 12 microsatellites and 13,198 SNPs respectively. Data on body mass at

---

**Fig. 1. inbreedR workflow.** For both microsatellite and SNP datasets, the program provides utilities for data conversion and checking, estimation of identity disequilibrium, derivation of key parameters relating to HFC theory and exploration of sensitivity to the number of loci deployed. Further details are provided in the main text.
weaning, a fitness proxy, are also available for the same individuals.

```r
library(inbreedR)
data("mouse_msats") # microsatellite data
data("mouse_snps") # snp data
data("bodyweight") # fitness data
```

### Data conversion and checking

The working format of `inbreedR` is an individual x locus matrix or `data.frame` in which rows represent individuals and each column represents a locus. If an individual is heterozygous at a given locus, it is coded as 1, whereas a homozygote is coded as 0, and missing data are coded as NA. We provide a converter function from a common two-column-per-locus (allelic) format to the working format, as well as a function to check for common formatting errors within the input matrix. Guidelines for extracting genotype data from VCF files are given in the vignette.

```r
# transforms microsatellite data into (0/1)
mouse_msats <- convert_raw(mouse_msats)
# checks the data
check_data(mouse_msats, num_ind = 36, num_loci = 12)

# checks the data
check_data(mouse_snps, num_ind = 36, num_loci = 13198)
```

### Identity disequilibrium

The package provides functions to calculate \( g_2 \) for both microsatellites and SNPs. The `g2_microsats()` function implements the formula given in David et al. (2007). For large datasets (e.g. SNPs) the `g2_snps()` function implements a computationally feasible formula described in Appendix S1. For both microsatellites and SNPs, `inbreedR` also calculates confidence intervals by bootstrapping over individuals (Table 1). It also permutes the genetic data to generate a \( P \)-value for the null hypothesis of no variance in inbreeding in the sample (i.e. \( g_2 = 0 \)). The `g2_snps()` function provides an additional argument for parallelization which distributes bootstrapping and permutation across cores.

```r
g2_mouse_msats <- g2_microsats(mouse_msats, nperm = 1000, nboot = 1000, CI = 0.95) 
g2_mouse_snps <- g2_snps(mouse_snps, nperm = 100, nboot = 100, CI = 0.95, parallel = FALSE, ncores = NULL)
```

### Results of both functions can be plotted as histograms with CIs (Fig. 2)

```r
par(mfrow = c(1,2))
plot(g2_mouse_msats, main = "Microsatellites", col = "cornflowerblue", cex.axis = 0.85)
plot(g2_mouse_snps, main = "SNPs", col = "darkgoldenrod1", cex.axis = 0.85)
```

![Fig. 2. Output of the \( g_2 \) functions for the microsatellite and SNP datasets showing the distribution of \( g_2 \) estimates from bootstrap samples over individuals together with their 95% CIs. The empirical \( g_2 \) estimate is marked as a black dot along the CI.](image)

Another approach for estimating ID is to divide the marker panel into two random subsets, compute the correlation in heterozygosity between the two and repeat this hundreds or thousands of times in order to obtain a distribution of heterozygosity–heterozygosity correlation coefficients (Baloux, Amos & Coulson 2004). This approach is intuitive and has been shown to be equivalent to \( g_2 \) in its power to detect non-zero variance in inbreeding (Kardos, Allendorf & Luikart 2014) although it can be criticized on the grounds that samples within the HHC distribution are non-independent. Moreover, \( g_2 \) is preferable because it directly relates to HFC theory (eqn 2). The `HHC()` function in `inbreedR` calculates HHCs together with confidence intervals, specifying how often the dataset is randomly split into two halves with the `reps` argument.

```r
HHC_mouse_msats <- HHC(mouse_msats, reps = 1000) 
HHC_mouse_snps <- HHC(mouse_snps, reps = 100)
```

### Table 1. Output of the \( g_2 \) functions showing \( g_2 \) values and their 95% confidence intervals, standard errors and \( P \)-values for 36 mice genotyped at 12 microsatellites and 13 198 SNPs.

<table>
<thead>
<tr>
<th></th>
<th>( g_2 )</th>
<th>Lower CI</th>
<th>Upper CI</th>
<th>SE</th>
<th>( P )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsats</td>
<td>0.022</td>
<td>−0.008</td>
<td>0.065</td>
<td>0.019</td>
<td>0.076</td>
</tr>
<tr>
<td>SNPs</td>
<td>0.035</td>
<td>0.022</td>
<td>0.050</td>
<td>0.008</td>
<td>0.010</td>
</tr>
</tbody>
</table>

The results can be outputted as text (Table 2) or plotted as histograms with CIs (Fig. 3).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Lower CI</th>
<th>Upper CI</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsats</td>
<td>0.194</td>
<td>0.062</td>
<td>0.453</td>
<td>0.128</td>
</tr>
<tr>
<td>SNPs</td>
<td>0.976</td>
<td>0.961</td>
<td>0.987</td>
<td>0.007</td>
</tr>
</tbody>
</table>

The results can be outputted as text (Table 2) or plotted as histograms with CIs (Fig. 3).

```
par(mfrow=c(1,2))
plot(HHC_mouse_microsats, main = "Microsatellites", col = "cornflowerblue", cex.axis=0.85)
plot(HHC_mouse_snps, main = "SNPs", col = "darkgoldenrod1", cex.axis=0.85)
```

Fig. 3. Output of the HHC function showing the distribution of heterozygosity–heterozygosity correlation coefficients for the microsatellite and SNP datasets. Also shown are the mean HHcs as black dots and their 95% CIs. The two distributions are very different, microsatellites being positive but with the 95% CI overlapping zero, and SNPs being well in excess of 0.9 with a much greater precision. This reflects the enhanced power of the larger SNP dataset to capture variance in $f$ among individuals.

**HFC parameters**

Assuming that HFCs are due to inbreeding depression, it is possible to calculate both the expected correlation between heterozygosity and inbreeding level ($r^2(h,f)$) and the expected correlation between a fitness trait and inbreeding ($r^2(W,f)$) as described in eqn 1. These calculations are implemented in `inbreedR` using the functions `r2_hf()` and `r2_Wf()`. Both functions include an `nboot` argument to run bootstrapping over individuals and estimate confidence intervals. Similar to the `glm()` function, the distribution of the fitness trait can be specified using the family argument, as shown below:

```r
# r2 between inbreeding and heterozygosity
hf <- r2_hf(genotypes = mouse_microsats, nboot = 100, type = "msats")
# r2 between inbreeding and fitness
Wf <- r2_Wf(genotypes = mouse_microsats, trait = bodyweight, family = gaussian, type = "msats", nboot = 100)
```

**Workflow for estimating the impact of inbreeding on fitness using HFC**

Szulkin, Bierne & David (2010) in their Appendix S1 provide a worked example of how to estimate the impact of inbreeding on fitness within an HFC framework. Below, we show how the required calculations can be implemented in `inbreedR`. We start with the estimation of identity disequilibrium ($g_2$) and calculation of the variance of standardized multilocus heterozygosity ($\sigma^2(h)$), followed by the estimation of the three correlations from eqn 1. Example code for the microsatellite dataset is shown below and the results for both microsatellites and SNPs are given in Table 3.

```r
# g2 and bootstraps to estimate CI
g2 <- g2_microsats(mouse_microsats, nboot = 1000)
# calculate sMLH
het <- sMLH(mouse_microsats)
# variance in sMLH
het_var <- var(het)
# Linear model
mod <- lm(bodyweight ~ het)
# regression slope
beta <- coef(mod)[2]
# r2 between fitness and heterozygosity
Wh <- cor(bodyweight, predict(mod))^2
# r2 between inbreeding and sMLH
# including bootstraps to estimate CI
hf <- r2_hf(genotypes = mouse_microsats, type = "msats", nboot = 1000)
# r2 between inbreeding and fitness
# including bootstraps to estimate CI
Wf <- r2_Wf(genotypes = mouse_microsats, trait = bodyweight, family = gaussian, type = "msats", nboot = 1000)
```

**Table 2.** Output of the HHC function, showing mean HHcs with 95% confidence intervals and standard deviations for 36 mice genotyped at 12 microsatellites and 13 198 SNPs.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Lower CI</th>
<th>Upper CI</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsats</td>
<td>0.194</td>
<td>0.062</td>
<td>0.453</td>
<td>0.128</td>
</tr>
<tr>
<td>SNPs</td>
<td>0.976</td>
<td>0.961</td>
<td>0.987</td>
<td>0.007</td>
</tr>
</tbody>
</table>

**Table 3.** Parameters central to interpreting HFCs for the microsatellite and SNP datasets. $g_2$ is the empirical point estimate of $g_2$, $\sigma^2(h)$ is the variance in sMLH, $\hat{\beta}_{\text{in}}$ is the regression slope of sMLH in a linear model of the fitness trait, $r^2_{\text{in}}$ is the squared correlation of the fitness trait and sMLH, $r^2_{\text{hf}}$ is the expected squared correlation of sMLH and inbreeding, and $r^2_{\text{Wf}}$ is the expected squared correlation between sMLH and fitness. 95% confidence intervals are shown in squared brackets for the estimates from the package. Note that $r^2_{\text{in}}$ is an expected correlation derived from the ratio of $g_2/\sigma^2(h)$ and may slightly exceed one due to missing values; we therefore constrain the estimate between 0 and 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$g_2$</th>
<th>$\sigma^2(h)$</th>
<th>$\hat{\beta}_{\text{in}}$</th>
<th>$r^2_{\text{in}}$</th>
<th>$r^2_{\text{hf}}$</th>
<th>$r^2_{\text{Wf}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsats</td>
<td>0.022</td>
<td>0.078</td>
<td>1.601</td>
<td>0.121</td>
<td>0.280 [0.052]</td>
<td>0.434 [0.088]</td>
</tr>
<tr>
<td>SNPs</td>
<td>0.035</td>
<td>0.33</td>
<td>2.634</td>
<td>0.139</td>
<td>1 [0.89, 1]</td>
<td>0.132 [0.014]</td>
</tr>
</tbody>
</table>
Sensitivity to the number of markers

Sampling subsets of loci from an empirical genetic dataset and estimation of a statistic of interest based on these subsets can give insights into the power provided by a given marker panel (Miller et al. 2013; Hoffman et al. 2014; Stoffel et al. 2015). However, although subsampling markers (with replacement) from an empirical dataset allows exploration of trends in the magnitude of a statistic, the precision (variation) of the same statistic will be biased. This is due to the increasing non-independence of resampled marker sets as they approach the total number of markers. For example, given a dataset of 20 genetic markers, repeatedly subsampling 18 markers and calculating $g^2$ will always lead to lower variation in the estimates than subsampling sets of 5 markers. To circumvent this problem, the simulate_g2() function simulates genotypes from which subsets of loci can be sampled independently. The simulations can be used to evaluate the effects of the number of individuals and loci on the precision and magnitude of $g^2$. The user specifies the number of simulated individuals (n_ind), the subsets of loci (subsets) to be drawn, the heterozygosity of non-inbred individuals (H_nonInb, i.e. expected heterozygosity in the base population) and the distribution of $f$ among the simulated individuals. The $f$ values of the simulated individuals are sampled randomly from a beta distribution with mean (meanF) and variance (varF) specified by the user (e.g. as in Wang 2011). This enables the simulation to mimic populations with known inbreeding characteristics or to simulate hypothetical scenarios of interest. For computational simplicity, allele frequencies are assumed to be constant across loci and the simulated loci are unlinked. Genotypes (i.e. heterozygosity/homozygosity at each locus) are assigned stochastically based on the $f$ values of the simulated individuals. Specifically, the probability of an individual being heterozygous at any given locus ($H$) is expressed as $H = H_0(1-f)$, where $H_0$ is the user-specified heterozygosity of a non-inbred individual and $f$ is an individual’s inbreeding coefficient drawn from the beta distribution.

```
sim_g2_mouse_microsats <- simulate_g2(n_ind = 50, 
    H_nonInb = 0.5, meanF = 0.2, varF = 0.03, 
    subsets = c(5, 10, 15, 20, 25, 30, 35, 40, 
    45, 50), reps = 100, type = "microsats")

sim_g2_mouse_snps <- simulate_g2(n_ind = 50, 
    H_nonInb = 0.5, meanF = 0.2, varF = 0.03, 
    subsets = seq(from = 1000, to = 10000, 
    by = 1000), reps = 100, type = "snps")
```

The results can be visualized by showing the mean and CI of $g^2$ plotted against the number of loci used (Fig. 4).

```
par(mfrow = c(1, 2), mar=c(5,15,3,1.2))
plot(sim_g2_mouse_microsats, main = "Microsatellites" 
    cex.axis=1.5, cex.main = 1.5, cex.lab = 1.5)
plot(sim_g2_mouse_snps, main = "SNPs", 
    cex.axis=1.5, cex.main = 1.5, cex.lab = 1.5)
```

Bear in mind that $g^2$ values calculated from the simulated data may overestimate precision due to the assumption of unlinked loci. However, in practice, the number of linked SNPs in most real datasets will be small compared to the number of unlinked SNPs (Szulkin, Bierne & David 2010) and hence $g^2$ should not be substantially affected.

Finally, it is of interest to infer how well genetic marker heterozygosity reflects the inbreeding level $f$ and whether this correlation could be increased by genotyping individuals at a larger set of markers. The simulate_r2_hf() function can be used to compare the precision and magnitude of the expected squared correlation between heterozygosity and inbreeding ($r^2(h,f)$) for a given number of genetic markers.

Fig. 4. Output of the simulate_g2() function. Different sets of microsatellites and SNPs were simulated and stochastically drawn from distributions based on a mean (SD) inbreeding level f of 0.2 (0.03) assuming that a non-inbred individual has a heterozygosity of 0.5. The two plots show the $g^2$ statistics from all samples including their means and 95% CIs.
The results can again be plotted as a series of $r^2(h,f)$ estimates together with their means and CIs (Fig. 5).

\[ \text{sim}_\text{r2}_\text{mouse_microsat} \leftarrow \text{simulate}_\text{r2}_\text{hf}(n\_\text{ind} = 50, \right.
\]
\[ H\_\text{nonlinb} = 0.5, \text{meanP} = 0.2, \text{varP} = 0.03, \]
\[ \text{subsets} = \text{c}(5, 10, 15, 20, 25, 30, 35, 40, 45, 50), \text{reps} = 100, \text{type} = \text{"msats"}) \]

\[ \text{sim}_\text{r2}_\text{mouse_snps} \leftarrow \text{simulate}_\text{r2}_\text{hf}(n\_\text{ind} = 50, \right.
\]
\[ H\_\text{nonlinb} = 0.5, \text{meanP} = 0.2, \text{varP} = 0.03, \]
\[ \text{subsets} = \text{seq}(\text{from} = 1000, \text{to} = 10000, \right.
\]
\[ \text{by} = 1000), \text{reps} = 100, \text{type} = \text{"snps"}) \]

The results can again be plotted as a series of $r^2(h,f)$ estimates together with their means and CIs (Fig. 5).

Increasing the total number of genetic markers should not affect the proportion of linked markers and should thus not affect $g_2$. To test this, we evaluated the sensitivity of $g_2$ to marker number by repeatedly sampling random subsets of between 100 and 13,000 SNPs from the full mouse dataset and calculating the respective $g_2$ values. For each subset, markers were sampled without replacement to avoid non-independence, which is why the number of repetitions decreases with increasing marker number. The mean $g_2$ was found to be stable across all subset sizes, suggesting that, for our dataset, the expected $g_2$ does not vary appreciably with marker density (Fig. 6).

In general, the number of locus pairs in strong linkage is expected to be very low compared to the number of non-linked pairs (Szulkin, Bierne & David 2010). As $g_2$ averages over all pairs of loci, this point estimate should therefore be relatively insensitive to the inclusion of linked markers as long as all markers are broadly distributed.
mates and their confidence intervals. We used the indep-
2007) to remove one SNP from each pair with an
within
and HFCs for empirical data, straightforward simulations
inbreeding depression. In addition to the quantification of ID
estimate the power of a genomic inbreeding measure to detect
ID and HFCs. A good example is a
recent study of red deer, in which Huisman
et al. Kardos, Allendorf & Luikart 2014). A good example is a
above thresholds ranging from 0
combe on selfing or inbreeding rates estimated
linked markers need not be pruned as long as the genomic
distributions of the marker and trait loci are comparable.
However, if the goal of a study is to infer characteristics of
marker and trait loci are comparable. Thus, low MAF SNPs may be disproportionately error
prone when the depth of sequence coverage is not high enough
to capture multiple copies of the minor allele. On the other
hand, filtering out low MAF SNPs may distort the allele fre-
quency spectrum and lead to the loss of valuable information
(Hoffman et al. 2014).

Finally, LD and ID have been seen as alternative
hypotheses to explain HFCs (Hansson & Westerberg 2008).
However, LD often goes hand in hand with ID and is
therefore a relevant variance component when the aim is to
estimate \( \sigma^2(\text{IBD}_C) \) (Bierne, Tsitrone & David 2000;
Szulkin, Bierne & David 2010). As most HFC studies
should be interested in estimating \( \sigma^2(\text{IBD}_C) \) through \( g_2 \),
linked markers need not be pruned as long as the genomic
distributions of the marker and trait loci are comparable.
However, if the goal of a study is to infer characteristics of
a pedigree from \( g_2 \) (such as self-fertilization rates), it might
be useful to reduce physical linkage among markers using
PLINK (Purcell et al. 2007) or other methods to ensure
their independence (David et al. 2007). Further investiga-
tion would be needed to evaluate the impact of pruning
linked markers on selfing or inbreeding rates estimated
through \( g_2 \).

### Computation times

Computation times will be negligible for most microsatellite
datasets but somewhat longer for very large SNP datasets.
On a standard laptop (Intel Core i5 2.60 GHz, 8 GB
RAM), running the \texttt{g2_snps()} function for our exam-
ple SNP dataset (36 individuals genotyped at 13 198 loci)
with 1000 bootstraps takes 1 min 12 s without paralleliza-
and 38 s with parallelization on three cores. For com-
parison, we also simulated a large SNP dataset with 3500
individuals at 37 000 loci (similar to Huisman et al. 2016)
and ran this on a 40 core server with 1000 bootstraps,
which took 73 h.

### Availability

The current stable version of the package requires R 3.2.1 and
can be downloaded from CRAN as follows:

```
install.packages("inbreedR")
```

In the future, we will aim to extend the functionality of
\texttt{inbreedR} and the latest development version can be
downloaded from GitHub.

---

**Fig. 7.** Estimates of \( g_2 \) with confidence intervals for subsets of SNPs
pruned based on different LD thresholds. We used PLINK to remove
one SNP from each marker pair with an \( r^2 \) above the respective
threshold. As we used a sliding window spanning the full dataset instead of
local regions on a chromosome, the retained datasets contained a maxi-
mum of 4363 (\( r^2 > 0.99 \)) and a minimum of 1095 (\( r^2 > 0.5 \)) SNPs. The
magnitude and precision of \( g_2 \) does not vary noticeably for our dataset
when pruning strongly linked SNPs.

**Final remarks**

The \texttt{inbreedR} package implements a framework to esti-
mate the impact of variation in inbreeding on marker heterozy-
gosity and fitness, which has been suggested to be routinely
reported in HFC studies (Szulkin, Bierne & David 2010;
Kardos, Allendorf & Luikart 2014). A good example is a
recent study of red deer, in which Huisman et al. (2016) quan-
tified identity disequilibria through \( g_2 \) in several datasets to
estimate the power of a genomic inbreeding measure to detect
inbreeding depression. In addition to the quantification of ID
and HFCs for empirical data, straightforward simulations
within \texttt{inbreedR} provide a way to explore the effect of the
number of genetic markers on \( g_2 \) and the expected correlation
between marker heterozygosity and inbreeding. This is impor-
tant for evaluating the power of a given dataset to measure
inbreeding depression and could also facilitate the planning of
future projects by allow exploration of the effects of sample size
and marker number on the power to detect ID and HFCs.

Although \( g_2 \) and related parameters can provide insights
into whether an HFC is due to inbreeding or not, the user
should be aware that spurious HFCs can occur due to popul-a-
tion structure (Slate et al. 2004), which should therefore be
appropriately dealt with beforehand. For instance, genetically
distinct populations could be analysed separately. Also, it is
worthwhile considering whether SNPs should be filtered based
on their minor allele frequencies (MAF) prior to analysis. One
the one hand, genotyping by sequencing approaches rely on
sufficient depth of coverage to call SNPs with reasonable con-
dence. Thus, low MAF SNPs may be disproportionately error
prone when the depth of sequence coverage is not high enough
to capture multiple copies of the minor allele. On the other
hand, filtering out low MAF SNPs may distort the allele fre-
quency spectrum and lead to the loss of valuable information
(Hoffman et al. 2014).

Finally, LD and ID have been seen as alternative
hypotheses to explain HFCs (Hansson & Westerberg 2008).
However, LD often goes hand in hand with ID and is
therefore a relevant variance component when the aim is to
estimate \( \sigma^2(\text{IBD}_C) \) (Bierne, Tsitrone & David 2000;
Szulkin, Bierne & David 2010). As most HFC studies
should be interested in estimating \( \sigma^2(\text{IBD}_C) \) through \( g_2 \),
linked markers need not be pruned as long as the genomic
distributions of the marker and trait loci are comparable.
However, if the goal of a study is to infer characteristics of
a pedigree from \( g_2 \) (such as self-fertilization rates), it might
be useful to reduce physical linkage among markers using
PLINK (Purcell et al. 2007) or other methods to ensure
their independence (David et al. 2007). Further investiga-
tion would be needed to evaluate the impact of pruning
linked markers on selfing or inbreeding rates estimated
through \( g_2 \).

---

**Fig. 7.** Estimates of \( g_2 \) with confidence intervals for subsets of SNPs
pruned based on different LD thresholds. We used PLINK to remove
one SNP from each marker pair with an \( r^2 \) above the respective
threshold. As we used a sliding window spanning the full dataset instead of
local regions on a chromosome, the retained datasets contained a maxi-
mum of 4363 (\( r^2 > 0.99 \)) and a minimum of 1095 (\( r^2 > 0.5 \)) SNPs. The
magnitude and precision of \( g_2 \) does not vary noticeably for our dataset
when pruning strongly linked SNPs.
install.packages("devtools")
devtools::install_github("mastoffel/inbreedR")

Data accessibility
Both example datasets are included in the R package.

References