Title: Detecting non-random mating or selection in natural populations using multi-locus gene families

Authors

Gabe D O'Reilly [1]*; Oliver Manlik [2]; Sandra Vardeh [3]; Jennifer Sinclair [4]; Zack P Lawler [5]; William B Sherwin [1]

* Corresponding author

Gabe D O'Reilly; gabe.d.oreilly@gmail.com

Author affiliations

[1] University of New South Wales, Evolution and ecology research centre: School of BEES Sydney, NSW, AUS

[2] UAE University Al Ain, Abu Dhabi, AE

[3] Bundesamt für Naturschutz Bonn, Nordrhein-Westfalen, DE

[4] Cape Bernier Vineyard, Bream Creek, TAS, AUS

[5] The University of Newcastle Newcastle, NSW, AUS

Abstract

New sequencing technologies have opened the door to many new research opportunities, but these advances in data collection are not always compatible with some important methods for data analysis. F_{IS} has been a staple calculation in the field of population genetics. F_{IS} can be used to measure either a departure from random mating, or measure underlying selective pressures for or against heterozygote genotypes. However, when using Next Generation Sequencing (NGS) technology on multi-locus gene families it is often impossible to discern which allelic variants are present at each locus. This in turn makes it impossible to calculate either locus-specific expected heterozygosity, or observed heterozygosity, both of which are required to calculate F_{IS}. This is unfortunate because

there are many important multi-locus gene families such as: the major histocompatibility complex (MHC) in animals; homeobox genes in fungi; or the self-incompatibility genes in plants. Without the ability to calculate F_{IS} from NGS of multi-locus gene families, we need a new multi-locus measure that will allow us to detect the underlining mating, and selective patterns present in such multi-locus genes. This paper provides such a novel multi-locus measure, called ${}^1H_{IS}$. We demonstrate the accuracy of the ${}^1H_{IS}$ equation using simulated data, and two datasets taken from natural populations of dolphins and penguins. The introduction of this new measure is particularly important because of the great interest in mating patterns and selection of multi-locus gene families, such as MHC.

Key words:

Population Genetics, Inbreeding, Selection, Simulation

Introduction

Determining mating structure and selection in a population are the primary aim of much population genetics research because such knowledge is a critical part of all population biology and allows us to effectively manage populations. There are numerous methods to quantify mating structure, but F_{IS} is the method that has seen the most use and is generally the standard for quantification of a population's mating structure. F_{IS} is often called the inbreeding coefficient, but F_{IS} also has other applications that are not related to inbreeding, described below. F_{IS} compares the proportion of heterozygotes expected in a randomly mating population (H_e, equation 1) to the actual number of heterozygotes observed in a study population (H_o, equation 2) (Halliburton, 2004). H_e is calculated from the proportions of alleles in the population and is commonly used as a measure of genetic diversity.

$$H_e = 1 - \sum_{i=1}^{V} {P_i}^2 \qquad [Equation\ 1]$$

Where capital '*V*' is the number of variants (in this case allelic types) for that locus and P_i is the proportion of the *i*th allele in the population ($\sum_{i=1}^{V} P_i = 1$).

$$H_o = Proportion\ of\ population\ that\ are\ heterozygous\ at\ that\ locus \qquad [Equation\ 2]$$

The equation for F_{IS} is

$$F_{IS} = \frac{H_e - H_o}{H_e} \qquad [Equation\ 3]$$

This comparison gives an F_{IS} value between -1 and +1 that indicates how the number of heterozygotes in the population deviates from what is expected under random mating conditions. A positive F_{IS} value indicates that there are fewer heterozygotes than expected under random mating. For instance, an inbred population will often have a much lower proportion of heterozygotes than expected, and this deficit leads to a positive F_{IS} value; in the extreme case when there are no heterozygotes observed at all, despite available allelic variation, then $F_{IS} = +1$. In contrast, an outbred population will have a much higher proportion of heterozygotes than expected under random mating, and this excess leads to a negative F_{IS} value. Thus F_{IS} for selectively neutral genes can be used to determine mating patterns. However, the heterozygote excess or deficit that F_{IS} measures can be due to either mating pattern or selection, so if the population is already known to be randomly mating, F_{IS} can be useful for detecting signatures of selective pressures. It should be noted that random genetic drift can also be a factor affecting F_{IS}, because it is possible for an excess or deficit of heterozygotes to occur from chance, causing F_{IS} to deviate from zero.

Such random factors can be a source of variation for F_{IS} in randomly mating populations, although it is unlikely they would be a source of extreme deviation of F_{IS} from zero. F_{IS} can be applied to investigate selection for or against heterozygous individuals in genes that may not be selectively neutral. F_{IS} can be calculated by two methods: either for a single locus or multiple un-linked/independent loci that do not share common alleles. To calculate F_{IS} on multiple loci, first calculate F_{IS} independently at each locus, then average across those loci to get a single F_{IS} value. This method is often called 'Multiple F_{IS}'. Multiple F_{IS} is useful because sampling multiple loci lessens the between-locus variance of F_{IS}.

However, F_{IS} and Multiple F_{IS} are difficult to use on multi-locus gene families, which can share common alleles across several loci (Ellis *et al*, 2005; Zagalska-Neubauer, 2010). Multi-locus gene families are of particular interest when investigating mating patterns and selection (Sommer, 2005). For example, multi-locus gene families such as the major histocompatibility complex (MHC) have been associated with fitness and various fitness components, including mate choice(Yamazaki *et al*, 1988), immune defence (Klein, 1986; Altizer, 2003) and reproductive success (Kalbe *et al*, 2009; Thoss *et al*, 2011; Sepil *et al*, 2013). This makes MHC an important multi-locus gene family to study for examining mating patterns and selection within a population. With current sequencing methods, it is often impossible to determine which alleles are present at which loci in multi-locus gene families, unless a model species is being studied (even then, it can be difficult) (Babik, 2010). Generally, the sequencing output will just give relative abundance of each variant per individual summed over all loci at which the alleles appear (Figure 1) (Manlik, 2016; Vardeh, 2015). Not knowing the exact location of each allele is a problem because, as for any other multiple F_{IS}, in multi-locus gene families F_{IS} must be calculated independently for each locus

then averaged to produce the F_{IS} of that multi-locus gene family. As shown in Figure 1, there is considerable ambiguity even with a very small multi-locus gene family containing only two loci. With such ambiguity, it becomes impossible to accurately calculate $H_e$or H_o per locus, and therefore impossible to calculate F_{IS}. Figure 1 also depicts the assumption that we know exactly how many loci make up the multi-locus gene family; however, outside of model organisms or heavily studied multi-locus gene families, this is often not the case. Not knowing the exact number of loci adds even more ambiguity to the calculation of F_{IS}.

F_{IS} is also difficult to use when analysing polyploid species. Similar to the issues with multi-locus gene families, current sequencing methods cannot always distinguish which variant comes from which chromosome, which becomes a problem when studying loci that are repeated across polyploid homologues. Additionally, much like not knowing how many loci are in a multi-locus gene family, not knowing how many homologue chromosomes there are could also add to this ambiguity. We will refrain from referring to this polyploid issue directly for the rest of the paper, but all the solutions we apply to multi-locus gene families can be applied in the same manner to polyploid data.

This paper aims to introduce methods/algorithms that can deal with this ambiguity that multi-locus gene families introduce into the calculation of multi-locus F_{IS}, and instead give a reasonable estimation of a population's mating structure (or selection on heterozygotes). We also aim to explain a method that can give a reasonable estimate of number of loci in a multi-locus gene family for a non-model species. This paper sets out to devise an adequate solution to the problem of $\mathrm{F_{IS}}$ in multi-locus gene families by developing:

1. An equation to calculate mating structure, or selection on heterozygotes, within multi-locus gene families.
2. An algorithm to estimate the number of loci in multi-locus gene families.

Materials and Methods

Equation

Our method is based on the rationale that when there is either inbreeding or selection against heterozygotes, there is expected to be less diversity of variants within each individual relative to the total diversity of variants across the population. The opposite is true of populations that experience outbreeding or selection for heterozygotes. The total amount of diversity an individual can hold will also be linked to the number of loci present. The method described below is based on these understandings, and with them we can construct a potential equation for calculating mating structure for multi-locus gene families by applying Shannon's information theory to the problem. Other approaches were attempted; however, they did not give suitable results (Supplement S2: Figure S2, Figure S3, and Figure S4). Shannon's information (${}^1H$) is a general measure of diversity, originally developed for telecommunications (Shannon, 1949), and since applied to population genetics (Sherwin *et al*, 2017; Manlik *et al* 2019b; O'Reilly *et al*, 2020). A potential F_{IS} analogue based on Shannon's information is called ${}^1H_{IS}$, and its two possible equations are as follows:

$$ {}^1H'_{IS} = -\left(\frac{\overline{{}^1H_I}(L+1)}{{}^1H_S} - L\right) \qquad [Equation\ 4a] $$

where the number of loci (or an estimate for number of loci) is *L*, and ^1H_I is the Shannon's information per individual based on the proportions of each variant within each individual p_i, using the equation $^1H_I = -\sum_{i=1}^{v} p_i \ \ln p_i$. Lower case 'v' is the total number of variants in the individual (that may or may not be alleles at the same locus $\sum_{i=1}^{v} p_i = 1$). Then to produce $\overline{^1H_I}$ one averages those Shannon's information values across all individuals to get $\overline{^1H_I}$. ^1H_S is based on using the total proportions of variants in the whole population P_i, to calculate Shannon's information as $^1H_S = -\sum_{i=1}^{V} P_i \ \ln P_i$, where capital '*V*' is the total number of variants in the population (that may or may not be alleles at the same locus $\sum_{i=1}^{V} P_i = 1$). In equation 4a, the foundation of $^1H_{IS}$ is the comparison between the diversity held within individuals ($\overline{^1H_I}$) and the diversity held within the total population (^1H_S), which is why Equation 4a takes the form of $\frac{\overline{^1H_I}}{^1H_S}$. When sampled individuals contain all the diversity found in the total population $\frac{\overline{^1H_I}}{^1H_S} = 1$, indicating a higher likelihood of heterozygotes, so we would expect a negative value of F_{IS}, and thus we would like a similarly negative value for $^1H'_{IS}$, hence the negative sign at the front of equation 4a. Additionally, number of loci acts as a way to weight ^1H_I to *L* , making the calculation more sensitive to differences between $\overline{^1H_I}$ and ^1H_S with more loci. With more loci, the numerator in the equation is inflated, possibly giving $\frac{\overline{^1H_I}(L+1)}{^1H_S}$ values beyond 1, which is then brought back to the -1 to +1 scale by $-L$. Equation 4a is a transformed into Equation 4b, by adding a correction using the genetic evenness of the population.

$$ {}^1H_{IS} = -\left(\frac{\overline{{}^1H_I}(L+1)}{{}^1H_S} - L\right) E_V \qquad [Equation\ 4b] $$

Genetic evenness (E_V) is a measure of how evenly distributed alleles are, with E_V reaching its maximum value when all alleles are equally frequent, where Max ${}^1H_S = ln(V)$. So $E_V = {}^1H_S/ln(V)$, where *V* is the number of variants in the population. Greater evenness of variant proportions would bring $\overline{{}^1H_I}$ closer to ${}^1H_S$, with the same mating pattern or selection, so the multiplication by evenness corrects for this effect. Equation 4b is used for ${}^1H_{IS}$ for the rest of this paper unless stated otherwise, because Equation 4b gave more accurate results.

To give a basic example of equation 4b applied to a population with a single locus with two allelic variants, where $E_V = 1$: If $E_V = 1$ with two variants, then ${}^1H_S = 0.690$. If the whole population are homozygotes, then ${}^1H_{IS} = -\left(\frac{0\ (1+1)}{0.690} - 1\right) 1 = 1$, indicating maximum inbreeding (a heterozygote deficit). If the whole population are heterozygotes, then ${}^1H_{IS} = -\left(\frac{0.69\ (1+1)}{0.69} - 1\right) 1 = -1$, indicating minimum inbreeding (a heterozygote excess). If the population has an even mix of homozygotes and heterozygotes (with every possible genotype evenly represented in the population), then ${}^1H_{IS} = -\left(\frac{0.345\ (1+1)}{0.690} - 1\right) 1 = 0$, indicating random mating (neither excess nor deficit of heterozygotes).

Estimation of number of loci

As mentioned above, it is common that the exact number of loci ("*L*") is not known when studying a multi-locus gene family in a non-model organism. It may be possible to estimate the number of loci by examining other research in similar organisms or multi-locus gene families; however, these approaches may not be helpful when dealing with a novel species, and because number of MHC loci can even vary within species (Bowen *et al,* 2004; Siddle *et al*, 2010). Nevertheless, it is possible to estimate the minimum number of loci if certain assumptions are made. To estimate the number of loci, we must assume that at least one individual, which has been sampled from the population, has a true singleton variant sequence post filtering (i.e. a sequence variant that only occurs once in a particular individual, across all its loci). It then becomes possible to calculate the minimum number of loci. The closer the sample data comes to fulfilling this assumption, the more accurate the estimation of number of loci will be. The data set will give relative proportions of variants over all loci as ratio, with the assumed singleton being "1" in that ratio. For example, the output for an individual with three variants (variants *B1*, *B2* and *B3*) might be presented as the ratio: 1:3:2, indicating that there are three times as many *B2* variants than *B1* variants, and two times as many *B3* variants than *B1* variants. If we assume that this individual has only one *B1* variant (ie the individual is heterozygous for the *B1* variant at a single locus), the sum of this ratio (1+3+2) would be twice the minimum number of loci (because each locus contains two alleles). In this case $2L \geq 6$, so there are at least $L \geq 3$ loci (L is rounded off to the nearest decimal point). With very high read depth, whichever individual gives the biggest

sum of the ratio will provide the most accurate estimate of the minimum number of loci, which will be the assumption for this initial explanation (Table 1).

It is common for NGS datasets to be formatted in number of sequence reads rather than ratios. In this case, for each individual simply divide each number of sequence reads per variant by the lowest value of number of sequences reads for any variant (post filtering), then round up any decimals to a whole number. This should result in ratios which can then be used for the above method to get a minimal estimate of number of loci. For example, if the read numbers for variants B1 B2 and B3 were 12, 35, 23 respectively, then divide the values by 12 (the lowest read value) to get 1:2.9:1.9, sum those to get 5.8 and round to the nearest integer, giving $2L = 6$, meaning there is a minimum of three loci. Similar methods have been applied previously, although they use presence/absence of alleles, rather than attempting to quantify abundance using sequence reads as our method does (Babik *et al,* 2009; Heimeier *et al* , 2009 ; *Sommer et* al, 2013). Some methods have also identified some loci as fixed in order to get better estimations alongside presence /absence methods (Stervander *et al*, 2020), and other methods have taken a computational approach to genotyping MHC (Stuglik *et al*, 2011).

While this approach is useful it should be noted that this method can only give a minimum estimate of the number of loci, but never overestimate unless there is some significant form of sequencing error that cannot be fixed during the filtering process. When dealing with real data, there may still be sequence misreads, or missing data, in which case it may be more appropriate to calculate *L* for each individual, and use the mode or average of those *L* values to minimise the impact of any sequence read errors greatly inflating the value of *L*. While

using the mode and average of individual *L* values will help when the data has sequence misreads or missing data, it may not overcome more systematic issues to do with NGS data, such as low read depth or NGS's inherent stochasticity in which variants it sequences (Smith & Peay, 2014; Qin *et al*, 2016), or other inherently random processes throughout genetic sequencing and data collection. The impact of NGS data's stochasticity will be explored later in this paper. We call this the 'One Individual' locus number estimate. An alternate method was also tested for estimating number of loci based on the assumption that at least one individual in the population would only have singletons, however this method was not as accurate and is presented in the supplement (Supplement S1: Equation S1 and Figure S1).

Creating a simulated dataset

To test ${}^{1}H_{IS}$ (Equation 4b) stochastic, forward time simulations were performed using the PYTHON package SIMUPOP (Peng and Kimmel, 2005) The full code used is available on request. The simulations were set up to give a range of F_{IS} values. Random mating was simulated to give F_{IS} of approximately 0 (Though some minor variation due to random drift did occur). Other values of F_{IS} were obtained by manipulating the mating structure and selection within the simulated populations. To increase F_{IS} in the simulated populations, we applied two simulation treatments. Firstly, selective pressure in favour of homozygotes was applied by increasing mating chance of homozygotes. Secondly, mating was restricted to smaller sub-populations of ten, mimicking small families. To decrease F_{IS} in the simulated populations, we applied selective pressure in favour of heterozygotes. Two different treatments of selective pressure were applied in favour of heterozygotes, one with mild selection, and one with stronger selection; this was done by increasing heterozygotes' chance of mating.

As well as manipulating mating structure and selection within the simulations, we also varied the number of loci, variant distribution, number of generations, and population size. Multi-locus gene families have varying numbers of loci, and because the number of loci is a key value in the calculation of ${}^{1}H_{IS}$, we ran simulations with different numbers of loci to see how that altered the accuracy of ${}^{1}H_{IS}$. The number of loci was set to 3, 5 or 10 loci. Variant distribution was manipulated to see if the presence of rare variants altered the accuracy of ${}^{1}H_{IS}$, especially regarding the estimated number of loci. Variant distribution had two treatments: 'Even', where each variant had an equal chance of occurrence at the start of the simulation; and 'Uneven', where one variant was given a 70% chance of occurring and the rest of the variants comprised the remaining 30% chance in equal proportions. There were ten different variants per locus in each of these variant distributions. For each treatment group, we ran separate simulations for three different numbers of generations – 10, 100, and 1000. Mutation and recombination were both set not to occur in our simulation. Population size was also altered per treatment group. The final treatment parameters were:

- Mating: Small Family and Random Mating, with expectations of positive and zero F_{IS} respectively
- Selection: Mild Selection for Homozygotes(60% chance for homozygotes to mate, 50% chance for heterozygotes to mate); Mild Selection for Heterozygotes (20% chance for homozygotes to mate, 60% chance for heterozygotes to mate); and Strong Selection for Heterozygotes (0% chance for homozygotes to mate, 100% chance for heterozygotes to mate); with expectations of positive, mildly negative and strongly

negative F_{IS} respectively. Note that all selection treatments were applied only to random mating populations.

- Number of loci: 3 loci, 5 loci, and 10 loci
- Variant distribution: Even, and Uneven; with ten variants per locus each
- Generations: 10 generations, 30 generations, and 50 generations
- Population size: 40 individuals, 400 individuals

All combinations of values of the parameters were tested, giving a total of 180 treatment groups. There were 100 replicates of each treatment. The data from these simulations were first used to calculate multiple F_{IS} for these simulated populations, because the exact number of loci and which variants were allelic to them was known. Next, the data were converted to a format that resembled data with all the limitations of a real study (i.e., No information on which variants are at which loci and no information on number of loci), and Equation 4b was applied. For the simulated dataset, there were no missing variants in the data (i.e., if an individual actually had a particular variant, that variant always appeared in the data).

Unfortunately, with current sequencing methods on multi-locus gene families, it is not currently possible to know if all loci are fixed across the population (fixed meaning only one variant at a specific locus across all individuals). Therefore, we also decided to calculate the variance of Shannon's information between individuals, which is the variance of ${}^1H_I$, and so can be calculated on any dataset where ${}^1H_{IS}$ is used. When the same variants are present in all individuals (and thus there is low or no variance of Shannon's information between

278 individuals), it typically means that those variants are fixed at certain loci throughout the
279 population.

280 Impact of read depth on our locus-number estimation method

281 Next generation sequencing (NGS) data is produced by a stochastic sampling process, during
282 which variants detected in an individual are selected at random to be recorded. In single-
283 locus sequences this is not an issue, because if the sequencing results show two variants we
284 know that there are two variants (i.e. alleles) at that locus. Moreover, if the sequencing only
285 reveals one variant, the probability of bypassing an equally proportioned second variant is
286 quite low, so we can be confident that the sample is a homozygous for that locus (site).
287 However, when dealing with large multi-locus gene families, there is an increased chance for
288 the replicates in the sequencing sampling process (read depth) to either completely miss a
289 variant or to sample some variants more than others despite equal proportions in the
290 individual. This problem is exacerbated when the read depth is lower than the number of
291 loci in the individual, because it will become impossible to sample every single variant (in
292 their relative proportions), because there simply will not be enough replicates to represent
293 each variant in an individual. In the context of our method, this would create problems with
294 estimating number of loci. To further test the robustness of our locus-number estimation
295 method, some additional simulations were done with an added layer of obfuscation to the
296 final dataset. Code was written to stochastically sample variants within individuals for the
297 final dataset, similarly to an NGS sampling process (Code available on request). We then
298 applied our locus-number estimation to this new simulated NGS dataset to see how it would
299 affect the accuracy of our locus-number estimation. This new dataset was reduced in scope

compared to the main dataset used, by removing strong selection treatments. Our artificial read depth value was set to 30.

Assessment of simulation results

Simulated data results were assessed by comparison of the ${}^{1}H_{IS}$ measurement to the F_{IS} measurement of the same population. These comparisons were done with linear regression, as well as by calculating Root Mean Squared Error (RMSE, equation 5).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(({}^{1}H_{IS})-(F_{IS}))^2}{n}} \qquad [Equation\ 5]$$

Where *n* is the number of values.

Our data from all treatments had a large proportion of values clustered around F_{IS} = 0, which would greatly bias regression results towards replicates that were near F_{IS} = 0. To correct for this, values of F_{IS} were binned at intervals of 0.1 from an F_{IS} range of -1.05 to +1.05. Twenty values were randomly sampled within each bin; however because the -1.05 to -0.95 bin and the -0.95 to -0.85 bin had twenty values total when combined, they were made to be a single bin. This binning ensures that the range of F_{IS} is weighted equally in the regression. Regression results without this binning can be found in the supplement (Supplement S4). There were no datapoints that could be in two bins simultaneously.

If all loci were monomorphic (i.e. there is only one variant in the population), ${}^{1}H_{IS}$ will not be accurate, and F_{IS} will not work at all. This would be seen as invariance of genetic patterns across all individuals, so probably these analyses would not be attempted. However, in an intermediate case in which only some loci were fixed, what would be the effect on these methods? Fixation of different variants at individual loci for different

variants (locus-specific fixation) would not be known in real NGS data for multilocus gene-families, but the variance of ${}^1H_I$ can be calculated, which acts as a proxy measure for fixation. Variance of ${}^1H_I$ would be zero if there is total or locus-specific fixation (all individuals in the population have the exact same variant proportions through the population, such as complete fixation across all loci in the family, or locus-specific fixation for each locus). To identify and deal with cases which are close to fixation for all loci, we removed replicates that had less 1×10^{-10} variance in ${}^1H_I$ before binning. Again, these are cases where researchers are unlikely to wish to use ${}^1H_{IS}$ or F_{IS}.

Assessment of ${}^1H_{IS}$ on the simulated NGS data set was analysed using the same methods above for the full simulated dataset. The data filtering process was slightly different however, because replicates with low ($< 1 \times 10^{-10}$) variance of richness were removed as opposed to replicates with low ($< 1 \times 10^{-10}$) variance of individual Shannon's information (as was done for the complete simulated data). The rationale behind this was that due to the low read depth it was likely that some variants would not be sequenced.

Dolphin and Penguin Data

While simulations and theory can establish whether the method works as intended under a known, wide range of mating and selection patterns, it is crucial to investigate whether a method can overcome the obstacles of real, imperfect datasets. Therefore, we have applied these methods to MHC class I data from two dolphin populations (*Tursiops* sp.), Shark Bay (SB), and Bunbury (BB) (Manlik 2016; Manlik *et al*, 2016); as well as MHC from three penguin populations (*Eudyptula minor*), Perth (PER), Albany (ALB), and Esperance (ESP) (Vardeh, 2015). We have also compared these results of the dolphin and penguin MHC data to F_{IS}

343 results of microsatellite data (Vardeh, 2015; Manlik *et al*, 2019b) from those same
344 populations, as a partial verification of the results of equation 4b. Additionally, F_{IS} was also
345 calculated on what appeared to be a single-locus MHC dataset of 75 female dolphins from
346 SB, using MHC II DQB. This study (Manlik et al. 2019b) was most likely on a single locus—
347 MHC II DQB. Manlik *et al* did not detect any patterns in the MHC II DQB sequences that
348 indicated multiple allelism (i.e. having more than two alleles or sequence variants per
349 amplicon/individual), gene duplications, stop codons, or frameshifts. Additionally, comparing
350 MHC DQB sequences of seven mother–father–offspring trios did not reveal any patterns that
351 were inconsistent with single-locus Mendelian inheritance.

352 In the dolphins, MHC I genetic variants were amplified and sequenced using Illumina MiSeq
353 paired-end sequencing technology at the Ramaciotti Centre of the University of New South
354 Wales. The primer pairs used in the amplification process targeted MHC variants that had
355 been previously described as being part of MHC I, exon 2 (Flores-Ramirez *et al*, 2000) in gray
356 whales. However, due to the complex multi-locus nature of MHC, for the study it was not
357 possible to assign MHC amplicon sequences to a particular locus. After quality-filtering in
358 MOTHUR (version 1.34.0) (Schloss *et al*, 2009), true sequence variants were identified
359 following the general filtering process outlined by Sommer *et al* (2013) (see Manlik, 2016).
360 MHC II genetic variants followed the same procedure, but the primer pairs used in the
361 amplification process targeted (Manlik *et al*, 2019a). The full MHC data processing and
362 filtering can be found in Manlik (2016).

363 The final sample for dolphin MHC I from SB had 24 individuals and BB had 11 individuals. The
364 final results to which we applied our method were sequence variant percentage per

individual. We used these proportions for all subsequent calculations, $\overline{{}^{1}H_I}$ was calculated based on the proportions of each variant in each individual p_i , (equation ${}^{1}H_I = -\sum_{i=1}^{v} p_i \log p_i$), then averaging those Shannon's information values across all individuals to get $\overline{{}^{1}H_I}$ in equation 4b. ${}^{1}H_S$ was calculated by averaging proportions of each sequence variant across individuals, then calculating Shannon's information based on these new proportions P_i. To estimate number of loci, for each individual we divided each P_i by the lowest P_i, then summed these ratios to give a new value. That value was then divided by two and rounded to zero decimal places for the individual with the most variants. However, misreads can heavily influence the maximum value obtained from the individual with the most variants, and this can lead to incorrect results. So, in addition to that One Individual locus-number estimation, we also calculated ${}^{1}H_{IS}$ using the average and the mode from all individuals. Average and mode Locus-number calculations were rounded to the nearest integer because loci only exist as whole numbers.

We also applied our method to the Penguin MHC data, that was summarised as number of sequence variant reads per individual. The data had sequences reads for 8766 different MHC variants, after filtering to remove any variant that only occurred once in the whole population (Vardeh, 2015). We then further filtered to remove any sequence read that did not make up at least 10% of an individual penguin's reads. Sequences that were filtered out of one penguin could still occur in data for other penguins. This was done primarily to remove low read counts within each penguin. The problems with misreads influencing One Individual locus-number estimation discussed above with the dolphin data holds true for the penguin data, so again, in addition to a One Individual locus-number estimation, we

calculated ${}^{1}H_{IS}$ using the average and the mode of *L* (Table 3). Locus-number calculations were rounded to the nearest integer because loci do not naturally occur in fractions.

When assessing the results of the dolphin and penguin data we cannot obtain replication within the individual population, so statistical methods cannot be applied. Instead we assessed whether the direction of departure from random mating (positive or negative) was consistent between the MHC ${}^{1}H_{IS}$ and the microsatellite F_{IS} , as well as MHC I ${}^{1}H_{IS}$ and single-locus MHC II F_{IS} In the dolphins. In doing these comparisons, we must bear in mind that they would only be expected to show deviations from zero in the same direction if both were controlled by the same processes, such as no selection but some inbreeding, or random mating with the same selection on all loci, which is unlikely. For the Dolphin populations, Manlik (2016) assumed the MHC genes were under selection, and the microsatellite loci were thought to be neutral.

Results

Simulated dataset results

Results were analysed to investigate if we could come to the same conclusions about a population mating structure using a ${}^1H_{IS}$ value that we would using a F_{IS} value. Simulated results were analysed as a combined dataset (with all treatments together, Figure2), as well as when separated by different treatment parameters such as number of loci and allele distribution (Figure 3 to Figure 6). The comparison of ${}^1H_{IS}$ values with their corresponding F_{IS} values across the whole binned dataset, showed a strong regression fit, close to the expected 45° line (Figure 2). Examining only simulations that altered the number of loci that were set in each simulation, showed that ${}^1H_{IS}$ results performed well in all cases, but better with a larger number of loci (Figure 3). Three-locus treatments only showed a range of F_{IS} values from ~-0.5 to 1, five locus treatments from ~-0.5 to 1, and ten locus treatments showed the full range from -1 to 1. Simulations given one of two variant distribution treatments showed that ${}^1H_{IS}$ performed well in both cases, but better in the 'Uneven' variant distribution treatment (Figure 4). ${}^1H_{IS}$ in Even treatments showed a reduced range of F_{IS} values, from ~-0.5 to 1, whereas Uneven treatments showed the full range of F_{IS} values from -1 to 1.

Simulation results were also analysed by separating data based on the demographic parameters: population size and generations of breeding. Simulations were set to run for one of three generation times, giving other treatment parameters more time to affect the data. There was good regression fit in all cases, though slightly weaker with the longest generation time (Figure 5). As generation time within each simulation increased, number of

421 replicates with low ${}^1H_I$ variance also increased. Simulations were run with each of two
422 population sizes, which marginally influenced accuracy of ${}^1H_{IS}$, and the range of values for
423 ${}^1H_{IS}$ and F_{IS} (Figure 6). Note that in the larger population sizes, values tended to form
424 clusters, which related to initial values of variables other that population size: Small families
425 simulated for ten generations (F_{IS}= ~0.45 cluster); Small families simulated for 30 and 50
426 generations (F_{IS}= 0.6+ cluster); random mating and selection treatments (F_{IS}= ~0 cluster).

427 ## Simulated NGS data

428 To investigate the effect of low read depth, additional simulations were also analysed on
429 data that has been obfuscated in a similar way to real NGS data, due to low read depth.
430 These simulated NGS-like data increased the error of our locus-number estimates, and thus
431 gave worse regression results than their non-NGS like counterpart in Figure 2, though still
432 with good R-squared and RMSE (Figure 7).

433 ## Dolphin Data

434
435 All values and results from the ${}^1H_{IS}$ calculations, along with F_{IS} results from the
436 microsatellite data are listed in Table 2. Shark Bay (SB) microsatellite data for the same
437 population showed results that agree with the sign of our ${}^1H_{IS}$ method for MHC I in the
438 same population. For SB the positive ${}^1H_{IS}$ values suggest inbreeding or selection for
439 homozygotes, which is consistent with the F_{IS}, based on microsatellites. However, the ${}^1H_{IS}$
440 gave values an order of magnitude larger than F_{IS}. Also at SB, the F_{IS} value of MHC II DQB
441 showed a negative F_{IS} value, indicating a disagreement with the ${}^1H_{IS}$ results and F_{IS} from
442 the microsatellites. For Bunbury (BB) the MHC I ${}^1H_{IS}$ results based on the average or mode
443 locus-number estimates are consistent with the microsatellite F_{IS} value in both direction and

magnitude. However, the One Individual locus-number estimate, based on the BB ${}^1H_{IS}$ is not comparable to the microsatellite F_{IS} value – both with respect to direction and magnitude (Table 2).

Penguin Data

MHC sequence data were collected for three populations of little penguins (*Eudyptula minor*) in Western Australia (Vardeh, 2015). Results from ${}^1H_{IS}$ calculations, along with F_{IS} results from microsatellite data, are tabulated in Table 3. Each individual penguin had relatively little diversity of variants ($\overline{{}^{\mathbf{1}}\boldsymbol{H_I}}$ in Table 3). In contrast the populations showed a relatively large amount of diversity of MHC variants across individuals (${}^{\mathbf{1}}\boldsymbol{H_S}$ in Table 3). F_{IS} values based on microsatellite data agree with the sign of the MHC ${}^1H_{IS}$ values for the same population (Table 3), and both estimates indicate a heterozygote deficit. Notably, results for the ALB and ESP populations gave ${}^1H_{IS}$ values that are at least an order of magnitude larger than F_{IS}, although both F_{IS} and ${}^1H_{IS}$ suggested that the populations have a deficit of heterozygotes.

Discussion

F_{IS} is an important tool for the management and investigation of a population's genetic structure and adaptation, so it will be useful that our ${}^1H_{IS}$ method can overcome the limitations of F_{IS} on multi-gene families, such as MHC genes. On the basis of simulations and real data of natural populations ${}^1H_{IS}$ showed a strong relationship to F_{IS} (Figures 2-7; Tables 2 and 3), making ${}^1H_{IS}$ a useful tool for analysing mating patterns and selection in data from multi-locus gene families or polyploid species, when conventional F_{IS} is unable to be calculated. Simulations showed that Equation 4b worked well under a wide variety of mating structures and selection parameters. The fit between ${}^1H_{IS}$ and F_{IS} was good irrespective of the number of loci and the evenness of variants (Figures 3 and 4). However, it is worth noting that in our simulations, the number of loci and the evenness of variants strongly affected the range of values of ${}^1H_{IS}$ and F_{IS}. When there were only three loci, both F_{IS} and ${}^1H_{IS}$ did not go below ~-0.5 (Figure 3). This may be due to the selection scheme in our simulation, which implemented selection only during selection of parents, and not through offspring survival. The result was generated from a single generation of random mating without selection at the end of the simulation, which would bring F_{IS} towards zero. Compared to treatments with three loci, treatments with ten loci would usually have a wider range of F_{IS} and ${}^1H_{IS}$ values, and so some ten-locus replicates would maintain their low F_{IS} and ${}^1H_{IS}$ values, whereas this is less likely to happen with three-locus treatments (Figure 3). The ${}^1H_{IS}$ to F_{IS} comparison showed slightly more favourable regression result and lower RMSE when the variant distribution was 'Uneven', as well as showing a slightly better fit to the expected 1:1 (45°) line (Figure 4). This is likely partly due to 'Uneven' treatments having

480 the full range of -1 to +1 F_{IS}, and 'Even' treatments not going below -0.5 F_{IS}. This restricted
481 range could be because uneven allele distribution could give a wider range of F_{IS} values, as a
482 result of both H_e and H_o being very small, so that slight deviations could make a large
483 change in equation 4b, resulting in the full range of values from -1 to +1. Figure 5 shows
484 that relationship between F_{IS} and ${}^1H_{IS}$ gave high R-squared values for all generation times
485 trialled, although slightly better at shorter generation times (30 and 10). However, it should
486 be noted that despite the greater scatter, the departure from the expected 1:1 (45°) line
487 decreased as generation time increased in Figure 5. As generation time in our simulation
488 treatments increased, there was the potential for genetic drift to alter genotype
489 proportions, including creation of fixed loci with no variants. This would lower Shannon's
490 information ${}^1H_S$ (O'Reilly *et al,* 2020) as well as H_e , because $Max\ {}^1H_S = Ln(V)$, and
491 $Max\quad H_e = 1 - 1/V$ where V is the number of genetic variants in the population. Values
492 of ${}^1H_I$ and H_o would be secondarily restricted. It is unclear why this would cause slightly
493 better agreement between-F_{IS} and ${}^1H_{IS}$.

494 When the population size was 400 both F_{IS} and ${}^1H_{IS}$ values tended to cluster within
495 treatments (low variance of F_{IS} and ${}^1H_{IS}$ in 'small families' treatments under different
496 'generation time' treatments), as well as having a lower range of F_{IS} values (Figure 6). Again,
497 we believe this is due to the variance in population demographics being lessened in a larger
498 population size (Hedrick, 1994). This would explain why these population size of 400
499 treatments clustered within treatments and did not extend into more negative F_{IS} values.

500 An extreme result of drift is fixation of one or more loci. Population wide fixation is easy to
501 observe without any sophisticated methods because there would be zero genetic diversity,

so ${}^1H_S$ and He are zero therefore and ${}^1H_{IS}$ and F_{IS} are undefined (Table 4, first row). It is unlikely that a researcher would be interested in calculating either statistic from such data. A more subtle situation where ${}^1H_{IS}$ will give inaccurate results is when there is locus-specific fixation, which occurs when each different locus is fixed for different variants (Table 4, second row). When applied to a dataset with such a fixation pattern, F_{IS} would again be undefined, whereas ${}^1H_{IS}$ will give negative values. Because ${}^1H_{IS}$ is not locus-specific, it will not be able to detect such a pattern of fixation, and will instead interpret the individuals to be maximally diverse across the population, and assume a that some form of selection or demographic process is driving that diversity to give a negative ${}^1H_{IS}$ value. A very extreme case of locus-specific fixation (where every single locus is completely fixed across the population, as in Table 4, second row) can be detected by looking at variance of ${}^1H_I$ across the population, because it will be zero in such a case. But the more insidious cases, where say half the loci are fixed, can be very difficult to detect, and would give ${}^1H_{IS}$ a negative bias on such datasets. While in our study we removed values with low variance of ${}^1H_I$, ${}^1H_{IS}$ did actually work well in some instances where ${}^1H_I$ was 0. When F_{IS}=-1, and variance of ${}^1H_I$ was 0, ${}^1H_{IS}$ did tend to give values around the -0.9 range. However, these data, while showing correct results, were filtered out of our final dataset based on the criteria set out in the methods section. This situation only occurred in ~0.003% of our simulations and seems to only be the case when every individual in the population has the exact same heterozygote genotype.

There are several reasons for caution when estimating number of loci, but there are appropriate steps to take to help minimise these factors. Firstly, as mentioned above, the

524 estimation of number of loci assumes that the sample will have at least one individual
525 possessing a singleton variant, so that the number of loci can be calculated for Equation 4b.
526 This requirement can cause Equation 4b to not be accurate in situations where the data
527 would not expect to have singletons, such as a population with few, but equally abundant
528 variant sequences and many loci. This makes sense, as it is less likely for any individual to
529 have a singleton variant, if their genotype is dominated by 1 or 2 variants across many loci.
530 Secondly, NGS data also poses some problems with accuracy of our locus-number
531 estimation. Due to the stochasticity of NGS, it is not always going to output the correct allele
532 proportions needed to give an accurate estimate.

533 While our locus-number estimate is helpful if there is no prior data on a study population,
534 we would not recommend our one-individual method as a substitute to a robust
535 independent investigation of number of loci in a species. The one-individual method for
536 locus estimation was sensitive to misreads, however the mode and mean loci estimation
537 methods were far more robust to the stochasticity of real NGS data.

538 Assessing the dolphin and penguin results is difficult, because there may be different
539 selective and demographic pressures on MHC genes used to measure $^1H_{IS}$, compared to
540 the microsatellites used to measure F_{IS}. Therefore, the differences in $^1H_{IS}$ and F_{IS} results in
541 Table 2 and Table 3 could be explained either by error or by selection. If the selective
542 differences are the cause of deviation, it would imply that selection is driving MHC I diversity
543 into more extreme heterozygote deficits in SB (or less extreme heterozygote excess in BB)
544 than those for the presumably nearly-neutral microsatellites which may only be responding
545 to mild inbreeding. This would require further investigation to identify such selective

546 pressures on MHC I. However, such an interpretation is strengthened by the disagreement
547 of the two F_{IS} values for SB: the microsatellite F_{IS} and the MHC II DBQ F_{IS} (Table 2). These
548 values are likely due to different selective pressures acting on microsatellites and MHC II,
549 with the microsatellites (and nearby linked genes) possibly being neutral, affected only by
550 inbreeding, while the MHC II may have been subject to selection that favoured
551 heterozygotes.

552 Between the dolphin populations, the main demographic difference was in population size,
553 with SB having ~3000 individuals, and BB having ~250 individuals (Manlik *et al*, 2016).
554 Although the BB population is smaller, it is also more open to immigration from other
555 populations, which is thought to have increased in recent generations (Manlik *et al*, 2019a),
556 so it is reasonable that both F_{IS} and ${}^{1}H_{IS}$ are close to zero. It is known that some
557 inbreeding occurs in Shark Bay (Frère, 2010), so positive F_{IS} there is as expected for the
558 microsatellites. Again, the higher positive ${}^{1}H_{IS}$, compared to F_{IS}, could be due to selection
559 on MHC or to error of the method. There could possibly be selective effects acting on MHC I,
560 which would have to be against MHC heterozygotes to elevate the apparent heterozygote
561 deficit in MHC relative to microsatellites; selection for and against MHC heterozygotes is
562 known in other species (Sommer, 2005). Alternatively, there may have been mis-estimation
563 of the number of loci. Once again, Manlik was attempting to amplify a single MHC locus in
564 MHC I but could not confirm that this was achieved (Manlik, 2016). Possibly we have under-
565 estimated the number of loci, which would have depressed our estimated value of ${}^{1}H_{IS}$.
566 Nevertheless, the direction of departure from random mating is consistent across the two
567 dolphin populations, except when using the One Individual estimate for number of loci in the

BB dolphin population, which gave a positive ${}^1H_{IS}$ value, where the F_{IS} value was negative using microsatellite data; or with the MHC II DBQ dataset. It is also known that the MHC II DBQ nucleotide diversity in SB is very high compared to BB (Manlik *et al,* 2019b), so if there is a selective pressure for heterozygotes or outbreeding, this could help explain MHC II DBQs negative F_{IS} value because heterozygotes would tend to be maintained in the population.

Compared to the dolphins, the penguins also show that the direction of departure from random mating is consistent across the populations, but the penguins show an alternative explanation for the difference in magnitude of the ${}^1H_{IS}$ and F_{IS} values. Our analysis of the penguin dataset may have encountered the limitation described above, of difficulty of identifying true singletons for the locus-number estimation. We tested methods from the literature that are designed to help with singleton estimation, however they were not useful in this case (Supplement S3). Once data were filtered, individual penguins showed very little diversity within them, and gave an estimate of number of loci of 1-4, which seems quite low for an MHC multi-locus gene family, although Vardeh (2015) was attempting to amplify a single member of the gene family. If our estimate of number of loci was too low, the value of equation 4b would be depressed, and if it were too high, the value would be elevated. Elevation seems more likely, given that ${}^1H_{IS}$ gave positive values (indicating an excess of homozygote loci) which agrees with the direction of heterozygote deficit indicated by the microsatellite dataset, but with much greater magnitude. However, this elevation may not have come from our locus-number estimate, as for two of the three of the locus-number estimations used in Table 3, the estimate was one locus, therefore not possibly being an underestimation.

590 The use of ${}^{1}H_{IS}$ unlocks potential for evolutionary and ecological studies investigating
591 mating structure or selection using current and old data sets on multi-locus gene families,
592 especially of non-model species. This can augment traditional F_{IS} studies on single locus
593 genes. Thus, multi-locus gene family data sets can now be used to gain an understanding of
594 mating structure or selective pressures on these extremely important gene-families in wild
595 populations. Such conclusions could not only give historical context to the populations
596 studied, but also be used to guide future studies on related populations, especially in
597 conservation applications. The power of ${}^{1}H_{IS}$ comes from four possibilities:
598 1 - Researchers will be able to design studies that not only look at diversity in multi-locus
599 gene families but also analyse the mating structure/selective pressures on those same gene
600 families.

601 2 – Researchers will be able to more directly study specific multi-locus gene families that are
602 known to have an impact on mating and adaptation (such as MHC genes) and their
603 population wide effects.

604 3 – This method could be applied retrospectively to datasets collected before that method
605 existed, thus allowing researchers to utilise old MHC datasets to gain new insights into
606 previously studied populations.

607 4 – The new method is also directly applicable to cases where the entire genome is
608 replicated, such as polyploidy.

Acknowledgements

I would like to acknowledge the numerous people who have helped with the revision and editing process: Lee Ann Rollins, Alex Sentinella, Juliet Byrnes, Adriano Alarcón and David Dor. I would like to also Thank Rose "SuperShark" Hammer for drawing the dolphins in Figure 1.

References

Altizer, S., Harvell, D., & Friedle, E. (2003). Rapid evolutionary dynamics and disease threats to biodiversity. Trends in Ecology & Evolution, 18(11), 589-596.

Babik, W., Taberlet, P., Ejsmond, M. J., & Radwan, J. (2009). New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular ecology resources*, *9*(3), 713-719.

Babik, W. (2010). Methods for MHC genotyping in non-model vertebrates. *Molecular Ecology Resources*, *10*(2), 237-251.

Bowen, L., Aldridge, B. M., Gulland, F., Van Bonn, W., DeLong, R., Melin, S., ... & Johnson, M. L. (2004). Class II multiformity generated by variable MHC-DRB region configurations in the California sea lion (Zalophus californianus). *Immunogenetics*, *56*(1), 12-27.

Chao, A., & Jost, L. (2015). Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution*, *6*(8), 873-882.

Ellis, S. A., Morrison, W. I., MacHugh, N. D., Birch, J., Burrells, A., & Stear, M. J. (2005). Serological and molecular diversity in the cattle MHC class I region. Immunogenetics, 57(8), 601-606.

630 Flores-Ramirez, S., Urban-Ramirez, J., & Miller, R. D. (2000). Major histocompatibility
631 complex class I loci from the gray whale (Eschrichtius robustus). *Journal of Heredity*, *91*(4),
632 279-282.

633 Frère, C. H., Krützen, M., Kopps, A. M., Ward, P., Mann, J., & Sherwin, W. B. (2010).
634 Inbreeding tolerance and fitness costs in wild bottlenose dolphins. *Proceedings of the Royal*
635 *Society B: Biological Sciences*, *277*(1694), 2667-2673.

636 Halliburton R (2004) Introduction to population genetics. Pearson/Prentice Hall, Upper
637 Saddle River

638 Hedrick, P. W. (1994). Evolutionary genetics of the major histocompatibility complex. *The*
639 *American Naturalist*, *143*(6), 945-964.

640 Heimeier, D., Baker, C. S., Russell, K., Duignan, P. J., Hutt, A., & Stone, G. S. (2009). Confirmed
641 expression of MHC class I and class II genes in the New Zealand endemic Hector's dolphin
642 (Cephalorhynchus hectori). *Marine Mammal Science*, *25*(1), 68-90.

643 Kalbe, M., Eizaguirre, C., Dankert, I., Reusch, T. B., Sommerfeld, R. D., Wegner, K. M., &
644 Milinski, M. (2009). Lifetime reproductive success is maximized with optimal major
645 histocompatibility complex diversity. *Proceedings of the Royal Society B: Biological*
646 *Sciences*, *276*(1658), 925-934.

647 Klein, J. (1986). *Natural history of the major histocompatibility complex*. Wiley.

648 Manlik, O (2016) Fitness & major histocompatibility complex diversity of two bottlenose
649 dolphin populations. Ph.D Thesis, University of New South Walkes, Sydney

Manlik, O., Krützen, M., Kopps, A. M., Mann, J., Bejder, L., Allen, S. J., ... & Sherwin, W. B. (2019a). Is MHC diversity a better marker for conservation than neutral genetic diversity? A case study of two contrasting dolphin populations. *Ecology and evolution*, *9*(12), 6986-6998.

Manlik, O., Chabanne, D., Daniel, C., Bejder, L., Allen, S. J., & Sherwin, W. B. (2019b). Demography and genetics suggest reversal of dolphin source-sink dynamics, with implications for conservation. *Marine Mammal Science*, *35*(3), 732-759.

O'Reilly, G. D., Jabot, F., Gunn, M. R., & Sherwin, W. B. (2020). Predicting Shannon's information for genes in finite populations: new uses for old equations. Conservation Genetics Resources, 12(2), 245-255.

Peng, B., & Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. Bioinformatics, 21(18), 3686-3687.

Qin, L. X., Tuschl, T., & Singer, S. (2016). Empirical insights into the stochasticity of small RNA sequencing. *Scientific reports*, *6*(1), 1-8.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... & Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, *75*(23), 7537-7541.

Sepil, I., Lachish, S., & Sheldon, B. C. (2013). MHC-linked survival and lifetime reproductive success in a wild population of great tits. *Molecular ecology*, *22*(2), 384-396.

Shannon, C. E. (1949). Communication theory of secrecy systems. *The Bell system technical journal*, *28*(4), 656-715.

671 Sherwin, W. B., Chao, A., Jost, L., & Smouse, P. E. (2017). Information theory broadens the
672 spectrum of molecular ecology and evolution. *Trends in ecology & evolution*, *32*(12), 948-
673 963.

674 Siddle, H. V., Marzec, J., Cheng, Y., Jones, M., & Belov, K. (2010). MHC gene copy number
675 variation in Tasmanian devils: implications for the spread of a contagious
676 cancer. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1690), 2001-2006.

677 Smith, D. P., & Peay, K. G. (2014). Sequence depth, not PCR replication, improves ecological
678 inference from next generation DNA sequencing. *PloS one*, *9*(2), e90234.

679 Sommer, S. (2005). The importance of immune gene variability (MHC) in evolutionary
680 ecology and conservation. *Frontiers in zoology*, *2*(1), 1-18.

681 Sommer, S., Courtiol, A., & Mazzoni, C. J. (2013). MHC genotyping of non-model organisms
682 using next-generation sequencing: a new methodology to deal with artefacts and allelic
683 dropout. *BMC genomics*, *14*(1), 1-17.

684 Stervander, M., Dierickx, E. G., Thorley, J., Brooke, M. D. L., & Westerdahl, H. (2020). High
685 MHC gene copy number maintains diversity despite homozygosity in a Critically Endangered
686 single-island endemic bird, but no evidence of MHC-based mate choice. *Molecular*
687 *ecology*, *29*(19), 3578-3592.

688 Stuglik, M. T., Radwan, J., & Babik, W. (2011). jMHC: Software assistant for multilocus
689 genotyping of gene families using next-generation amplicon sequencing. *Molecular ecology*
690 *resources*, *11*(4), 739-742.

691 Thoss, M., Ilmonen, P., Musolf, K., & Penn, D. J. (2011). Major histocompatibility complex
692 heterozygosity enhances reproductive success. *Molecular Ecology*, *20*(7), 1546-1557.

693 Vardeh, S (2015) Population Genetics, demography and population viability of little penguins
694 (*eudyptula minor*) in Australia. Ph.D Thesis, University of New South Walkes, Sydney

695 Yamazaki, K., Beauchamp, G. K., Kupniewski, D., Bard, J., Thomas, L., & Boyse, E. A. (1988).
696 Familial imprinting determines H-2 selective mating preferences. *Science*, *240*(4857), 1331-
697 1332.

698 Zagalska-Neubauer, M., Babik, W., Stuglik, M., Gustafsson, L., Cichoń, M., & Radwan, J.
699 (2010). 454 sequencing reveals extreme complexity of the class II Major Histocompatibility
700 Complex in the collared flycatcher. *BMC evolutionary biology*, *10*(1), 395.

701 Data accessibility

702 Simulation code to generate data as well as data generated from our simulation run (which the
703 results in this paper are based off) are available at the following GitHub repository:
704 https://github.com/GubbaFlubba/Detecting-non-random-mating-or-selection-in-natural-
705 populations-using-multi-locus-gene-families

706 Author Contributions

707 Gabe O'Reilly: Wrote the paper, programmed the simulations, developed the method,
708 analysed the data.

709 Oliver Manlik: Primary contributor for all the dolphin data.

710 Sandra Vardeh: Primary contributor for all the penguin data.

Zack P Lawler: Assisted with testing methodology.

Jennifer Sinclair: Contributor for the penguin data.

William Sherwin: supervised the research and provided input at every step of its development.

All authors contributed to revisions and editing the paper.

Tables and Figures

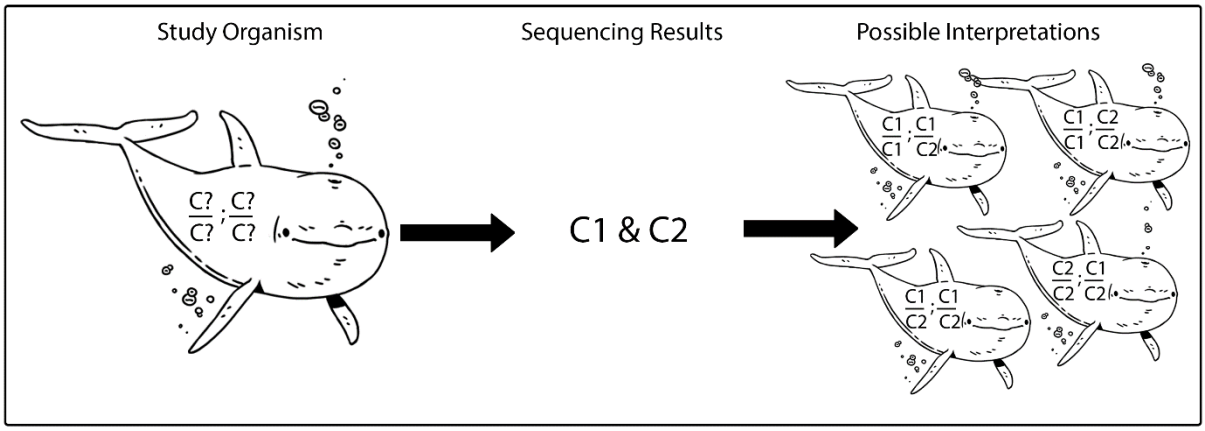

Figure 1: showing the results from sequencing a multi-locus gene family of two unlinked loci for an individual (variants C1 and C2), and the ambiguity those results can give.

	Variants:				Sum	Estimated number of loci
	B1	*B2*	*B3*	*B4*		
Individual 1:	1	1	2	1	=5	3
Individual 2:	1	4	3	2	=10	5
Individual 3:	1	0	1	0	=2	1
Individual 4:	1	0	1	3	=5	3

Table 1: Example data set on a population with four individuals showing relative proportions of variants as ratios. The ratios are adjusted so that the least frequent variant is represented by "1", then summed for each individual, the highest value among these summed values is rounded to the next even number then divided by two to gain a minimum estimate of number of loci. In the data shown for individual 1, five is divided by two then rounded up to give minimum $L \geq 3$. This rationale is explained in the main text.

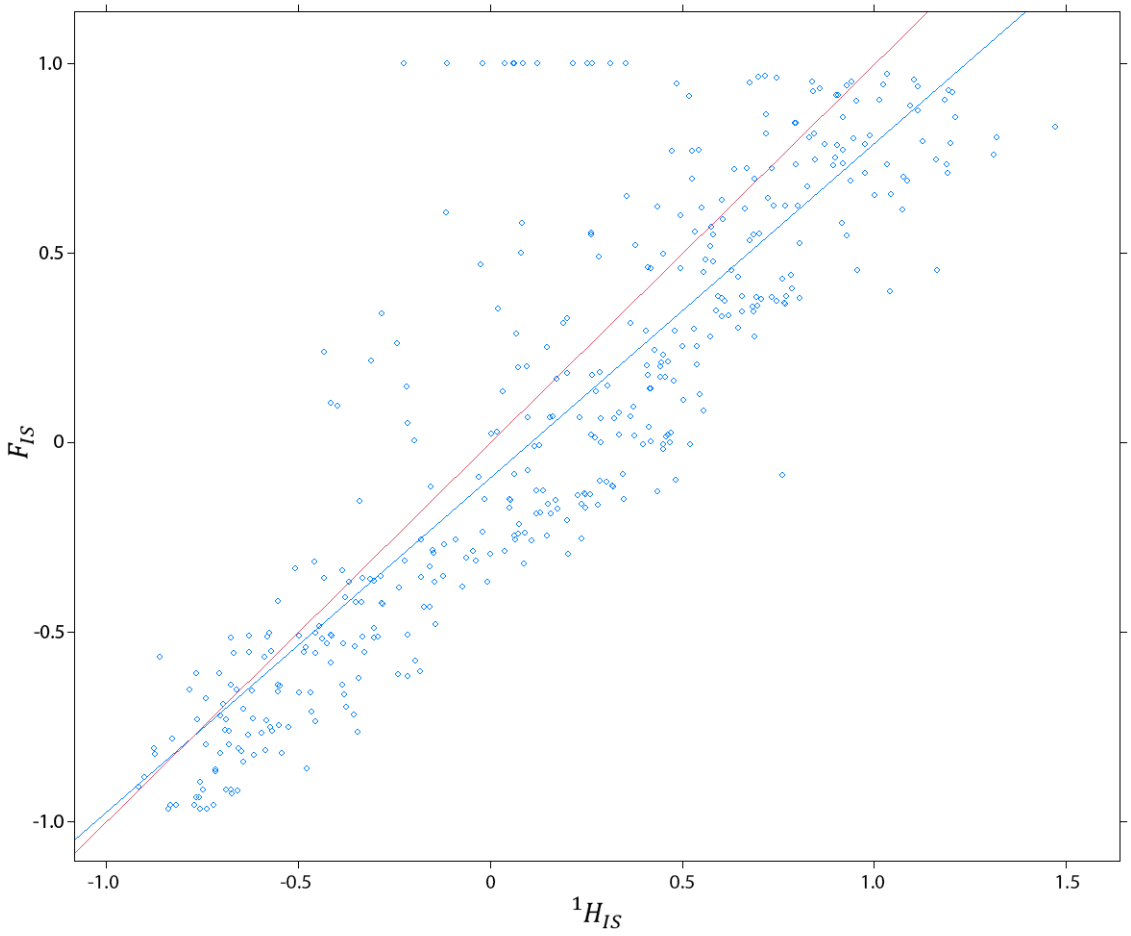

Figure 2: Regression of F_{IS} on $^1H_{IS}$ in simulated data that has had replicates with low ^1H_I variance removed. F_{IS} ranges were manipulated via 'mating' and 'selection' treatment parameters shown in the methods section. The total binned data, with all treatments together are shown. Blue line indicates a regression slope, the Red line indicates the expected 1:1 slope for perfect agreement between the methods. Regression analysis showed an R-squared of 0.756, p = < 0.001 and RMSE= 0.398. Non-binned data can be found in the supplement (Figure S5).

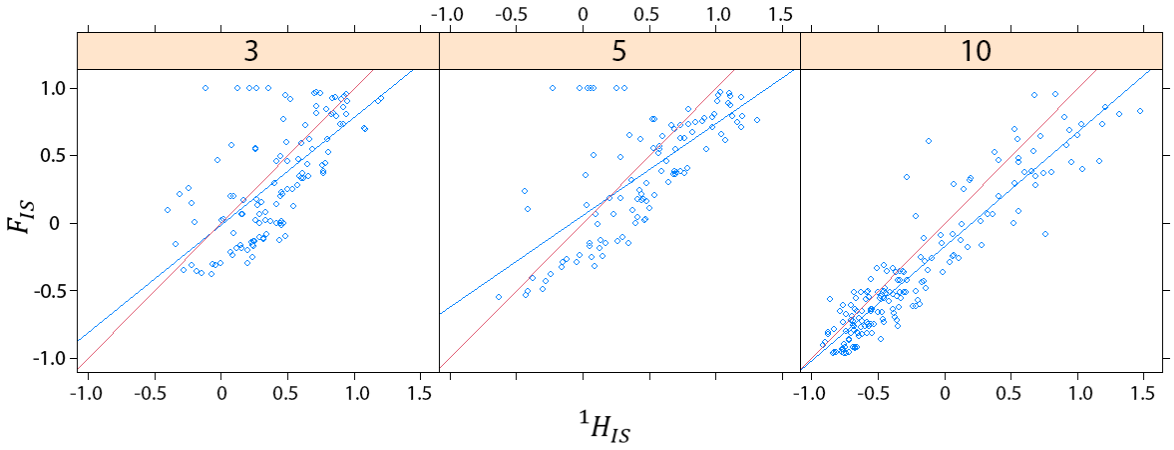

Figure 3: How number of loci affects the regression of F_{IS} on $^1H_{IS}$. Comparison of $^1H_{IS}$ results to their corresponding F_{IS} results from simulated binned data that has had replicates with low ^1H_I variance removed. The F_{IS}ranges were manipulated via 'mating' and 'selection' treatment parameters shown in the methods section. The three panels show treatments with differing numbers of loci set up in the simulation, indicated above in each panel. Blue line indicates a regression slope, the Red line indicates the expected 1:1 slope. In treatments with three loci, $^1H_{IS}$ showed an R-squared of 0.445, p-value = < 0.05 and RMSE = 0.334. In treatments with five loci, $^1H_{IS}$ showed an R-squared of 0.452, p-value = < 0.05 and RMSE = 0.368. In treatments with ten loci, $^1H_{IS}$ showed an R-squared of 0.861, p-value = < 0.05, and RMSE = 0.255. Non-binned data can be found in the supplement (Figure S6).

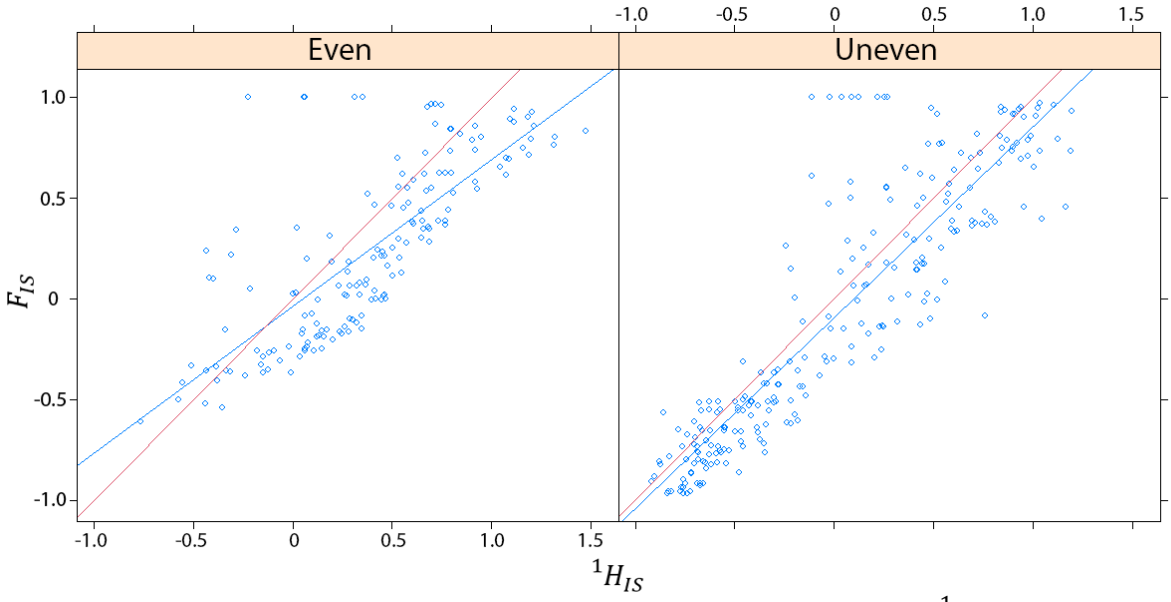

Figure 4: How allele variant distribution affects the regression of F_{IS} on $^1H_{IS}$. Comparison of $^1H_{IS}$ results to their corresponding F_{IS} results from simulated binned data that has had replicates with low ^1H_I variance removed. The F_{IS}ranges were manipulated via 'mating' and 'selection' treatment parameters shown in the methods section. The two panels show treatments with differing distribution of variants in the simulation, indicated above in each panel. Blue line indicates a regression slope, the Red line indicates the expected 1:1 slope. In treatments with an Even variant distribution, $^1H_{IS}$ showed an R-squared of 0.593, p-value = < 0.05, and RMSE = 0.333. In treatments with an Uneven variant distribution, $^1H_{IS}$ showed R-squared of 0.795, p-value = < 0.05, and RMSE = 0.300. Non-binned data can be found in the supplement (Figure S7).

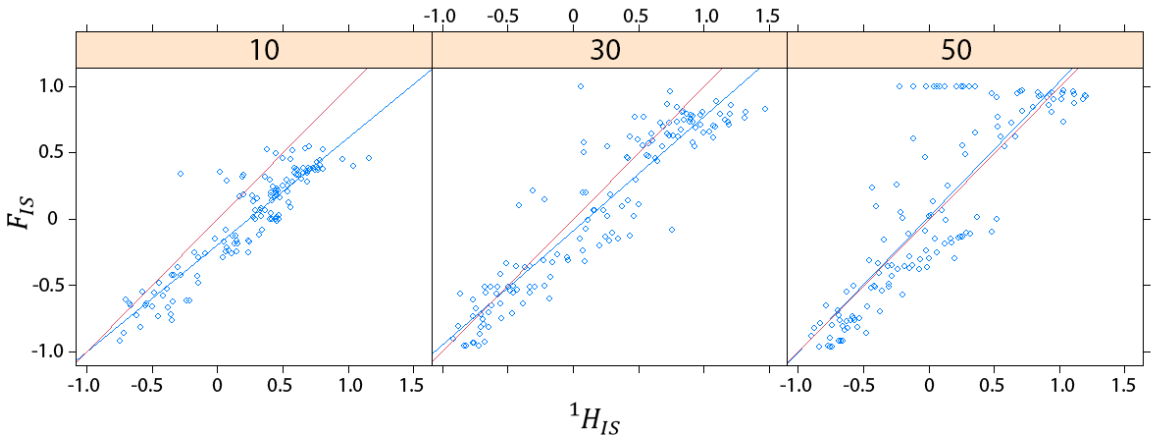

Figure 5: How number of generations simulated affects the regression of F_{IS} on $^1H_{IS}$ comparison. Comparison of $^1H_{IS}$ results to their corresponding F_{IS} results from simulated binned data that has had replicates with low ^1H_I variance removed. The F_{IS} ranges were manipulated via 'mating' and 'selection' treatment parameters shown in the methods section. The three panels show treatments with differing numbers of generations simulated, indicated above in each panel. Blue line indicates a regression slope, the Red line indicates the expected 1:1 slope. Ten-generation data had an r-squared of 0.827, p-value = < 0.05 and RMSE = 0.299. Thirty-generation data had an r-squared of 0.855, p-value = < 0.05, and RMSE = 0.273. Fifty-generation data had an r-squared of 0.723, p-value = < 0.05, and RMSE = 0.362. Non-binned data can be found in the supplement (Figure S8).

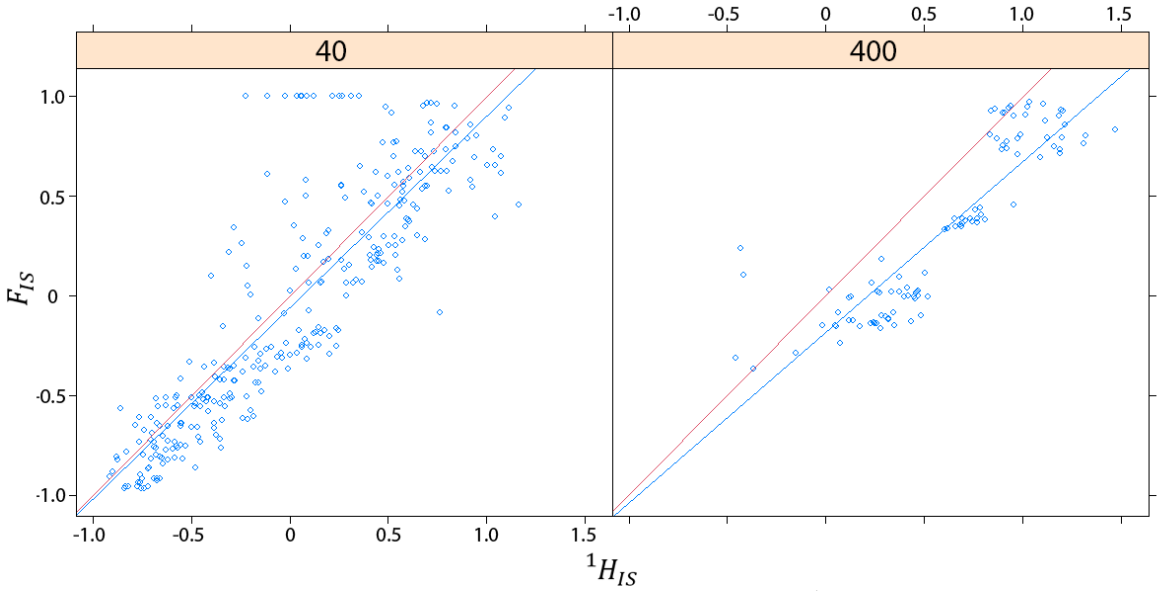

Figure 6: How population size affects the regression of F_{IS} on $^1H_{IS}$ comparison. Comparison of $^1H_{IS}$ results to their corresponding F_{IS} results from simulated binned data that has had replicates with low ^1H_I variance removed. The F_{IS} ranges were manipulated via 'mating' and 'selection' treatment parameters shown in the methods section. The two panels show treatments with differing population sizes in the simulations, indicated above in each panel. Blue line indicates a regression slope, the Red line indicates the expected 1:1 slope. In population sizes of 40, $^1H_{IS}$ showed an r-squared of 0.749, p-value = < 0.05, and RMSE = 0.304. In population sizes of 400, $^1H_{IS}$ an r-squared of 0.769, p-value = < 0.05, and RMSE = 0.340. In 400 population size treatments, there was a reduced range of F_{IS} values, from ~-0.5 to 1, whereas 40 population size showed the full range of F_{IS} values from -1 to 1. Non-binned data can be found in the supplement (Figure S9).

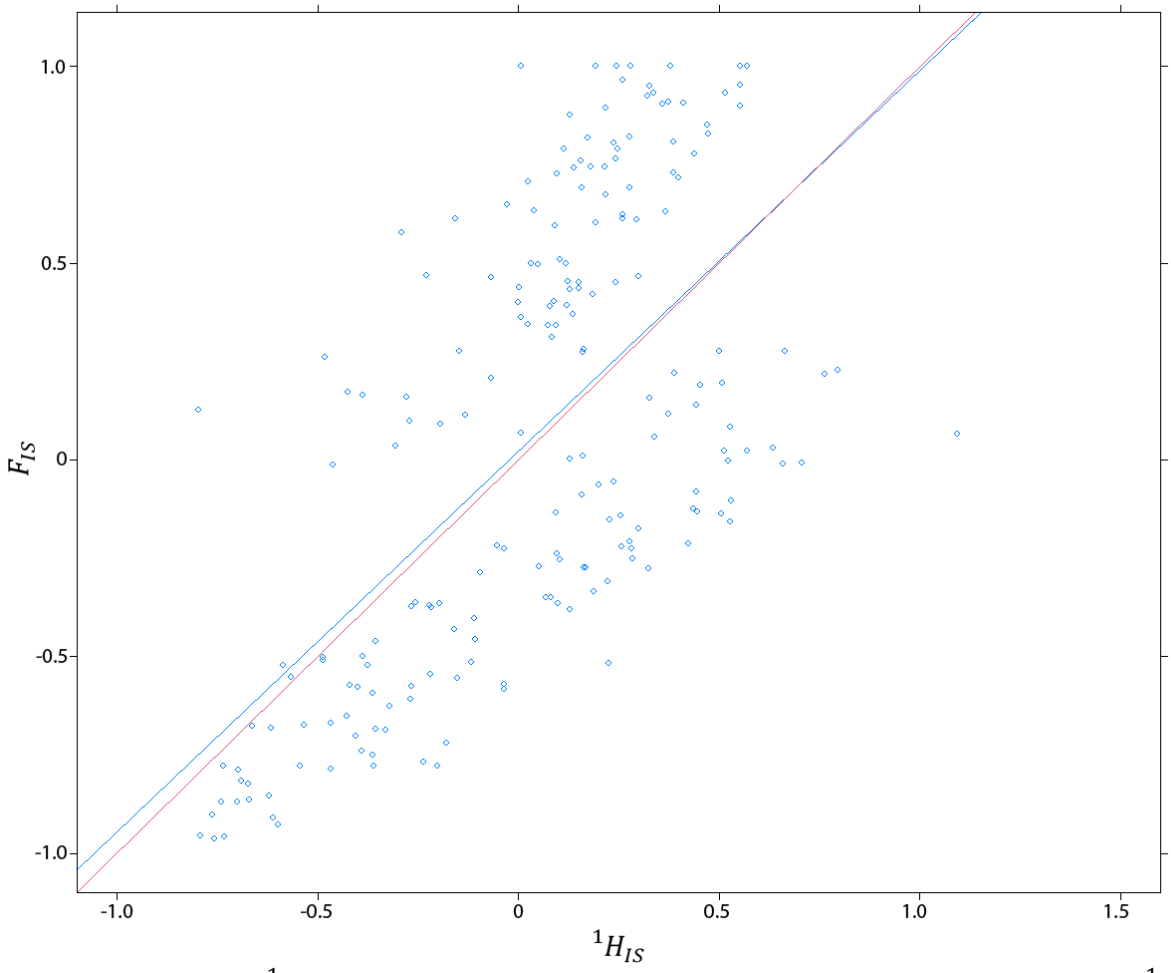

Figure 7: Results of $^1H_{IS}$ regressed against F_{IS} on binned NGS like data. Comparison of $^1H_{IS}$ results to their corresponding binned F_{IS} results from simulated NGS data that has had replicates with low ^1H_I variance removed. The F_{IS}ranges were manipulated via 'mating' and 'selection' treatment parameters shown in the methods section. Blue line indicates a regression slope, the Red line indicates the expected 1:1 slope. $^1H_{IS}$ showed an R-squared of 0.407, p-value = < 0.05, RMSE = 0.441. Non-binned data can be found in the supplement (Figure S10).

Population	MHC I sequences						Microsatellite data F_{IS}*	MHC II DQB Single locus F_{IS}
	Loci number Estimation Method	**Locus number Estimate** (non-rounded)	$\overline{^1H_I}$	^1H_S	E_s	$^1H_{IS}$		
Shark Bay (SB)	Average	6 (5.5)	1.787	2.371	0.746	0.540	0.0327	-0.024
	Mode	5 (4.5)				0.356		
	One Individual	9 (8.5)				1.091		
Bunbury (BB)	Average	3 (3.3)	1.119	1.48	0.617	-0.015	-0.0376	NA
	Mode	3 (3.0)				-0.015		
	One Individual	10 (10)				1.038		

*Table 2: Heterozygote deficit or excess in MHC variants and microsatellites in dolphin populations – Locus number estimates, $\overline{^1H_I}$ values, ^1H_S values, E_s values and $^1H_{IS}$ values for each population and locus number estimation method. * F_{IS} values are estimated from microsatellite data from the same populations (Manlik, 2016; Manlik et al, 2019a).*

Population	**Loci number Estimation Method**	**Locus number Estimate** (non-rounded)	$\overline{{}^{1}H_I}$	${}^{1}H_S$	E_s	${}^{1}H_{IS}$	F_{IS} (based on microsatellite data)*
Perth (PER)	Average	1 (1.36)	0.362	1.865	0.750	0.459	0.342
	Mode	1 (1)				0.459	
	One Individual	4 (3.54)				2.27	
Albany (ALB)	Average	1 (1)	0.411	2.279	0.888	0.568	0.001
	Mode	1 (1.02)				0.568	
	One Individual	3 (2.88)				2.02	
Esperance (ESP)	Average	1 (0.89)	0.246	2.345	0.889	0.702	0.093
	Mode	1 (1)				0.702	
	One Individual	3 (3.22)				2.29	

*Table 3: Heterozygote deficit or excess of MHC and microsatellites in penguin populations - Locus number estimates, $\overline{{}^{1}H_I}$ values, ${}^{1}H_S$ values and ${}^{1}H_{IS}$ values for each population and locus number estimation method. The MHC data were filtered to remove sequence reads that did not make up at least 10% of the sequence reads per individual. * F_{IS} values are estimated from microsatellite data from the same populations from Vardeh, (2015).*

Scenario	Genotype of every individual (with 4 loci)	${}^{1}H_{IS}$ result	F_{IS} result
Total Fixation	$\frac{C1}{C1};\frac{C1}{C1};\frac{C1}{C1};\frac{C1}{C1}$	Undefined	Undefined / 0
locus specific fixation	$\frac{C1}{C1};\frac{C2}{C2};\frac{C3}{C3};\frac{C4}{C4}$	-1	Undefined / 0

797 *Table 4: Two different scenarios for cases in which ${}^{1}H_{IS}$ would give inaccurate results.*