Title:

Friends and Family: A software program for identification of unrelated individuals from molecular marker data

Authors:

Deon de Jager,* Petrus Swarts,* Cindy Harper,‡ Paulette Bloomer*

*Molecular Ecology and Evolution Programme, Department of Genetics, University of Pretoria, cnr Lynnwood Road and Roper Street, Hatfield, South Africa 0083

‡Veterinary Genetics Laboratory, Faculty of Veterinary Science, University of Pretoria, Soutpan Road, Onderstepoort, South Africa 0110

Keywords: Breeding programmes, conservation, Hardy-Weinberg equilibrium, kinship, relatedness, translocation

Corresponding author:

Deon de Jager

Room 7-34, Agricultural Sciences Building, University of Pretoria, cnr Lynnwood Road and Roper Street, Hatfield, South Africa 0083

Fax: +27 12 362 5327

Email: dejager4@gmail.com

Running title:

Friends and Family

Abstract

The identification of related and unrelated individuals from molecular marker data is often difficult, particularly when no pedigree information is available and the data set is large. High levels of relatedness or inbreeding can influence genotype frequencies and thus genetic marker evaluation, as well as the accurate inference of hidden genetic structure. Identification of related and unrelated individuals is also important in breeding programmes, to inform decisions about breeding pairs and translocations. We present Friends and Family, a Windows executable program with a graphical user interface that identifies unrelated individuals from a pairwise relatedness matrix or table generated in programs such as COANCESTRY and GenAlEx. Friends and Family outputs a list of samples that are all unrelated to each other, based on a user-defined relatedness cut-off value. This unrelated data set can be used in downstream analyses, such as marker evaluation or inference of genetic structure. The results can be compared to that of the full data set to determine the effect related individuals have on the analyses. We demonstrate one of the applications of the program: how the removal of related individuals altered the Hardy-Weinberg equilibrium test outcome for microsatellite markers in an empirical data set. Friends and Family can be obtained from https://github.com/DeondeJager/Friends-and-Family.

Introduction

Global human expansion and climate change has led to increased habitat fragmentation, resulting in reduced geographic range and decreased population sizes for many species (Ceballos *et al.* 2015; Di Marco & Santini 2015). Consequently, species have to be managed either within fragmented natural populations, fragmented re-introduced populations in small reserves, or in captivity, with the ultimate goal of the latter being re-introduction into the wild (Robert 2009). However, captive populations of species are prone to genetic drift, inbreeding and altered selection pressures, leading to a reduction in genetic diversity and reduced ability to integrate back into the wild (Fischer & Lindenmayer 2000; Robert 2009; Willoughby *et al.*

2015). Fragmented natural populations, and re-introduced populations in small reserves, are also susceptible to these genetic processes, albeit likely to a lesser extent than captive populations. Consequently, maintaining genetic diversity is vital to the sustainability of managed and captive populations, the success of re-introducing individuals into the wild and the overall conservation of a species. Willoughby *et al.* (2015) showed that the most effective way to limit loss of genetic diversity in captive populations is through a mean kinship breeding protocol – selecting the most unrelated individuals to breed with one another. This approach resulted in the slowest loss of genetic diversity, compared to random mating or selecting breeding individuals based on docility, a favourable behavioural trait in captive animals, but unfavourable in the wild (McPhee 2004). The study by Willoughby *et al.* (2015) illustrated the importance of implementing strategic breeding protocols in captive and managed populations to minimise both the loss of genetic diversity and the effects of selection, thereby increasing the chance of successful re-introduction and reducing the extinction risk of species.

Generally, researchers use neutral, independent genetic markers to study the structure and genetic diversity of populations. Two major criteria are that the markers should be unlinked (i.e. not show significant linkage disequilibrium) and in Hardy-Weinberg equilibrium (HWE) (Weir 1996). During initial marker development, a random sample from a population is used to test the markers. However, it is not always possible to employ a robust sampling strategy, particularly for endangered species, for species that are difficult to access or elusive, or when sampling costs are prohibitive. This could result in small sample sizes or biased sampling (from a single or only a few sampling locations). Consequently, a high proportion of related individuals could be sampled, particularly in species characterized by strong family structure.

For example, African Cape buffalo (*Syncerus caffer caffer*) are mammals that usually live in social herds of between 50 and 500 animals, consisting of smaller social groups of related females, or clans where a few males may also be present (Nowak 1999; Kingdon 2015). There is a strong hierarchical structure within herds and clans, with the dominant male getting priority during the mating season (Kingdon 2015). This leads to a relatively high number of siblings, half-sibs and related individuals in a herd. The family structure within the herd increases the chances of sampling related individuals, which could influence marker evaluation and other analyses, such as genetic structure (Rodriguez-Ramilo & Wang 2012). In such cases it is difficult to know whether the genetic markers developed would be in HWE if not for the unintentional biased sampling. Removing highly related or inbred individuals from the data set would allow researchers to mimic a random sample, albeit at the cost of a reduced sample size.

A data set containing close relatives, or that has high inbreeding levels, could influence the analysis of hidden genetic structure within populations when using unsupervised Bayesian clustering algorithms, as Rodriguez-Ramilo & Wang (2012) showed when using the software STRUCTURE. This is due to the software assuming that the population is in HWE with no relatives or inbreeding (Pritchard *et al.* 2000). In a simulated data set, the software consistently over-estimated the number of clusters (*K*) present in a population, as family structure was interpreted as population structure (Rodriguez-Ramilo & Wang 2012). However, when close relatives were removed using the program COLONY, the accuracy of STRUCTURE to estimate the correct *K* improved. Similar results were obtained with an empirical data set, not only with STRUCTURE, but also with other unsupervised Bayesian clustering software InStruct, BAPS and Structurama (Rodriguez-Ramilo & Wang 2012).

In the study by Rodriguez-Ramilo & Wang (2012), the authors used the program COLONY to identify related individuals and removed them from the data set. This program infers parentage and sibship by constructing pedigrees from molecular marker data and determining the likelihood of each pedigree (Jones & Wang 2010). COLONY allows the user to input parameters and data such as life history traits, reproductive strategies, ploidy, type of marker, genotyping error rates and more. However, it also requires knowledge of which samples are offspring and which are potential parents, which is not always available, especially in wild populations. In such cases, one would need to manually identify and remove related individuals from the data set, which can become difficult to execute accurately when sample sizes are large.

Furthermore, Puechmaille (2016) showed that STRUCTURE frequently incorrectly estimates the number of subpopulations when sampling across locations is uneven. However, equalising sample numbers across sampling localities greatly increased the ability of STRUCTURE to estimate the correct *K* (Puechmaille 2016).

Recently, Waples & Anderson (2017) evaluated different methods of identifying and removing related individuals from simulated and empirical data sets. This was accomplished by determining the precision of estimates such as allele frequency, population differentiation (F_{ST}) and effective population size before and after relatives were removed. In general, the authors found that removing siblings increased the root-mean-squared-error (i.e. reduced precision) of allele frequency calculations and F_{ST} estimations (Waples & Anderson 2017). Removing all siblings increased error in F_{ST} estimation by 10-fold, but removing only some siblings resulted in a reduction in error in cases of mixed and monogamous mating models (Waples & Anderson 2017). The authors suggest that removing relatives from a data set can be useful (i.e. when testing for population structure), but show that it is not always the best approach. Therefore, researchers should critically evaluate their results before and after relatives have been removed.

Here we describe the software program Friends and Family, developed to identify unrelated individuals from a table or matrix of pairwise relatedness values. The data set of unrelated individuals obtained through Friends and Family is intended to mimic a randomly mating population as closely as possible and can thus be used in genetic marker evaluation and to investigate hidden population structure without the confounding effects of family structure or inbreeding. It is critical that unbiased relatedness estimates are obtained, particularly in initial marker evaluation; otherwise more bias is introduced into the outcome (Wang 2011b). Friends and Family can also be used to assist in management of captive bred populations and implementation of breeding protocols by identifying unrelated individuals when no pedigree data are available.

In this paper, we first describe the functionality of Friends and Family, the input and output files and how the program performs the relatedness-based sorting. Thereafter, we use the program to remove relatives from an empirical microsatellite data set and test the effect this has on the HWE of the loci. Finally, we discuss the results and other potential applications of Friends and Family, followed by the conclusion.

Functionality

Friends and Family requires as input, a lower triangular matrix (hereafter referred to as a tri-matrix) or table of pairwise relatedness values generated from codominant genetic marker data, in a comma separated values (CSV) file. When a table is used as input, the program automatically converts the table to a tri-matrix and performs the analysis using this tri-matrix. The program uses several rounds of sorting to identify and remove related individuals from the tri-matrix based on a user-defined relatedness cut-off value. The cut-off value chosen should be based on the subject species: whether haploid or diploid, the mode of reproduction, life history traits, as well as the method used to estimate relatedness. It is important to note that Friends and Family is dependent on the relatedness estimator used and thus cannot distinguish between close inbreeding (breeding between closely related individuals) and pervasive inbreeding (due to genetic drift, small population size or population structure) (Wang 2011b), any more than the applied estimator can. Consequently, when referring to a “relative”, “family” or “related individuals/ samples”, we are referring to a sample that has a pairwise relatedness value that is above the given cut-off value. It is not necessarily a reference to a family member in the traditional sense.

In each round of sorting, the sample with the fewest number of relatives (and consequently the largest number of unrelated individuals, or “friends”), is identified and used as the reference sample. All samples with a pairwise relatedness value equal to or greater than the cut-off value are labelled as “family” and removed from the tri-matrix. If, at any point, there are two or more samples with the same number of “friends”, their average relatedness to all the remaining samples in the tri-matrix is calculated. The sample with the lowest average relatedness is then used as the reference sample for that round. The sorting ends when there are no more pairwise relatedness values that are equal to or exceed the given cut-off value. The program outputs three CSV files in a user-specified folder. The first is a “friends” file containing a list of all the unrelated samples. The second is a “family” file containing the samples that were removed in each round. The third is a “wa3b2” file that is generated by the program as a step during analysis that contains a square matrix of the pairwise relatedness input data.

Furthermore, the user can select the option to “output binary files”. These CSV files show, for each round, the status of each sample (friend or not), the number of relatives each sample has and their identities, as well as the average relatedness values of the samples with the fewest number of relatives if this calculation was required.

Testing

To confirm that Friends and Family was performing the sorting correctly, the results were compared to that of a manual analysis. A sample of 17 African Cape buffalo from a single herd, genotyped at 11 independent, autosomal microsatellite loci was used as the test data set. Pairwise relatedness was estimated using the Wang (2002) method in COANCESTRY (Wang 2011a). This method was chosen, as it is more robust than other estimators when unknown relatives are present in the sample, as would be the case when sampling a herd of African Cape buffalo, when the sample sizes are small and when the allele frequency of the population is unknown, which is more often the case than not (Wang 2002). Unrelated individuals were identified using Friends and Family, at cut-off values of -0.1, 0, 0.1, 0.25, 0.5, 0.75 and 1. This particular data set did not contain identical individuals and thus no pairwise relatedness values of 1 were present. Therefore, some pairwise values were manually changed to 1, to ensure that Friends and Family performed correctly when this value was present. The identities of related and unrelated individuals and the calculations of average relatedness, when applicable, at each cut-off value were compared to that of the manual analysis, which confirmed that the program performed as expected.

Application: Effect of removing relatives on Hardy-Weinberg equilibrium

To investigate the effect the removal of related individuals had on HWE, we analysed HWE of 11 microsatellite markers in 16 geographically isolated herds of African Cape buffalo before and after relatives were removed, in both a qualitative and quantitative manner. Herds one to four are from various national parks in southern Africa and herds five to 16 are from privately owned game ranches.

Qualitative analysis

Hardy-Weinberg exact tests were carried out in Genepop version 4.3 (Guo & Thompson 1992; Raymond & Rousset 1995; Rousset 2008) using the Markov chain method with 10 000 dememorizations, 1000 batches and 5000 iterations per batch. Complete enumeration of loci was carried out when applicable. Fisher's method for combining independent *P*-values across herds (sampling localities) was used to determine the significance of the results. The significance level for each test was Bonferroni-adjusted for multiple testing (Rice 1989).

To remove related individuals from each herd, pairwise relatedness was estimated using the Wang (2002) method in COANCESTRY. This method provides more accurate relatedness estimates, compared to other estimators such as maximum-likelihood (Thompson 1975; Thompson 1991; Milligan 2003), Queller & Goodnight (1989), Ritland (1996) and Lynch & Ritland (1999). This is especially true when allele frequencies are calculated relative to subpopulations, as opposed to the whole population, and when structure between subpopulations is minimised (Wang 2011b). Therefore, we pooled herds into four genetically similar clusters, based on factorial correspondence analysis [implemented in Genetix (Belkhir *et al.* 2004)], STRUCTURE analysis (Pritchard *et al.* 2000) and genetic distance [D_{EST}, (Jost 2008)] (results not shown). The intra-herd pairwise relatedness values for each herd was extracted from the COANCESTRY output and used as input for Friends and Family. Unrelated individuals from each herd were identified in Friends and Family, at a relatedness cut-off value of 0.25. This value was chosen as a compromise between extracting unrelated individuals and maintaining an appropriate sample size for genetic analyses for the majority of the herds. The most unrelated samples, as identified by Friends and Family, were manually extracted from the genotype data set of each herd, which effectively is the same as removing related individuals. These data sets were used to once again evaluate HWE at each locus, implemented in Genepop as before. The results of the HWE tests were compared before and after related individuals were removed (Table 1). The average relatedness (median), as well as the maximum, minimum and range for intra-herd relatedness was compared before and after relatives were removed to ensure relatedness was in fact reduced (Fig. 1).

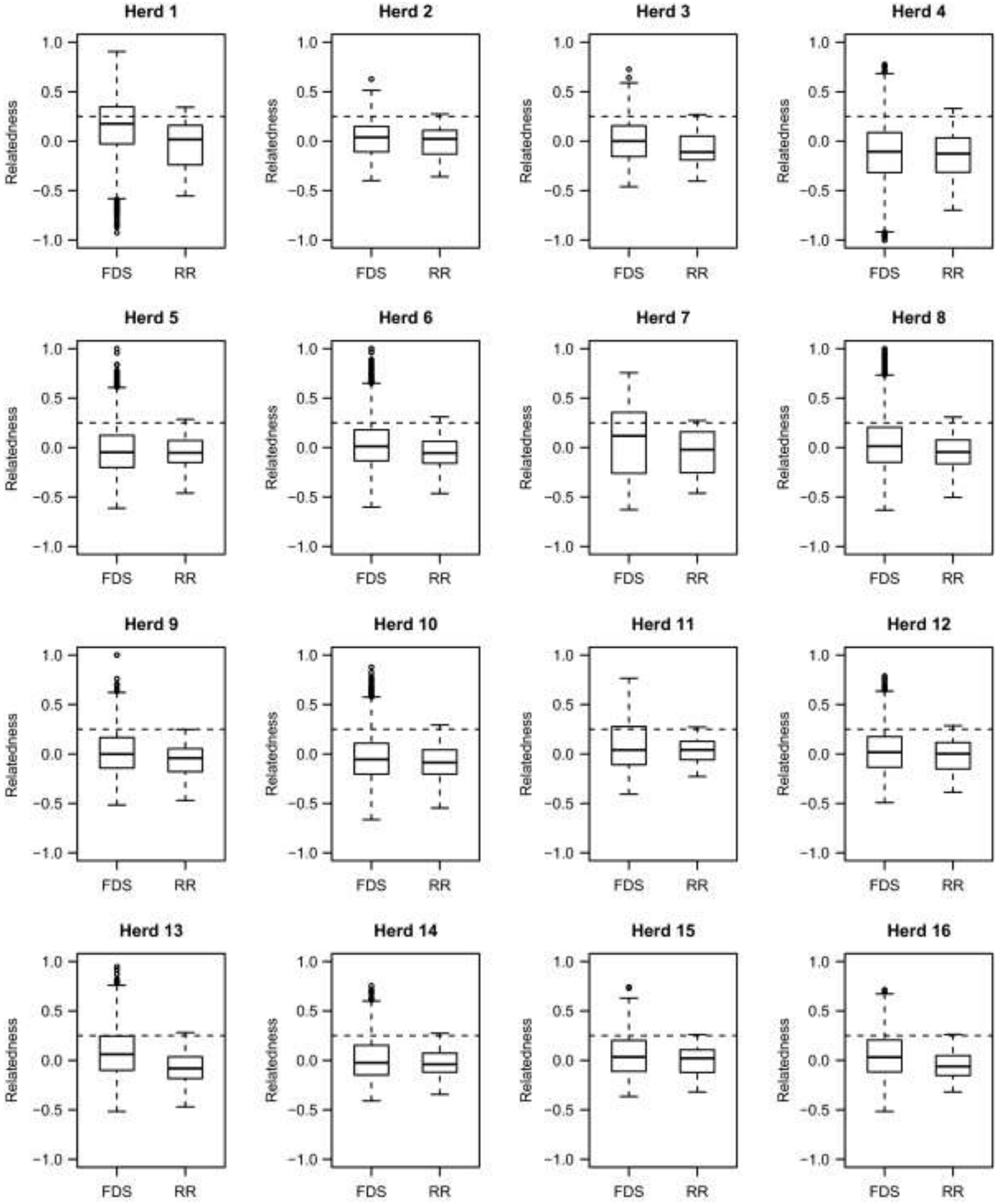


Figure 1. Boxplots showing the distribution of relatedness values of each herd before (full data set FDS) and after relatives were removed (RR). The horizontal dashed lines mark where relatedness is equal to 0.25. Open circles show outliers

Table 1 Change in Hardy-Weinberg equilibrium test statistic of 11 microsatellite markers after relatives were removed from each herd. Hyphen (-): Marker remained in HWE. X: Marker remained deviating from HWE. Tick (√): Marker changed to in HWE. Exclamation point (!): Marker changed to deviating from HWE. Question mark (?): HWE of marker could not be determined after relatives were removed, due to lack of diversity (as a result of small sample size).

Herd	BM1824	BM3205	BM3517	BM719	CSSM19	ETH10	ILSTS026	INRA006	SPS115	TGLA227	TGLA263
1	-	-	-	-	!	-	-	-	-	X	-
2	-	-	-	-	-	-	-	-	-	-	-
3	-	√	-	-	-	-	√	-	-	-	-
4	-	-	-	!	!	-	-	-	-	√	!
5	-	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	√	√	-
7	X	√	-	-	-	-	-	-	-	√	-
8	√	-	-	-	-	-	X	-	-	X	-
9	-	-	-	-	-	-	-	-	-	-	-
10	√	-	√	!	-	-	X	√	-	X	X
11	-	-	-	-	-	-	-	-	-	?	-
12	-	-	-	-	-	-	-	-	-	X	-
13	-	-	-	-	-	-	-	-	-	√	-
14	-	-	-	-	-	-	-	-	-	√	-
15	-	-	-	-	-	-	-	-	-	X	-
16	-	-	-	-	-	-	-	-	-	-	-
All	√	**X**	**-**	**-**	**-**	**-**	**X**	**-**	**-**	**X**	√

After the removal of relatives from each herd there were 15 instances where a marker was not in HWE with the full data set, but was in HWE with the reduced data set (Table 1). There were 12 instances where a marker was not in HWE and remained that way after relatives were removed (Table 1). There were five cases where a marker was in HWE, but changed to not being in HWE after the removal of relatives (Table 1). There was one instance where the HWE could not be determined by Genepop, after relatives had been removed (TGLA227, Herd 11, Table 1), as there were only two alleles present at this locus, one of which was represented by only one copy in the reduced data set. The majority of the observations (154) were markers that remained in HWE after relatives were removed (Table 1). The *P*-value of the test for HWE of each marker in each herd, before and after relatives were removed, is given in Tables S1 and S2, Supporting information. Sample size after the removal of relatives was reduced by 38 – 86% (mean = 68%). The fraction of samples retained after Friends and Family analysis is a function of the cut-off value (lower cut-off values are more stringent) and the relatedness and inbreeding levels within the data set.

The herds in the reduced data set all had lower median, as well as lower maximum relatedness values compared to the full data set, except for herd 11, where the median remained unchanged (Fig. 1). The maximum relatedness value in each herd was slightly higher than the cut-off value used in Friends and Family (0.25), except in herd 9 where it was equal to 0.25, which is most likely due to the reduction in sample size and subsequent re-estimation of relatedness.

Quantitative analysis

In order to quantitatively asses HWE, we used the inbreeding coefficient, F_{IS}, as a disequilibrium measure (Weir 1996). This approach has previously been shown to be a reliable measure of HWE when evaluating the effect of missing data on loci (Graffelman *et al.* 2014; Graffelman *et al.* 2015). The inbreeding coefficient of each locus in each population was calculated using the divBasic function in the R (R Core Team 2015) package diveRsity, with 1000 bootstrap iterations to obtain 95% confidence intervals (Keenan *et al.* 2013).

The F_{IS} values of the full data set were compared to two data sets where relatives had been removed (Table 2; Fig. S1 and S2, Supporting information). The first data set was the one used in the qualitative approach where relatives were removed using a cut-off value of 0.25 in Friends and Family. The second data set was one where relatives were removed using a cut-off value of 0.5. The second data set was included as Waples & Anderson (2017) showed that less stringent removal of siblings gives more reliable results compared to more stringent removal of siblings in allele frequency and F_{ST} analyses. Choosing a cut-off value of 0.5 in Friends and Family is a less stringent removal of relatives compared to 0.25. It has the effect of including more than one relative per family group in the final data set, since sibling relatedness theoretically ranges from 0.25 to 0.75. Therefore, in this case, using a cut-off of 0.25 equates to removing all relatives of a family group (except one) and using a cut-off of 0.5 equates to removing only some of the relatives in a family group. A locus was considered as deviating from HWE when the F_{IS} value differed significantly from zero, based on the 95% confidence intervals (Fig. S1 and S2, Supporting information).

In both data sets, the majority of loci did not deviate from HWE before and after removal of relatives (Table 2). In cases where loci no longer deviated from HWE after relatives had been removed, the F_{IS} values of only 7/12 cases moved towards zero in the 0.25 data set (compared to the full data set), whereas 8/9 cases moved towards zero in the 0.5 data set (Fig. S1 and S2, Supporting information). It is expected that the F_{IS} would move towards zero, as highly related individuals were removed from the data sets. There were also substantially fewer cases (4 vs 12) in the 0.5 data set where loci deviated from HWE where previously there was no deviation, compared to the 0.25 data set (Table 2).

Table 2 Summary of F_{IS} disequilibrium analysis.

HWE Change	Removing All Relatives (cut-off = 0.25)	Removing Some Relatives (cut-off = 0.5)
Deviating from HWE → HWE	12	9
HWE → Deviating from HWE	12	4
Deviating from HWE → Deviating from HWE	17	20
HWE → HWE	135	143
Total	176	176

Statistical power analysis

Sample size was substantially reduced in the data sets where relatives were removed, particularly in the 0.25 data set. Therefore, we tested the power of the loci in the full- and 0.25 data sets to detect varying levels of population differentiation (F_{ST}) using the program POWSIM (Ryman & Palm 2006). We used all the herds combined as the base population, so that all alleles would be represented. The observed allele frequencies, number of loci, number of populations (herds) and sample sizes were used as input for the simulations, except where sample sizes were set equal to the lowest sample size in the data set. We tested the power of the loci to detect F_{ST} levels of approximately 0.07 (the average F_{ST} of the herds), 0.01 and 0.001. To obtain these F_{ST} levels, the effective population size was set to 1000, with 150, 20 and 2 generations of drift for each F_{ST} respectively. Fisher's exact test (using the default iteration factors: 1000 dememorizations, 100 batches, 1000 iterations per batch) was used as the statistical test to detect power and the simulations were replicated 1000 times.

Table 3 shows both the full data set and the reduced data set retain enough power to detect F_{ST} levels of 0.07 and 0.01 when samples sizes are unequal and equal to the smallest sample size. However, it is evident that substantial loss of power occurs when trying to detect an F_{ST} of 0.001 in the 0.25 data set with unequal sample sizes, as well as when sample sizes are equal in both the full data set and the 0.25 data set (Table 3).

We further tested the power of the markers to differentiate between unrelated individuals in each herd before and after removal of relatives and between siblings in the full data sets by calculating the combined non-exclusion probabilities of identity and sib-identity in the program Cervus (Kalinowski *et al.* 2007). The ability of a set of markers to distinguish between individuals (and especially siblings) is particularly important when relatedness estimations are to be carried out, for example in planning translocations or designing

Table 3 F_{ST} power analysis of markers before and after removal of relatives.

	Unequal Sample Sizes			Equal Sample Sizes*		
F_{ST}	**0.0723**	**0.01**	**0.001**	**0.0721**	**0.01**	**0.001**
Full Data Set	1**	1**	0.996**	1**	1**	0.201^{NS}
0.25 Data Set	1**	1**	0.276^{NS}	1**	0.726^{NS}	0.085^{NS}

*n(Full Data Set) = 17; n(0.25 Data Set) = 6.

**For Fisher’s exact test, a confidence level of 0.95 was used.

Table 4 Non-exclusion probabilities of identity and sib-identity for the full data set and the reduced data set (0.25 relatedness cut-off). Sample sizes are given in parentheses.

	Non-exclusion *Pr*(Identity)		**Non-exclusion *Pr*(Sib-identity)**
	Full Data Set	**0.25 Data Set**	**Full Data Set**
Herd 1	6.86E-06 (79)	1.03E-06 (13)	5.22E-03 (79)
Herd 2	2.18E-10 (21)	7.54E-11 (13)	1.59E-04 (21)
Herd 3	8.04E-11 (35)	1.14E-11 (16)	1.64E-04 (35)
Herd 4	3.00E-08 (95)	4.81E-09 (36)	7.47E-04 (95)
Herd 5	2.21E-12 (153)	5.54E-13 (31)	5.00E-05 (153)
Herd 6	1.74E-11 (308)	4.68E-13 (42)	9.12E-05 (308)
Herd 7	9.52E-10 (21)	2.90E-10 (9)	2.08E-04 (21)
Herd 8	3.30E-11 (262)	4.87E-13 (38)	9.24E-05 (262)
Herd 9	4.74E-11 (57)	1.23E-11 (16)	1.53E-04 (57)
Herd 10	9.79E-12 (164)	7.18E-13 (44)	7.43E-05 (164)
Herd 11	1.55E-10 (17)	1.43E-10 (6)	1.41E-04 (17)
Herd 12	1.63E-10 (54)	2.53E-11 (17)	1.57E-04 (54)
Herd 13	2.88E-10 (99)	7.87E-12 (23)	2.07E-04 (99)
Herd 14	5.49E-11 (35)	1.02E-10 (12)	1.35E-04 (35)
Herd 15	1.06E-10 (22)	1.03E-10 (10)	1.22E-04 (22)
Herd 16	2.96E-10 (37)	4.58E-11 (14)	2.12E-04 (37)

breeding programmes. Table 4 shows that the power to distinguish individuals increased in all herds when relatives were removed, except in herd 14. The non-exclusion probability of sib-identity with this set of markers is low enough in all herds to distinguish siblings (Table 4).

Discussion and other potential applications

The program Friends and Family accurately identifies unrelated individuals from pairwise relatedness data at various cut-off values ranging from -0.1 to 1. Related individuals were then removed from the data sets and the resulting reduced data sets had lower median relatedness values, further validating the functionality of the program (Fig. 1). Even though a cut-off value of 0.25 was used to illustrate the reduction in relatedness, the majority of the maximum relatedness values were slightly higher than 0.25 after relatives had been removed (Fig. 1). This is most likely due to sampling error, as a result of the reduction in sample size, leading to an overestimation of pairwise relatedness (Wang 2011b).

As an example of how Friends and Family could be applied during novel marker development or marker evaluation, we compared whether marker loci were in HWE before and after relatives had been removed from a data set in both a qualitative and quantitative manner.

In the qualitative analysis, the observation that a certain marker is not in HWE in the “full data set” (Table 1), but is in HWE in the “relatives removed” data set could be the result of a number of factors (Waples 2015). In this particular case, we suggest the deviation from HWE was a result of biased sampling. A large proportion of the samples were most likely taken from a family within that population, as strong family structures exist in African Cape buffalo herds. Therefore, the data set likely represents a family and not the population as a whole. It is also possible that the management styles of private game ranches affected HWE at some loci, as different management styles have been shown to affect genetic diversity in impala antelope (Grobler & Van der Bank 1994). In other words, non-random mating, whether due to natural family structure or game management practices, could explain deviation from HWE in this case. Removing related individuals from the data set reduces the effect of such confounding factors.

In the quantitative analysis, we used the inbreeding coefficient, F_{IS}, as a disequilibrium measure for deviation from HWE. We found that in the majority of cases loci did not deviate from HWE before and after relatives had been removed. It was expected that F_{IS} would move towards zero after relatives had been removed and this was shown to be the trend in cases where loci deviated from HWE in the full data set but did not deviate in the reduced data set. However, during more stringent removal of relatives (0.25 data set) 5/12 cases showed F_{IS} moving away from zero, whereas with less stringent removal (0.5 data set) only 1/9 cases moved away from zero (Table 2; Fig. S1 and S2, Supporting information). In addition to this, only four cases were observed where a locus deviated from HWE in the 0.5 data set after removal of relatives when it had not deviated in the full data set (Table 2). Whereas, in the 0.25 data set there were 12 such cases. These results indicate that removing all relatives (0.25 cut-off) had a slightly more perturbing effect, as shown by the F_{IS} analysis, whereas removing only some relatives (0.5 cut-off) produced results more consistent with theoretical expectations. A similar effect was observed by Waples & Anderson (2017).

The power analysis (Table 3) was conducted to test whether the particular microsatellite markers used were still powerful enough to detect certain levels of population differentiation (F_{ST}) with the smaller sample sizes. This served as a proxy for the power to detect deviation from Hardy-Weinberg equilibrium, as both calculations rely on the allele frequencies observed in the data set. The results showed that there was no loss in power at moderate F_{ST} values in either data set. Thus, we are confident that the difference in the results between the 0.25 and 0.5 analyses is not an effect of smaller sample sizes in the 0.25 data set. It has been observed that a cut-off of 0.25 does not always have such a perturbing effect on HWE in other species (P. Bloomer, Pers. Obs.). Thus, the results of such an analysis are likely influenced by the life history traits of the species in question. Therefore, we suggest the stringency of removal of relatives be assessed on a case-by-case basis and that researchers evaluate the power of their genetic markers before and after relatives are removed.

Once it is observed that a marker is in fact in HWE after the removal of related individuals from the data set, the researcher should return to the full data set to investigate the causes of deviation from HWE, which could be biological or technical (Waples 2015). Deviation of microsatellite markers from HWE before and after relatives have been removed is more likely a result of technical errors, such as null alleles or large allele dropout, which leads to an excess of observed homozygotes (Van Oosterhout *et al.* 2004; Séré *et al.* 2014). Table 1 shows that marker TGLA227 was not in HWE before and after relatives had been removed in five of 16 herds and across all herds. The estimated null allele frequency at this locus in these herds was between 12 and 21%, potentially explaining these observations (results not shown). This is further supported by the positive F_{IS} values at this locus (Fig. S1 and S2, Supporting information) indicating excess homozygosity, which is expected when null allele frequency is high. These potential errors can be investigated and corrected. If technical errors are not the cause of deviation from HWE, then natural forces, such as selection, might be acting on the locus. In this case removal of the marker from the data set should be seriously considered.

Another application of Friends and Family includes its employment in genetic structure analysis. The program can be used in a similar fashion as COLONY was used by Rodriguez-Ramilo & Wang (2012), particularly when pedigree data are not available, which when working with wild and endangered species is the case more often than not. Even when pedigree data are available, Friends and Family can be used as a quick method to remove related individuals before structure analysis is carried out. Furthermore, the program can be used to equalise sample numbers across locations while simultaneously reducing the average relatedness of the data set during genetic structure analysis in order to improve the accuracy of the analysis (Puechmaille 2016). This would be done by adjusting the cut-off value for each data set until an equal number of samples remain from each location. The result is that the most unrelated samples are kept in the subsample. Therefore, the confounding effects of unequal sampling, inbreeding and family structure during population structure analysis can all be minimised when using Friends and Family.

Finally, Friends and Family can be useful in captive or managed populations of endangered and other species. As shown by Willoughby *et al.* (2015), the most effective breeding protocol in managed populations is one where the least related animals are allowed to breed. In this way the inevitable reduction in genetic diversity and allelic richness of captive and managed populations is slowed to a rate slower than even that of random mating. Friends and Family provides a list of the most unrelated individuals in a data set via the "friends.csv" output file. This is also applicable when moving breeding individuals between populations. Researchers should asses the power of the loci used in these cases to distinguish between unrelated individuals and between siblings in the full data set, as was done here. The sibling distinguishing power of a set of markers is particularly important when estimating relatedness, and it was shown that the marker set used in this study was able to accomplish this (Table 4). In this case power to distinguish between unrelated individuals increased in the data sets with related individuals removed (Table 4). Despite the decrease in sample size, the increase in distinguishing power was not unexpected, as similar genotypes were removed from the data set when relatives were removed, thus making it easier to distinguish individuals. Friends and Family is a useful tool in determining which animals should reproduce, within or between populations, in order to enhance management of endangered species.

Conclusion

Friends and Family identifies unrelated individuals from pairwise relatedness data at a given cut-off value. The program is particularly useful when data sets are large and pairwise relatedness comparisons reach into the thousands. This information allows researchers to extract these unrelated individuals, effectively removing all or some related individuals from their data set, depending on the relatedness cut-off value used. This reduced data set can be used for various applications from marker evaluation, to genetic structure analysis, to informing breeding programmes and translocations of managed or endangered species.

Acknowledgements

We would like to acknowledge Susan Miller for valuable comments on an early version of the manuscript and thank the two anonymous reviewers for their comments and suggestions to improve the manuscript. We thank Amy Clarke for microsatellite genotyping. We acknowledge private landowners, South African National Parks and the Namibian Ministry of Environment and Tourism who supplied buffalo material, the details of which will be published elsewhere. We thank VGL funders for financial support and DdJ acknowledges the University of Pretoria for bursary support. The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

References

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2004) GENETIX 4.05: Population genetics software for Windows^{TM}. *Université de Montpellier II. Montpellier*.

Ceballos G, Ehrlich PR, Barnosky AD*, et al.* (2015) Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances*, **1**, e1400253.

Di Marco M, Santini L (2015) Human pressures predict species' geographic range size better than biological traits. *Global Change Biology*, **21**, 2169-2178.

Fischer J, Lindenmayer DB (2000) An assessment of the published results of animal relocations. *Biological Conservation*, **96**, 1-11.

Graffelman J, Nelson S, Gogarten SM, Weir BS (2015) Exact inference for Hardy-Weinberg proportions with missing genotypes: Single and multiple imputation. *G3: Genes|Genomes|Genetics*, **5**, 2365-2373.

Graffelman J, Sánchez M, Cook S, Moreno V (2014) Statistical inference for Hardy-Weinberg proportions in the presence of missing genotype information. *PLoS ONE*, **8**, e83316.

Grobler JP, Van der Bank FH (1994) Allozyme variation in South African impala populations under different management regimes. *South African Journal of Wildlife Research*, **24**, 89-94.

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361-372.

Jones OR, Wang J (2010) COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, **10**, 551-555.

Jost L (2008) GST and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015-4026.

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program Cervus accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, **16**, 1099-1106.

Keenan K, McGinnity P, Cross TF, Crozier WW, Prodöhl PA (2013) diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution*, **4**, 782-788.

Kingdon J (2015) *The Kingdon field guide to African mammals*. 2nd edn. Bloomsbury Publishing.

Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. *Genetics*, **152**, 1753-1766.

McPhee ME (2004) Generations in captivity increases behavioral variance: Considerations for captive breeding and reintroduction programs. *Biological Conservation*, **115**, 71-77.

Milligan BG (2003) Maximum-likelihood estimation of relatedness. *Genetics*, **163**, 1153-1167.

Nowak RM (1999) *Walker's Mammals of the World.* JHU Press.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.

Puechmaille SJ (2016) The program structure does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, **16**, 608-627.

Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. *Evolution*, **43**, 258-275.

R Core Team (2015) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Raymond M, Rousset F (1995) GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248-249.

Rice WR (1989) Analyzing tables of statistical tests. *Evolution*, **43**, 223-225.

Ritland K (1996) Estimator for pairwise relatedness and indvidual inbreeding coefficients. *Genetical Research*, **67**, 175-185.

Robert A (2009) Captive breeding genetics and reintroduction success. *Biological Conservation*, **142**, 2915-2922.

Rodriguez-Ramilo ST, Wang J (2012) The effect of close relatives on unsupervised Bayesian clustering algorithms in population genetic structure analysis. *Molecular Ecology Resources*, **12**, 873-884.

Rousset F (2008) genepop'007: A complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103-106.

Ryman N, Palm S (2006) POWSIM: A computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology Notes*, **6**, 600-602.

Séré M, Kaboré J, Jamonneau V*, et al.* (2014) Null allele, allelic dropouts or rare sex detection in clonal organisms: Simulations and application to real data sets of pathogenic microbes. *Parasites & Vectors*, **7**, 331-331.

Thompson EA (1975) The estimation of pairwise relationships. *Annals of Human Genetics*, **39**, 173-188.

Thompson EA (1991) 8 Estimation of relationships from genetic data. In: *Handbook of Statistics*, pp. 255-269. Elsevier.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) Micro-checker: Software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535-538.

Wang J (2002) An estimator for pairwise relatedness using molecular markers. *Genetics*, **160**, 1203-1215.

Wang J (2011a) Coancestry: A program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources*, **11**, 141-145.

Wang J (2011b) Unbiased relatedness estimation in structured populations. *Genetics*, **187**, 887-901.

Waples RS (2015) Testing for Hardy–Weinberg proportions: Have we lost the plot? *Journal of Heredity*, **106**, 1-19.

Waples RS, Anderson EC (2017) Purging putative siblings from population genetic data sets: A cautionary view. *Molecular Ecology*, **26**, 1211-1224.

Weir B (1996) *Genetic data analysis II: Methods for discrete population genetic data.* Sinauer Assoc Inc., Sunderland, MA, USA.

Willoughby JR, Fernandez NB, Lamb MC*, et al.* (2015) The impacts of inbreeding, drift and selection on genetic diversity in captive breeding populations. *Molecular Ecology*, **24**, 98-110.

Data Accessibility

Friends and Family is available as Microsoft Windows executable software with a full graphical user interface. The program, README and example input and output files are available at the following URL: https://github.com/DeondeJager/Friends-and-Family. The original genotype data (full data set) is available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.1c2d3.

Author Contributions

DdJ conceived the program and performed the analyses. PS designed the program and wrote the code. DdJ, CH and PB wrote the paper.

Supporting information

TABLE S1 The *p*-value of a test of deviation from Hardy-Weinberg equilibrium of 11 microsatellite loci in 16 African Cape buffalo herds, and across all herds. Sample sizes are given in parentheses

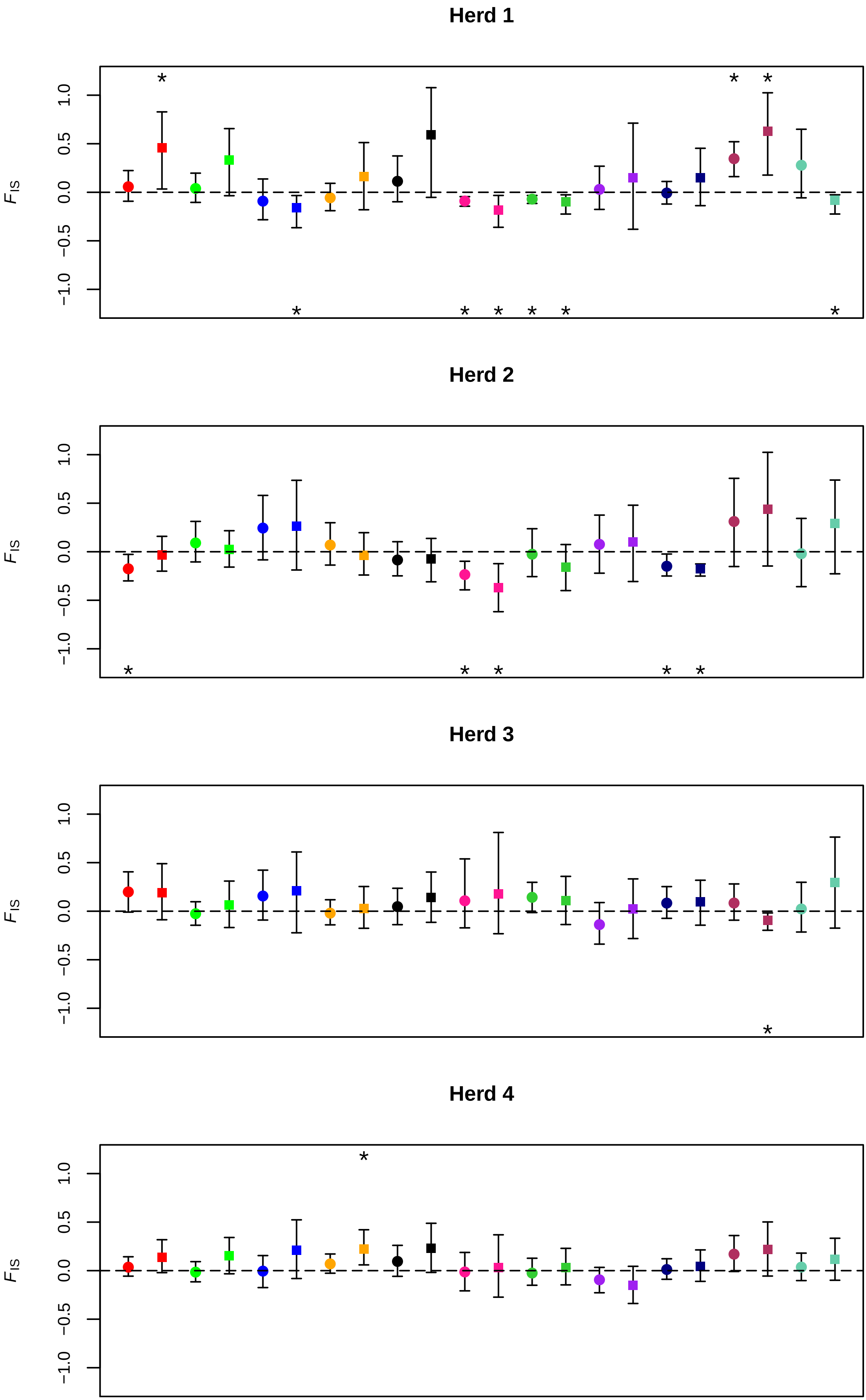
	BM1824	**BM3205**	**BM3517**	**BM719**	**CSSM19**	**ETH10**	**ILSTS026**	**INRA006**	**SPS115**	**TGLA227**	**TGLA263**
Herd 1 (79)	0.27	0.97	0.72	0.13	0.15	1	1	0.82	0.66	0***	0.05
Herd 2 (21)	0.98	0.81	0.2	0.58	0.79	1	0.17	0.77	0.69	0.2	1
Herd 3 (35)	0.09	0.01*	0.2	0.62	0.52	0.44	0.01**	0.35	0.06	0.22	0.69
Herd 4 (95)	0.78	0.08	0.73	0.23	0.09	1	0.55	0.29	0.55	0.02*	0.1
Herd 5 (153)	0.16	0.17	0.13	0.02	0.08	0.7	0.06	0.28	0.83	0.08	0.48
Herd 6 (308)	0.23	0.02	0.43	0.33	0.15	0.1	0.08	0.44	0.02*	0***	0.29
Herd 7 (21)	0.01*	0.01*	0.36	0.05	0.35	1	0.12	0.03	0.62	0.003**	0.25
Herd 8 (262)	0.02*	0.03	0.43	0.03	0.1	0.23	0***	0.34	0.03	0***	0.38
Herd 9 (57)	0.83	0.28	0.02	0.28	0.94	0.73	0.5	0.84	0.26	0.28	0.9
Herd 10 (164)	0.01*	0.01	0.01*	0.02	0.3	0.9	0***	0.01*	0.08	0***	0***
Herd 11 (17)	0.17	0.52	0.13	0.95	0.51	0.42	0.75	1	0.57	1	0.55
Herd 12 (54)	0.75	0.5	0.07	0.72	0.59	0.23	0.29	0.82	0.3	0***	0.13
Herd 13 (99)	0.13	0.05	0.12	0.36	0.6	0.25	0.68	0.12	0.06	0.002**	0.51
Herd 14 (35)	0.04	0.23	0.97	0.65	0.24	0.36	0.38	0.48	0.69	0.02*	0.35
Herd 15 (22)	0.6	0.82	0.75	0.99	0.37	1	1	0.18	0.35	0.01*	0.58
Herd 16 (37)	0.71	0.42	0.37	0.39	0.48	1	0.88	0.67	0.31	1	0.64
All	0.003**	0***	0.02	0.03	0.21	0.95	0***	0.2	0.04	0***	0.02*

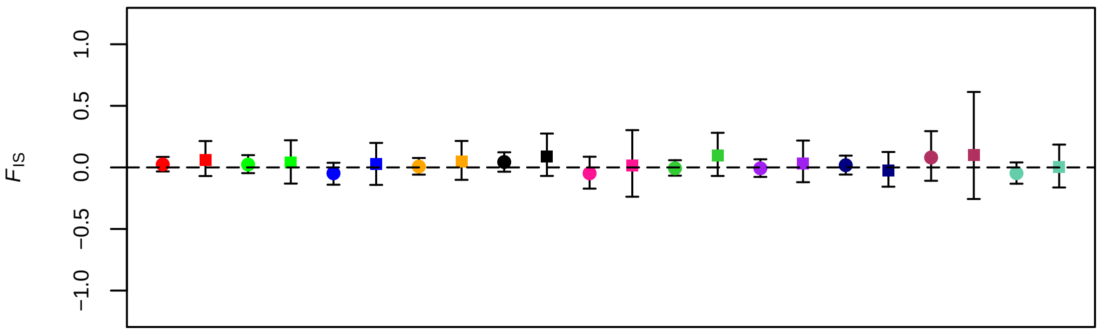
**p*<0.05; ***p*<0.01 ****p*<0.001 significant deviation from Hardy-Weinberg equilibrium after Bonferroni correction.

TABLE S2 The *p*-value of a test of deviation from Hardy-Weinberg equilibrium of 11 microsatellite loci in 16 African Cape buffalo herds, and across all herds, after relatives had been removed from each herd using a relatedness cut-off value of 0.25. Sample sizes are given in parentheses

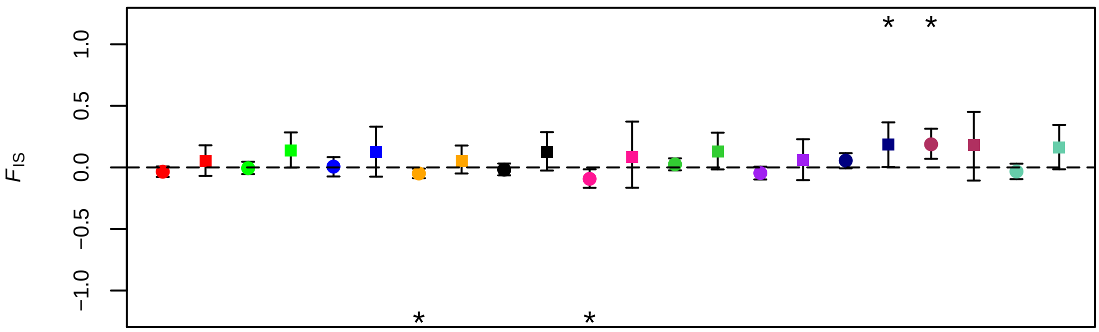
	BM1824	**BM3205**	**BM3517**	**BM719**	**CSSM19**	**ETH10**	**ILSTS026**	**INRA006**	**SPS115**	**TGLA227**	**TGLA263**
Herd 1 (13)	0.05	0.01	1	0.37	0.02*	1	1	0.58	0.02	0.01*	1
Herd 2 (13)	0.73	0.81	0.17	0.85	0.47	0.5	0.22	0.54	0.92	0.08	0.27
Herd 3 (16)	0.31	0.07	0.08	0.49	0.3	0.43	0.03	0.74	0.08	1	0.19
Herd 4 (36)	0.54	0.01	0.1	0.008**	0.02*	1	0.04	0.16	0.51	0.11	0.03*
Herd 5 (31)	0.21	0.17	0.51	0.06	0.18	0.9	0.52	0.91	0.7	0.36	0.94
Herd 6 (42)	0.25	0.04	0.41	0.2	0.16	0.51	0.08	0.54	0.05	0.09	0.08
Herd 7 (9)	0.02*	0.5	1	0.59	0.63	1	0.2	0.06	0.2	0.37	1
Herd 8 (38)	0.71	0.39	0.85	0.61	0.54	0.5	0.01*	0.77	0.19	0***	0.14
Herd 9 (16)	0.83	0.44	0.12	0.53	0.9	0.43	0.56	0.35	0.34	0.56	0.54
Herd 10 (44)	0.2	0.12	0.14	0.02*	0.6	0.64	0.003**	0.18	0.03	0.004*	0.005**
Herd 11 (6)	0.43	0.76	0.34	1	0.92	0.34	0.76	1	1	N/A	1
Herd 12 (17)	0.97	0.59	0.18	1	0.13	0.3	0.06	0.19	0.16	0.003*	0.4
Herd 13 (23)	0.17	0.04	0.09	0.52	0.87	0.13	0.82	0.49	0.25	0.06	0.1
Herd 14 (12)	0.06	0.26	0.73	0.81	0.42	0.25	0.42	0.46	0.32	0.13	0.16
Herd 15 (10)	0.93	0.52	0.75	0.94	0.19	1	1	0.76	0.18	0.01*	0.85
Herd 16 (14)	0.86	0.09	0.14	0.29	0.33	1	0.25	0.21	0.34	1	0.51
All	0.17	0.003**	0.14	0.23	0.1	0.95	0.002**	0.58	0.02	0***	0.06

*p<0.05; **p<0.01 ***p<0.001 significant deviation from Hardy-Weinberg equilibrium after Bonferroni correction.

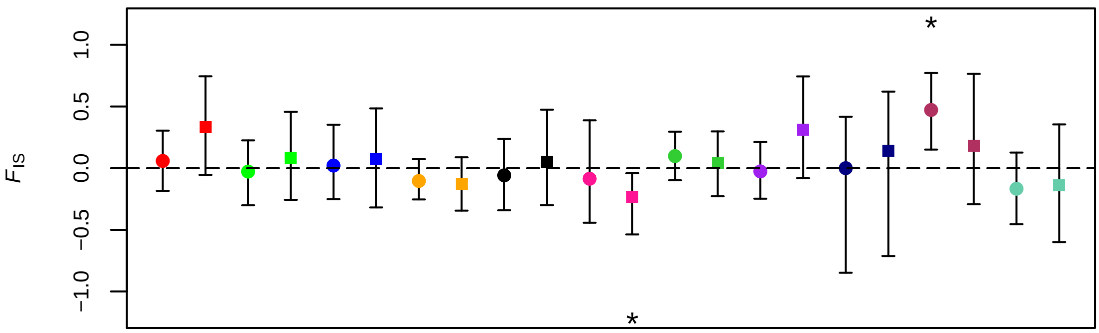

Herd 1
Herd 2
Herd 3
Herd 4
F_{IS}
1.0
0.5
0.0
−0.5
−1.0
*



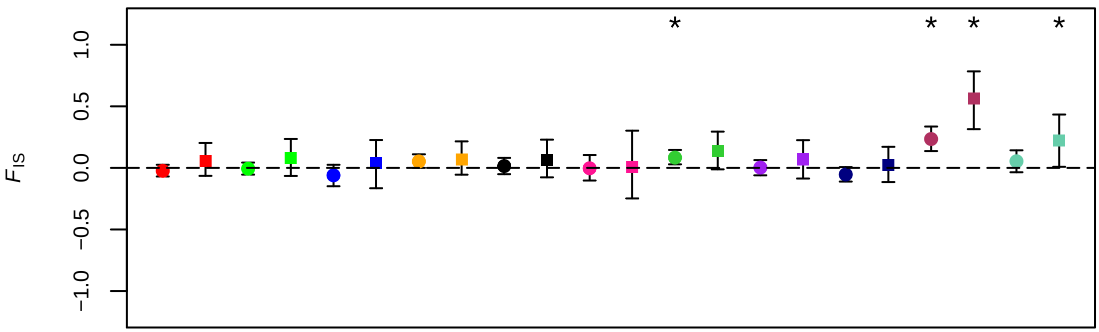
Herd 5



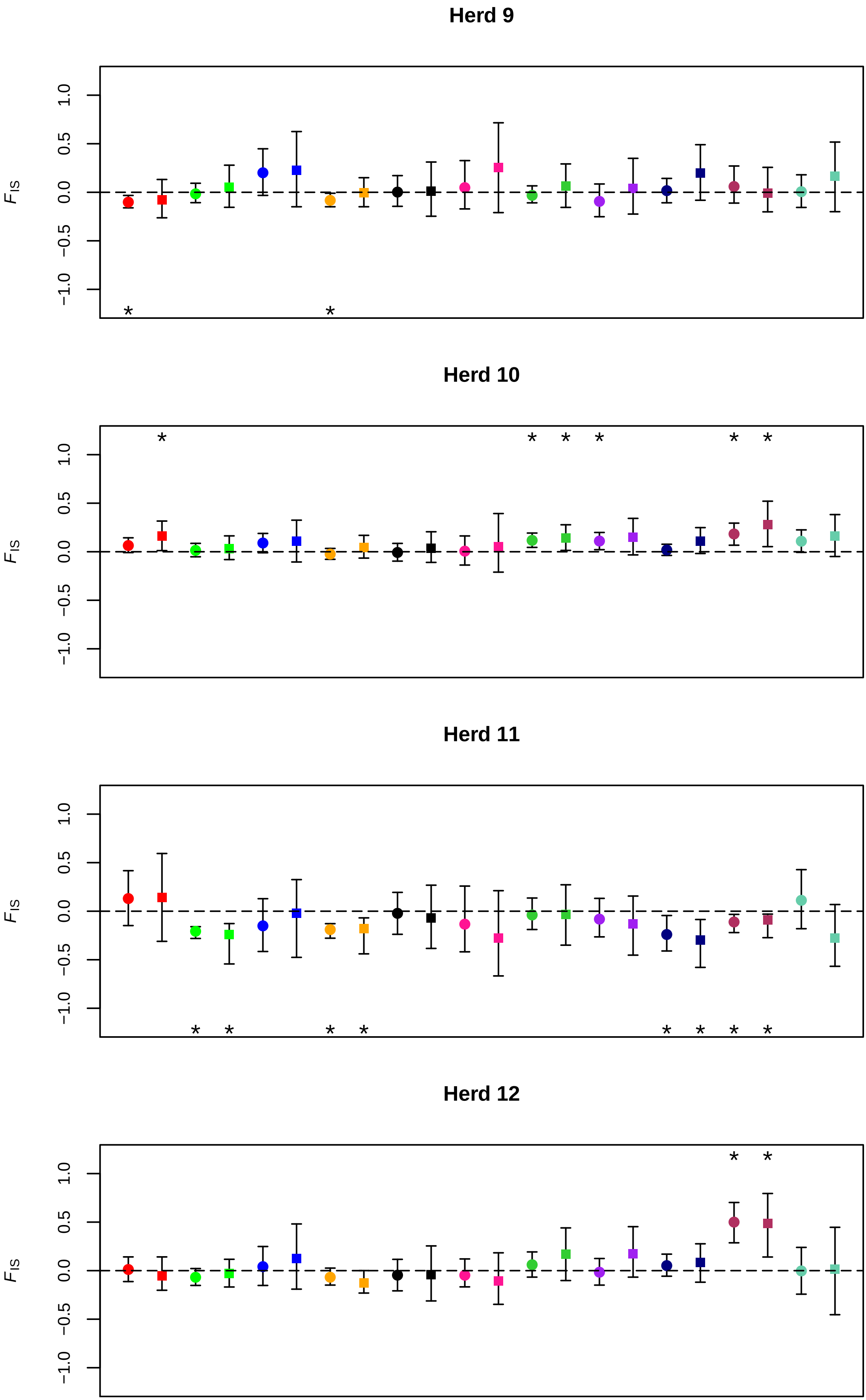
Herd 6



Herd 7



Herd 8


Herd 9
F_IS
1.0
0.5
0.0
−0.5
−1.0
* *
Herd 10
F_IS
1.0
0.5
0.0
−0.5
−1.0
* * * * * *
Herd 11
F_IS
1.0
0.5
0.0
−0.5
−1.0
* * * * * * * *
Herd 12
F_IS
1.0
0.5
0.0
−0.5
−1.0
* *

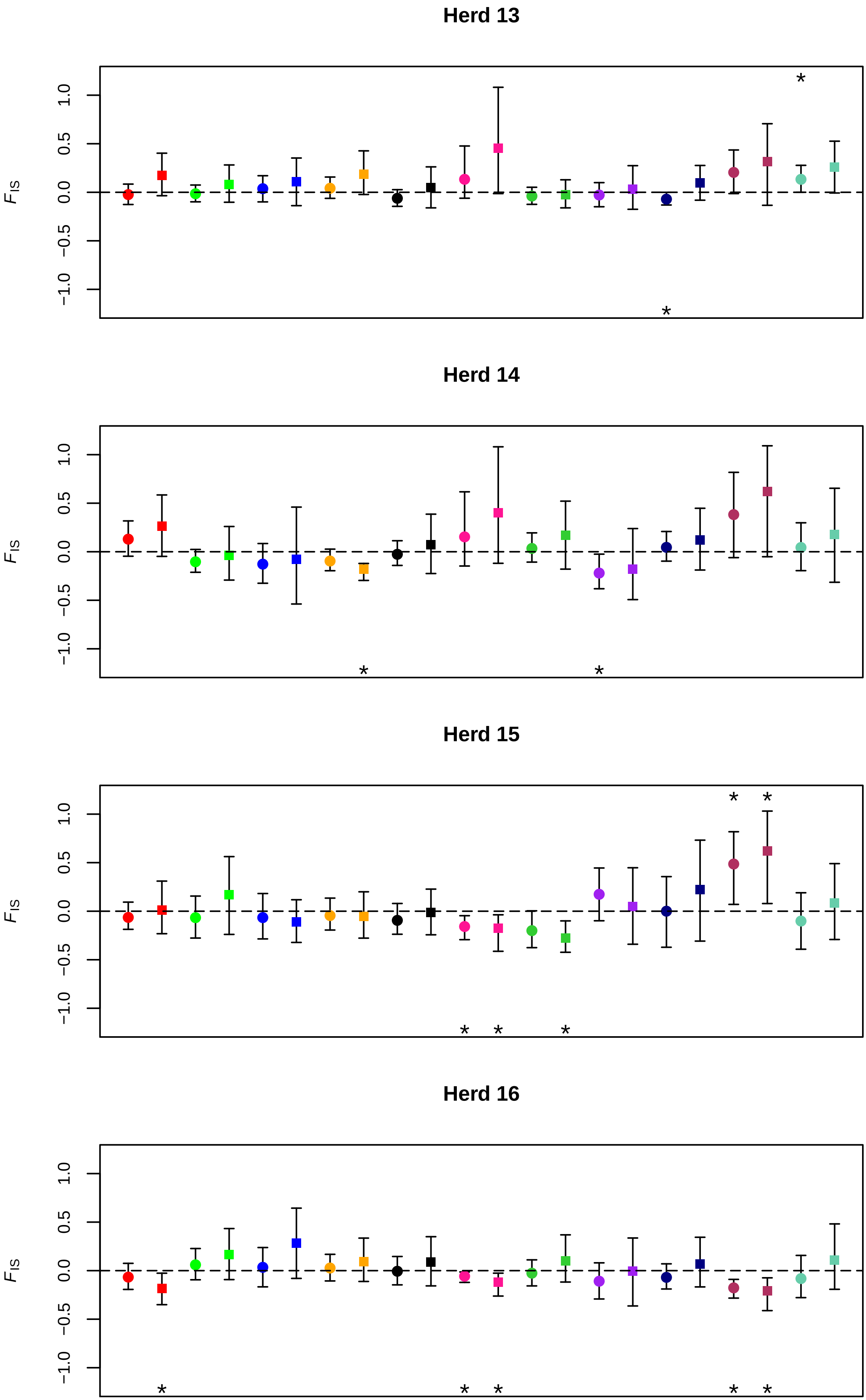
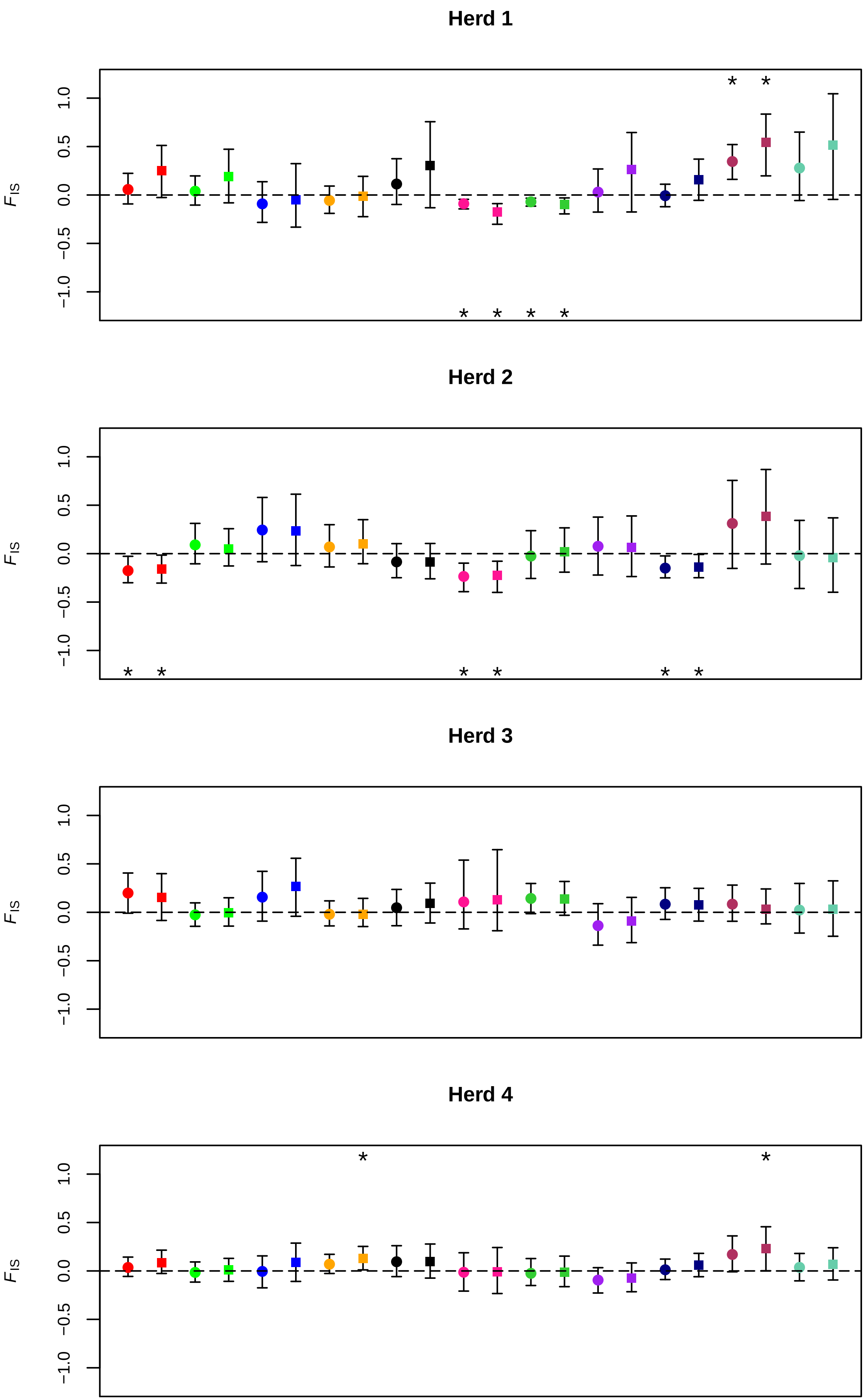
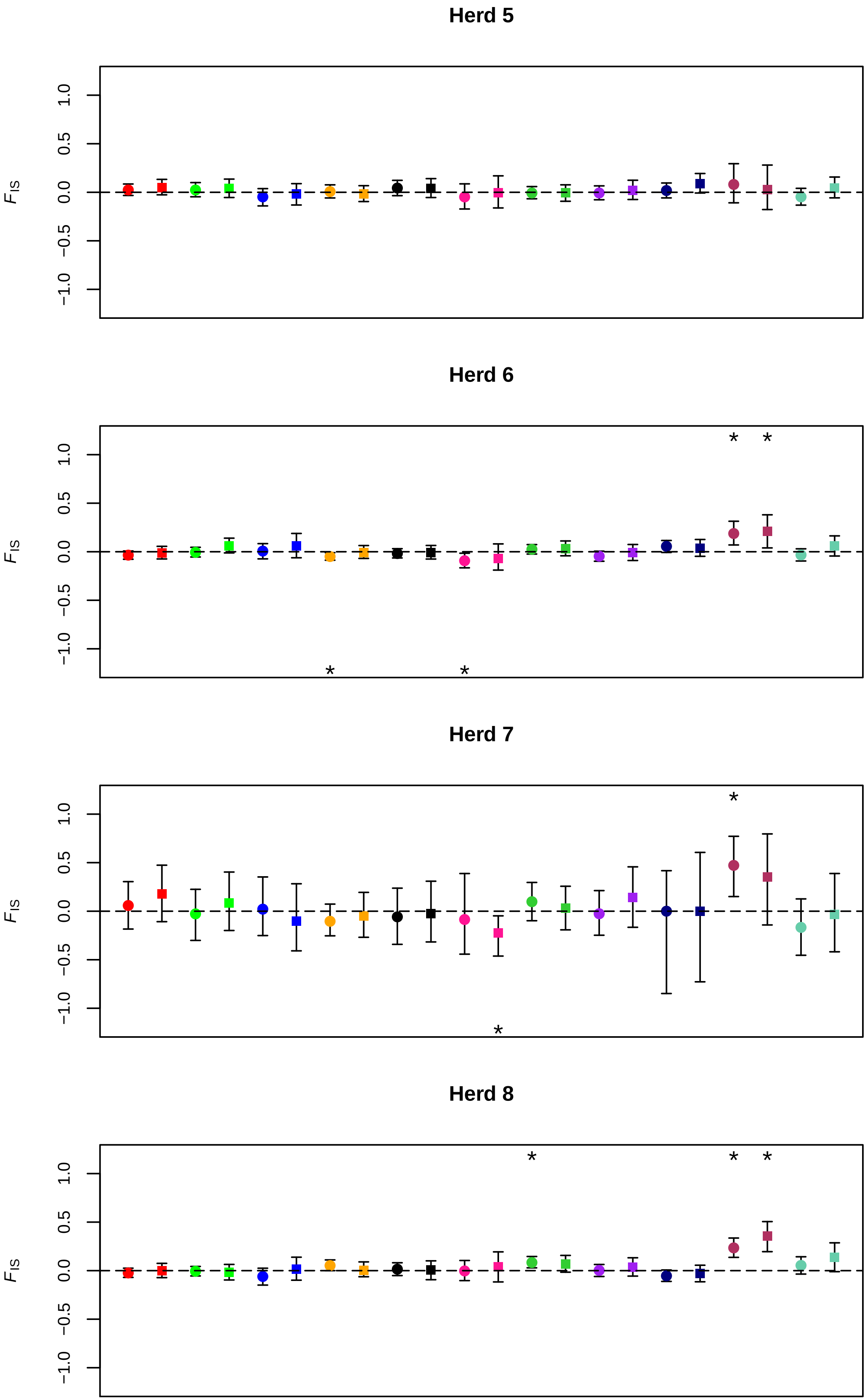

Herd 13
Herd 14
Herd 15
Herd 16
F_{IS}
1.0
0.5
0.0
−0.5
−1.0
*

Fig. S1 Scatterplot of inbreeding coefficients, F_{IS}, with lower and upper 95% confidence intervals, of each marker in each herd in the full data set (•) and after relatives were removed (■) using a relatedness cut-off of 0.25. Cases where F_{IS} deviated significantly from zero are indicated by an asterisk (*). Colours correspond to microsatellite loci: red – BM1824, green – BM3205, blue – BM3517, orange – BM719, black – CSSM19, deep pink – ETH10, lime green – ILSTS026, purple – INRA006, navy blue – SPS115, maroon – TGLA227, medium aquamarine – TGLA263.


Herd 1
Herd 2
Herd 3
Herd 4
F_{IS}
1.0
0.5
0.0
−0.5
−1.0
*

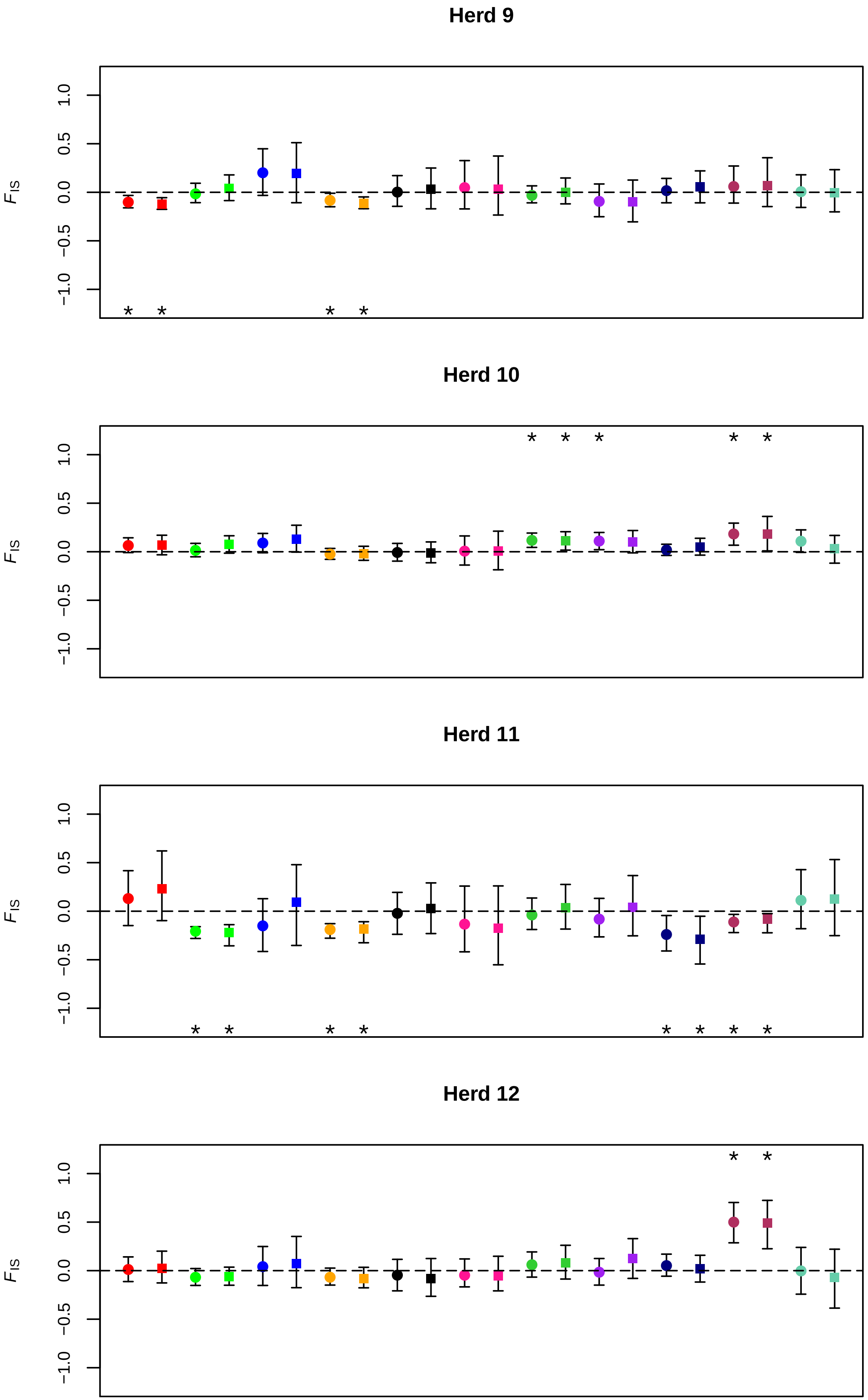


Herd 5

Herd 6

Herd 7

Herd 8


Herd 9
Herd 10
Herd 11
Herd 12
F_{IS}
1.0
0.5
0.0
−0.5
−1.0
*

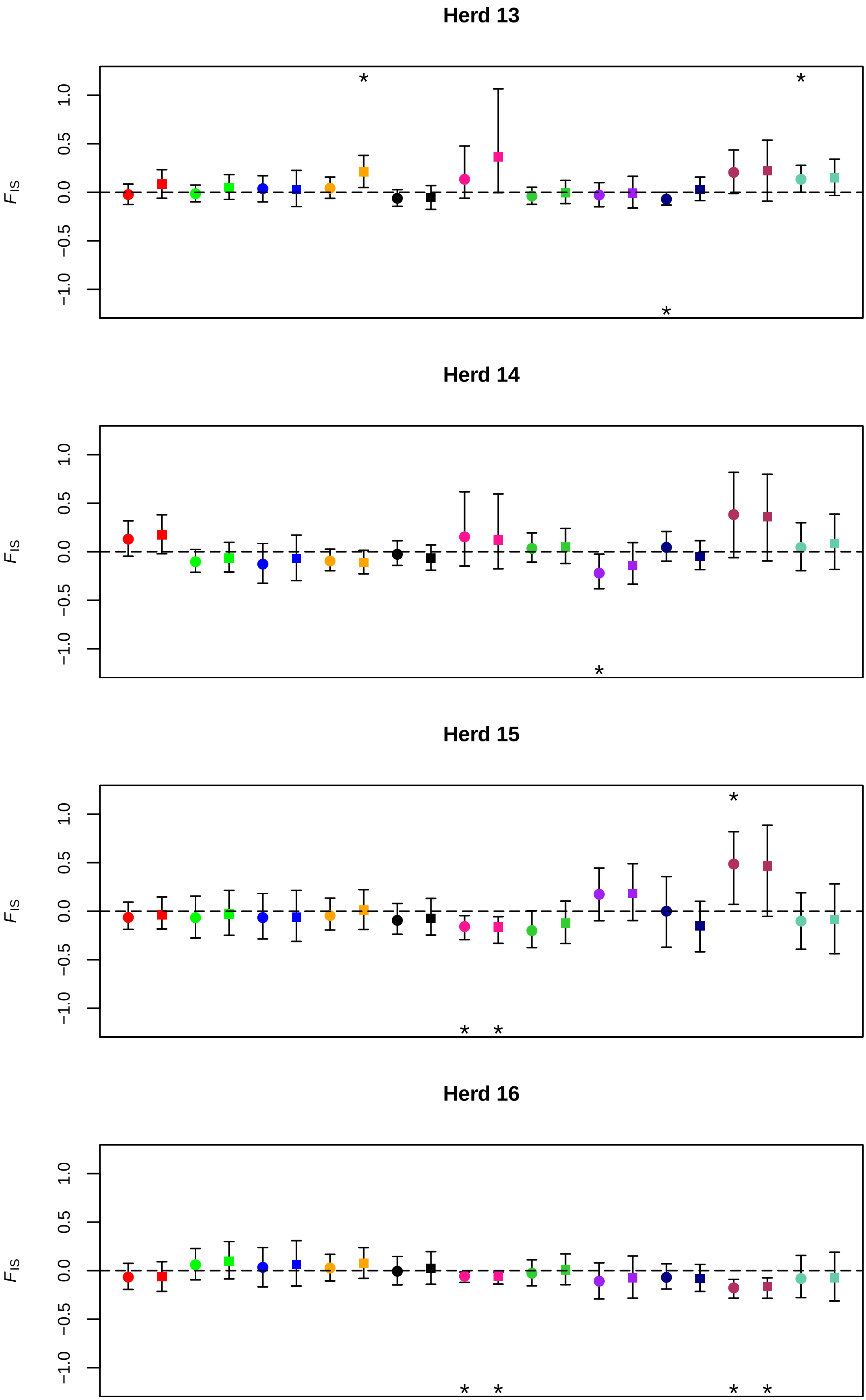

Herd 13
Herd 14
Herd 15
Herd 16
F_{IS}
1.0
0.5
0.0
−0.5
−1.0
*

Fig. S2 Scatterplot of inbreeding coefficients, F_{IS}, with lower and upper 95% confidence intervals, of each marker in each herd in the full data set (•) and after relatives were removed (■) using a relatedness cut-off of 0.5. Cases where F_{IS} deviated significantly from zero are indicated by an asterisk (*). Colours correspond to microsatellite loci: red – BM1824, green – BM3205, blue – BM3517, orange – BM719, black – CSSM19, deep pink – ETH10, lime green – ILSTS026, purple – INRA006, navy blue – SPS115, maroon – TGLA227, medium aquamarine – TGLA263.