Unbiased population heterozygosity estimates from genome-wide sequence data

Thomas L Schmidt[1]*, Moshe Jasper[1], Andrew R Weeks[1,2], Ary A Hoffmann[1]

[1]School of BioSciences, Bio21 Institute, University of Melbourne, Parkville, Victoria, Australia
[2]cesar Pty Ltd, Parkville, Victoria, Australia
**Correspondence: Tom Schmidt (toms@unimelb.edu.au), Bio21 Institute, University of Melbourne, Parkville, Victoria, Australia*

Abstract

1. Genetic variability within populations is a key parameter for the management of threatened species and for tracking invasion history. Heterozygosity (observed and expected) is commonly used to represent genetic variability and is increasingly being estimated with single nucleotide polymorphism (SNP) markers. While many SNP markers can provide precise estimates of genetic processes, the results of 'downstream' analysis of these markers may depend heavily on 'upstream' filtering decisions.
2. Here we explore the downstream consequences of sample size, rare allele filtering, missing data thresholds and known population structure on estimates of heterozygosity using a ddRADseq dataset of the mosquito *Aedes aegypti* and a DArTseq dataset of a threatened grasshopper, *Keyacris scurra*.
3. We show that estimates based on polymorphic markers only (i.e. SNP heterozygosity) are always biased by sample size, regardless of other filtering considerations. By contrast, results are consistent across sample sizes when calculations consider monomorphic as well as polymorphic sequence information (i.e. genome-wide or autosomal heterozygosity). We also show that when loci with missing genotypes are included, observed and expected heterozygosity estimates diverge in proportion to the amount of missing data permitted at each locus.
4. To ensure robust results across studies, we make two recommendations for estimating heterozygosity: (i) autosomal heterozygosity should be reported instead of (or in addition to) SNP heterozygosity; (ii) sites with any missing data should be removed when calculating heterozygosity. Applying these protocols to *K. scurra,* we show that autosomal heterozygosity estimates are consistent even when populations with different sample sizes and high levels of differentiation are analysed together. This should facilitate comparisons across studies and between observed and expected measures of heterozygosity.

Key-words: Heterozygosity, Single Nucleotide Polymorphisms (SNPs), RADseq, DArTseq, Filtering, Conservation, Population Structure, Genetic Mixing

Introduction

The power provided by single nucleotide polymorphism (SNP) markers detected using genome-wide sequencing approaches is leading to their increased use in conservation genetic studies (Garner et al., 2016). Although reference genomes are currently rarely available for threatened species, SNP assessments based on reduced representation sequencing are popular for investigating levels of genetic differentiation among remnant populations and for comparing levels and patterns of genetic variation within populations (e.g. Maroso, Franch, Dalla Rovere, Arculeo, & Bargelloni, 2016; Campbell et al., 2019) which in turn relates to the adaptive potential of populations (Ørsted, Hoffmann, Sverrisdóttir, Nielsen, & Kristensen, 2019) as well as patterns of inbreeding and relatedness (Mulvena et al., 2020). Results from SNP studies are being interpreted for use in management decisions that include genetic rescue, genetic mixing and founder selection in threatened species programs (e.g. Fitzpatrick et al., 2020). The relative ease of generating SNP genotypes is leading to their increased use by non-specialists, particularly through the availability of companies such as Diversity Arrays Technology (https://www.diversityarrays.com), which provide SNP genotypes through customised in-house processes (Gruber, Unmack, Berry, & Georges, 2017; Wright et al., 2019; Mulvena et al., 2020).

For these reasons, it is important to be aware of any biases inherent in the application of SNP markers in conservation genetics and in other applications. While potential biases have been considered for the detection of structure between populations (Gruber et al. 2017; Linck & Battey, 2019; Wright et al., 2019), there has been less focus on the estimation of genetic variability within populations. These estimates are important because they link to the evolutionary potential of populations, which is typically higher in populations with greater genetic variability (Hoffmann, Sgrò, & Kristensen, 2017; Ørsted et al., 2019). Genetic variability of populations is therefore crucial when making genetic management decisions for threatened species (Weeks et al., 2011; Hoffmann et al., 2020).

Genetic variation in populations is measured in several ways, the most common of which has traditionally been the proportion of polymorphic loci and patterns of observed and expected heterozygosity. Observed heterozygosity estimates are typically based on a substantial number of individuals being sampled from a population, but with large quantities of sequence data it may be possible to reduce the number of individuals needed to get meaningful comparisons across populations (Nazareno, Bemmels, Dick, & Lohmann, 2017). Sequence data can be used to identify average levels of heterozygosity assuming the apparent diversity at loci is not related to errors introduced during sequencing or genotyping, the latter of which requires adequate sequencing coverage to ensure that both strands of a diploid individual are sequenced (Nielsen, Paul, Albrechtsen, & Song, 2011).

For heterozygosity, h_i, the observed heterozygosity at locus *i* can be averaged for an individual across n loci as $\frac{\sum_{i=1}^{n} h_{i...n}}{n}$. This can be calculated from variation at polymorphic nucleotides only within loci (i.e. SNP heterozygosity) or at both polymorphic and monomorphic nucleotides (i.e. genome-wide or autosomal heterozygosity). Both SNP heterozygosity and autosomal heterozygosity appear in the literature; most population-focussed studies tend to focus on polymorphic markers (Jones et al., 2012; Chen, Cosgrove, Bowman, Fitzpatrick, & Clark, 2016; Bock, Kantar, Caseys, Matthey-Doret, & Rieseberg, 2018; Mathur, Tomeček, Heniff, Luna, & DeWoody, 2019; Surbakti et al., 2020) although others focus on autosomal heterozygosity (e.g. Hohenlohe et al., 2010), which is the parameter reported in studies comparing individual genomes (e.g. Gopalakrishnan et al., 2017; Westbury, Petersen, Garde, Heide-Jørgensen, & Lorenzen, 2019). It is not always clear whether the selected parameter is appropriate for a particular study, and the two parameters cannot be compared given that SNP heterozygosities will be orders of magnitude larger than autosomal heterozygosities. Moreover, SNP or autosomal heterozygosities derived from reduced representation sequencing may not be comparable across studies, given that they depend on the nature of the regions sequenced and (particularly in the case of SNP heterozygosity) the processes of identifying polymorphisms.

Observed heterozygosity of individuals in a population depends on two components, the amount of genetic variation and the level of inbreeding, which results in homozygosity (Ritland, 1996). The second parameter can be estimated when allele frequencies at polymorphic loci are known so that the observed heterozygosity at polymorphic loci can be compared to the expected heterozygosity, with the latter expected to be relatively higher when there is inbreeding. The accuracy whereby allele frequencies and hence inbreeding is estimated will depend on the number of individuals in a population that have been sampled and number of loci, with nucleotide diversity potentially providing good estimates of expected heterozygosity (Hohenlohe et al., 2010).

Apart from inbreeding, there are other reasons why the observed heterozygosity may be lower than the expected heterozygosity including sex linkage, deletion polymorphisms (resulting in null alleles) and the Wahlund effect (De Meeûs, 2018). Sex linkage effects can be eliminated if sexes have been distinguished prior to analysis particularly if the locations of loci on the genome is known (Coltman, Pilkington, Smith, & Pemberton, 1999). Wahlund effects can be determined through traditional population genetic analyses as they generally display a concordant pattern across all loci (Waples & Allendorf, 2015). However, deletion polymorphisms are difficult to detect unless whole genome data are available.

Sample size, rare allele filtering, missing data thresholds and structuring among populations typically vary from study to study. Our aim here is to explore how this variation influences SNP-based estimates of observed and expected heterozygosity, which could make comparisons across and even within studies difficult. We explore these questions with a pair of genome-wide datasets of the sort frequently used for assessing variation in wild

populations. We focus initially on a ddRADseq dataset from one population of a common species and then consider a DArTseq dataset from threatened species that covers multiple populations. We make some recommendations for assessing heterozygosity when study aims include comparisons of genetic variability across populations and with other studies.

Materials and Methods

Sequence data from the same population

To investigate the challenges in comparing heterozygosity estimates across populations and across studies, we start by considering a single, well-mixed population. We use double digest restriction-site associated (ddRAD) sequence data obtained from female *Aedes aegypti* mosquitoes sampled from a 0.125 km^2 area of Kuala Lumpur, Malaysia (Jasper, Schmidt, Ahmad, Sinkins, & Hoffmann, 2019). Note that as this ddRADseq dataset contains only females, and as *Ae. aegypti* mosquitoes do not have definable sex chromosomes but rather a small sex-determining region (Fontaine et al., 2017), we did not need to filter out genotypes at sex chromosomes.

We took subsamples from this population as follows.

Ten subsamples: For each of the following sample sizes (n = 10, 5, 4, 3, 2), ten random subsamples were taken from the 100 individuals without replacement. These subsamples were used to assess variation in heterozygosity estimates at each n.

Nested subsamples: The 100 individuals were randomly assigned to two groups, A and B, each of n = 50. Group A and Group B were then subsampled once at each of the following sample sizes (n = 50, 40, 30, 20, 10, 5), but where each subsample could only include individuals that were included at the next highest n. For instance, the subsample at n = 30 could only contain individuals that were present in the n = 40 subsample. Nested subsampling served to reduce variation among subsamples of different size from the same group. These subsamples were used to assess variation in heterozygosity estimates at $5 \leq n \leq 50$.

Sequence data from multiple populations

We consider the issue of multiple populations being included in a comparison by reanalysing a set of four populations of *Keyacris scurra* (Key's Matchstick Grasshopper) taken from a larger set of sequencing data derived from a Diversity Arrays Technology (DArT) approach. *Keyacris scurra* has recently been listed as endangered and is currently restricted in range to refugia in south-eastern Australia. These four populations have experienced very low gene

flow and are highly differentiated (pairwise F_{ST} = 0.14–0.28). The four populations were processed and sequenced together as part of the same project (Hoffmann et al., 2020). Note that no reference assembly is available for this species so the term "autosomal heterozygosity" here will also include sequence data from any differentiated sex chromosomes (or regions of chromosomes).

Sequence processing

For the ddRADseq dataset, aligned sequences were built into a Stacks.v2 (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) catalog with the program ref_map. For the DArTSeq dataset, sequence data were built into a *de novo* Stacks catalog using the program denovo_map, allowing for up to four mismatches within and between individuals. We analysed both datasets with the Stacks program "Populations", which was used to estimate observed and expected heterozygosity for a range of filtering settings described below.

Results

Estimates based on polymorphic loci (SNP heterozygosity)

Our first aim was to see how a dataset filtered with settings typically used for assessing genetic structure (i.e. variation between populations) might perform when used to estimate heterozygosity (i.e. variation within populations). Analysis of genetic structure will usually consider only polymorphic loci (SNPs). When filtering SNPs, the normal approach is to combine the entire data set, remove loci not genotyped in a sufficient number of individuals (typically 70-95%), and then filter out loci with a minor allele frequency (MAF) or minor allele count (MAC) that is not met globally (e.g. Lemopoulos et al., 2019; Mathur et al., 2019; Mulvena et al., 2020). Simulations suggest that a MAC ≥ 3 is optimal for detecting population structure (Linck & Battey, 2019), to avoid bias from singletons.

Population comparisons using filtered polymorphic loci: effects of sample size

We start with a simple comparison of how global (N) and local (n) sample size affects SNP heterozygosity estimates. For this we use the ten subsamples (n = 10, 5, 4, 3, 2), which we analyse first individually (i.e. with each subsample run in a separate analysis) and then together (i.e. where the ten subsamples are run in a single analysis). Filtering followed a standard approach, retaining a single SNP locus from each RADtag (--write-single-snp) which had no more than 20% missing data and that had a MAC ≥ 3. Observed and expected heterozygosities were estimated from these filtered polymorphic loci.

SNP heterozygosity estimates are shown in Figs 1a-d and indicate how this type of filtering approach presents problems for comparing estimates across studies. In all cases, heterozygosity was larger when fewer samples were used for estimation. Observed and expected heterozygosity estimates are both strongly affected by sample size, and effects are stronger when subsamples are analysed individually.

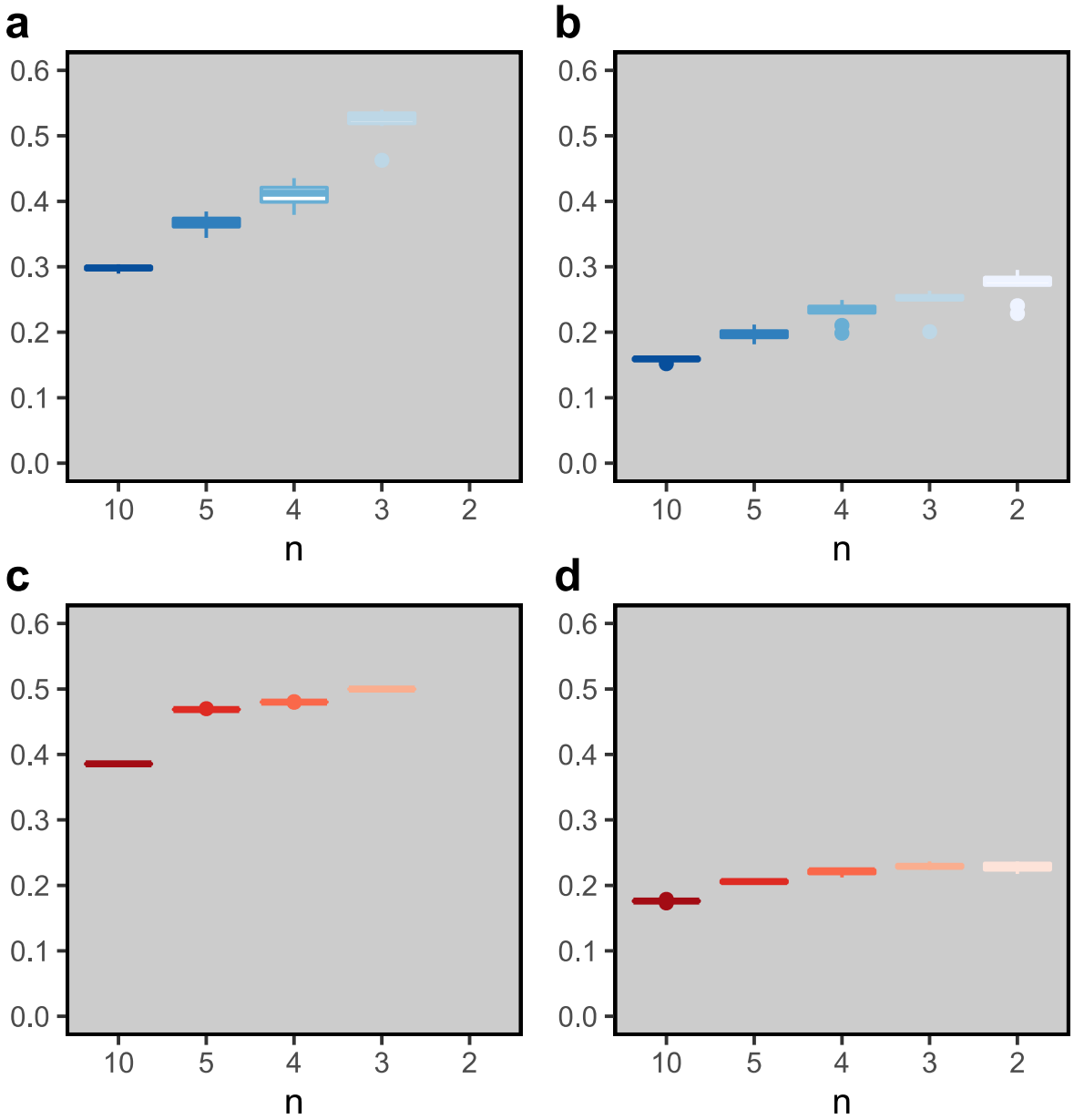

Fig. 1: Observed (blue) and expected (red) SNP heterozygosity estimated from ten subsamples of mosquitoes at each n, with MAC ≥ 3 filtering applied to the population subsamples individually (a, c) or together (b, d). Note that for MAC ≥ 3, individual analysis of n = 2 subsamples is not possible.

The source of this issue is that heterozygosity is generally lower for SNPs with rare alleles (for which most individuals are homozygous for the common allele) than for SNPs with multiple common alleles. For instance, for the n = 3 subsamples analysed individually (Fig. 1a, c) all SNPs have minor alleles at 0.5 frequency when MAC ≥ 3 is applied, leading to expected heterozygosity of 0.5. As additional samples are added, lower-frequency SNPs become increasingly more likely to be detected, leading to lower SNP heterozygosity estimates. When populations are analysed together, more SNPs with rare alleles are included due to global filtering, leading to lower estimates. However, even when populations were analysed together there were clear differences between SNP heterozygosity estimates for n = 10 and n < 10. These problems should diminish if singletons and doubletons are included because rare alleles in small populations are then more likely to be assessed as polymorphic, but they will nevertheless persist.

To investigate these effects at n ≥ 5, we use the nested subsamples from Groups A and B, first analysing each subsample from Group A individually, then analysing each pair of subsamples of equal n from A and B together (Fig. 2). Here we see that sample size bias is present throughout, and specifically shows a bias from global N (total sample size in the analysis) rather than local n (sample size of each population). Although these effects reduce as N increases, they are nevertheless evident even with samples of 30 or 40 from each population and N ≥ 80.

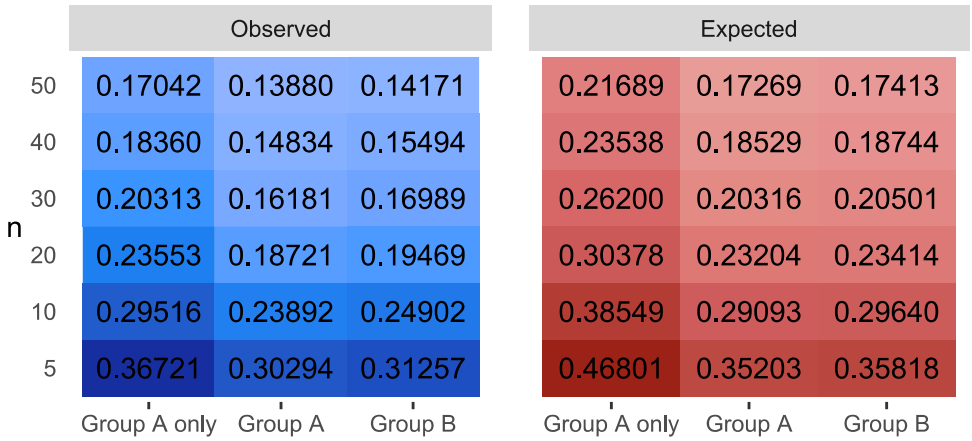

Fig. 2: Observed (blue) and expected (red) SNP heterozygosity estimated from mosquito samples of different size (n) and filtered at MAC ≥ 3. We compared estimates for Group A only with those of Group A and Group B combined in one analysis. Shading reflects similarity among numbers.

Population comparisons using filtered polymorphic loci: missing data thresholds

Previous filtering used a 20% maximum missing data threshold for each population as is commonly adopted in the literature. We investigated effects of missing data using the nested subsamples from Group A, filtered with thresholds of 0% (i.e. no missing data allowed), 10%, 20%, 30%, 40%, or 50%.

We see a considerable effect of missing data thresholds on both measures of SNP heterozygosity (Fig. 3). Samples of larger n were more strongly affected by choice of missing data threshold, with stringent filtering tending to produce higher estimates. When 50 individuals were used, filtering with a 10% missing data threshold (another common parameter setting) produced an estimate for observed heterozygosity 1.22 times higher than filtering with a 30% threshold (also a common parameter setting). This effect was less pronounced for expected heterozygosity (1.09 times higher) at n = 50, and this trend began to reverse at lower n.

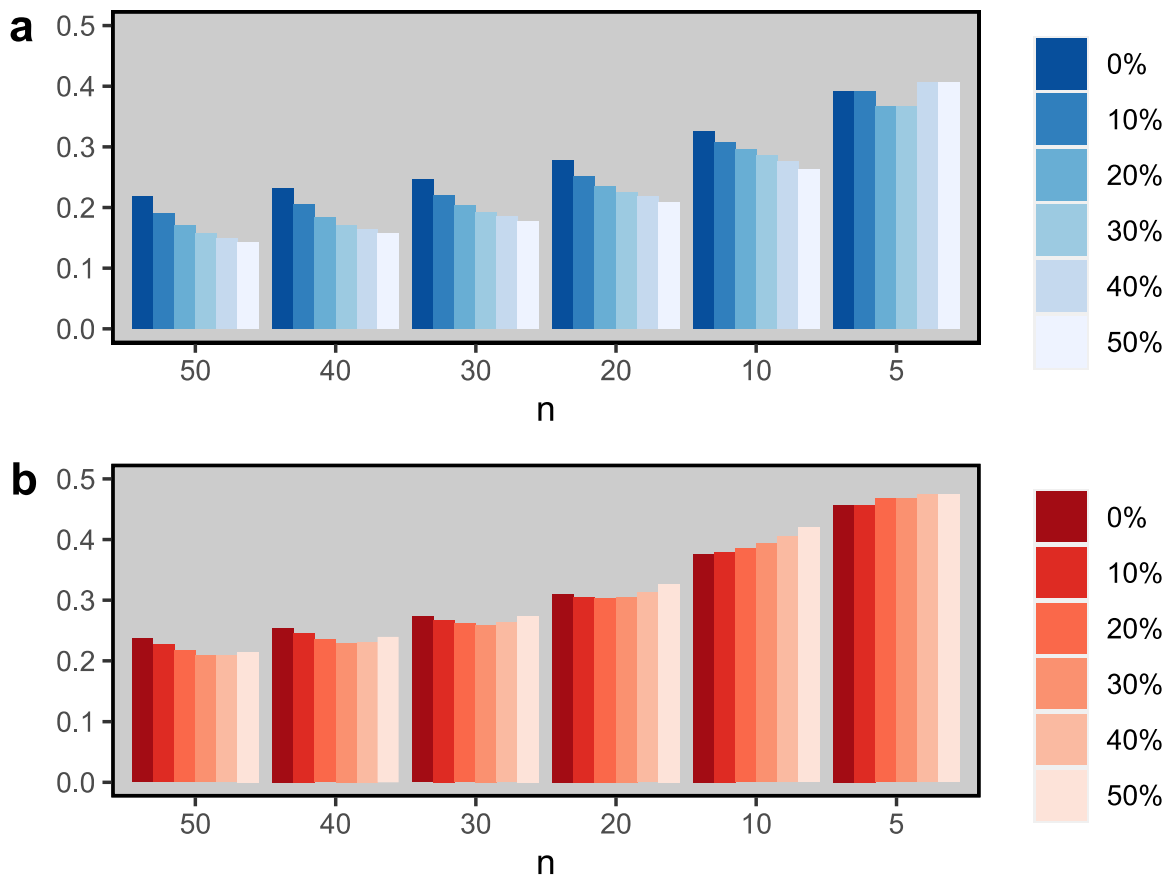

Fig. 3: Effect of missing data thresholds on observed (a) and expected (b) SNP heterozygosity estimates from the Group A mosquito samples filtered at MAC ≥ 3.

The higher observed heterozygosity estimates at 0% versus 50% thresholds might be expected if there is a correlation between errors and the degree of missing data at a locus. In that case, errors from monomorphic loci with high missing data could be read as low frequency polymorphisms, pushing down heterozygosity estimates when sample sizes are large. We expect this problem would increase if singletons and doubletons are included, as more low-frequency errors would be included in the estimates (c.f. Fig. S3).

Finally, the large population sizes in *Ae. aegypti* indicate this Malaysian sample is unlikely to contain inbred individuals, and thus we do not expect observed and expected heterozygosities to differ substantially. We note that when estimates of observed and expected heterozygosity are compared, these parameters are most similar when filtering with a 0% missing data threshold, and they start to diverge as this threshold is increased. Less stringent missing data thresholds may thus introduce artefacts of differential observed and expected heterozygosities, which may lead to incorrect inferences of local breeding patterns.

Population comparisons using all polymorphic loci

Considering these inconsistencies in SNP heterozygosity estimates when filtering datasets with 'typical' settings for genetic structure, we ran the above analyses again at a different setting using MAC ≥ 1 but this time including all SNPs rather than filtering by one SNP per RADtag. These analyses thus considered all polymorphic loci including singletons and doubletons. We used a maximum missing data threshold of 0% (or as specified), to avoid potential artefacts of differential observed and expected heterozygosities produced by including loci with missing data (Fig. 3).

Including all polymorphic loci reduced the bias from sample size; however, similar patterns of strong bias were still observed (see Figs S1-S3). Also, as predicted, missing data thresholds had a larger effect on estimates when singletons and doubletons were included (Fig. S3). This lends support to the expectation that correlations between errors and the degree of missing data at a locus drive heterozygosity estimates down when missing data thresholds are less stringent.

In light of these inconsistencies SNP heterozygosity appears prone to bias, regardless of whether filtering follows a typical protocol used for genetic structure or when considering every polymorphic locus. The bias from sample size is of particular concern, as for rare populations this cannot always be diminished by sequencing more individuals. This limits the potential for SNP heterozygosity estimates in one study to inform the results of another study.

Estimates based on all polymorphic and monomorphic loci (autosomal heterozygosity)

Given the above challenges, we next explored how sample size and missing data affect autosomal heterozygosity, which considers both monomorphic and polymorphic nucleotide information. We ran analyses on identical datasets to those used previously. For filtering, we used no MAC cut-off, and estimated heterozygosity across every site rather than every polymorphic site. In the output from the Stacks.v2 program "Populations", this corresponds to the entries in the "# All positions (variant and fixed)" subsection. We used a maximum missing data threshold of 0% (unless specified differently).

Population comparisons using all loci: effects of sample size

When considering variation at all nucleotides, observed and expected heterozygosity estimates are far less affected by sample size than SNP heterozygosity estimates (Figs 4, 5). Though there is some variability among subsamples when analysed individually, when analysed together estimates for n ≥ 10 for the multiple subsamples (Fig. 4) and nested subsamples (Fig. 5) indicated consistent observed heterozygosities of ~0.00039 and expected heterozygosities of ~0.00040. The similar estimates for these two parameters match expectations for this sample of Malaysian *Ae. aegypti*, where inbreeding is unlikely given the large size of mosquito populations and the spatial distribution of sampling.

These consistent estimates are expected when all nucleotides are taken into account because for smaller samples the higher frequency of heterozygotes at polymorphic sites will be offset by the lower number of polymorphic sites overall. Heterozygosity estimates from a set of individuals thus correlate directly with population heterozygosity because loci are not first filtered by polymorphism. In this sense, autosomal heterozygosity is a parameter that is both more robust to variation in study design and also a more accurate measure of genetic variation which can be used in comparisons across studies and organisms (e.g. Westbury et al., 2019).

Population comparisons with all loci: effects of missing data thresholds

Missing data thresholds had a smaller effect on autosomal heterozygosity than on SNP heterozygosity (Fig. 6; c.f. Figs 2, S3). Nevertheless, a similar problematic pattern is clear in the divergence between observed and expected heterozygosities when loci with missing data are included. At n = 50, filtering with a 0% missing data threshold estimated expected heterozygosity as 1.08 times larger than observed heterozygosity, while for a 20% threshold it was 1.29 times larger.

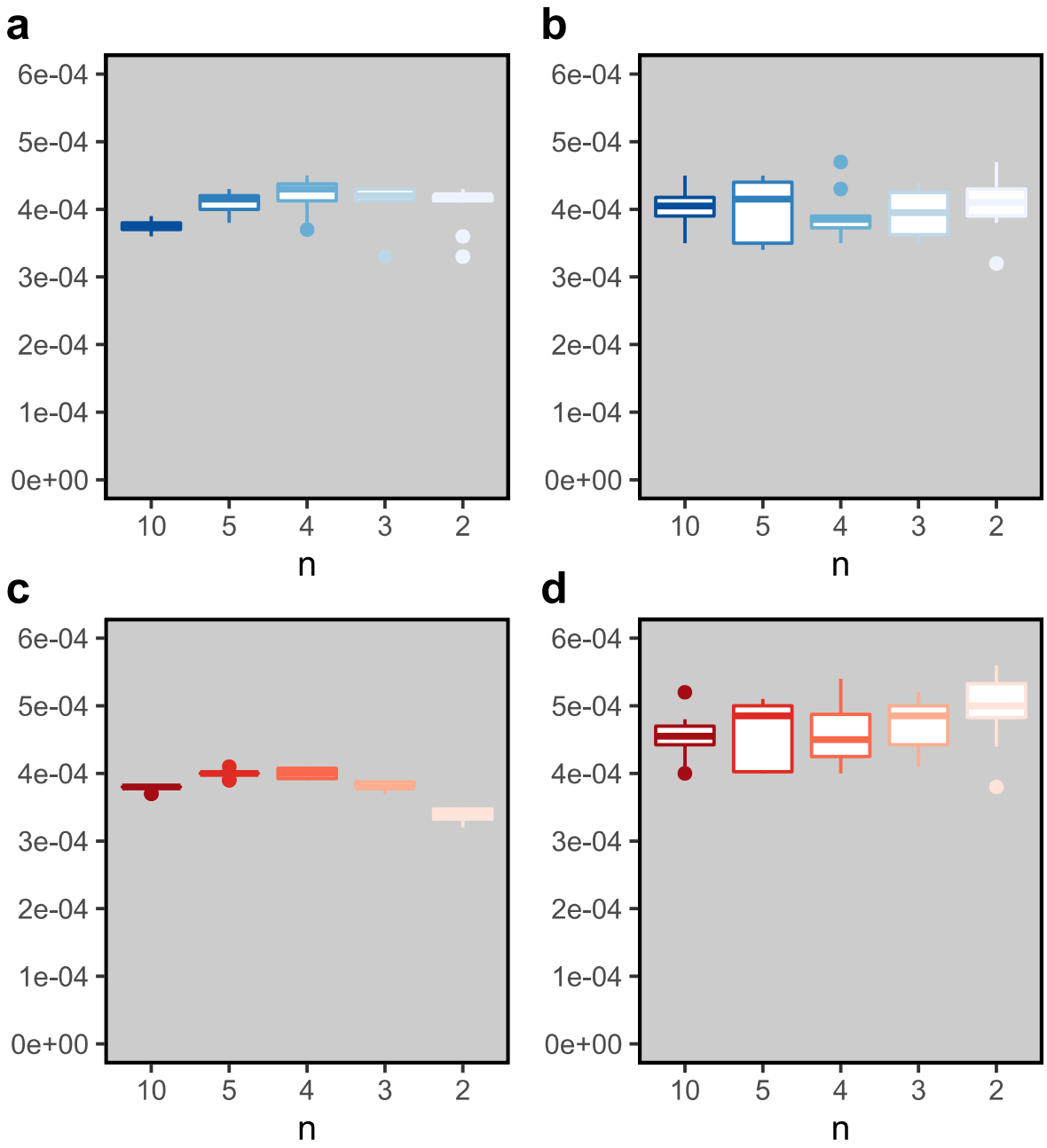

Fig. 4: Observed (blue) and expected (red) autosomal heterozygosity estimated from ten subsamples of mosquitoes at each n, with the population subsamples analysed individually (a, c) or together (b, d).

n	Observed: Group A only	Observed: Group A	Observed: Group B	Expected: Group A only	Expected: Group A	Expected: Group B
50	0.00039	0.00037	0.00038	0.00042	0.00039	0.00040
40	0.00038	0.00038	0.00039	0.00042	0.00040	0.00040
30	0.00038	0.00037	0.00039	0.00042	0.00040	0.00041
20	0.00038	0.00039	0.00041	0.00042	0.00042	0.00043
10	0.00038	0.00042	0.00043	0.00043	0.00045	0.00046
5	0.00033	0.00039	0.00041	0.00038	0.00041	0.00042

Fig. 5: Observed (blue) and expected (red) autosomal heterozygosity estimated from mosquito samples of different size (n). We compared estimates for Group A only with those of Group A and Group B combined in one analysis. Shading reflects similarity among numbers.

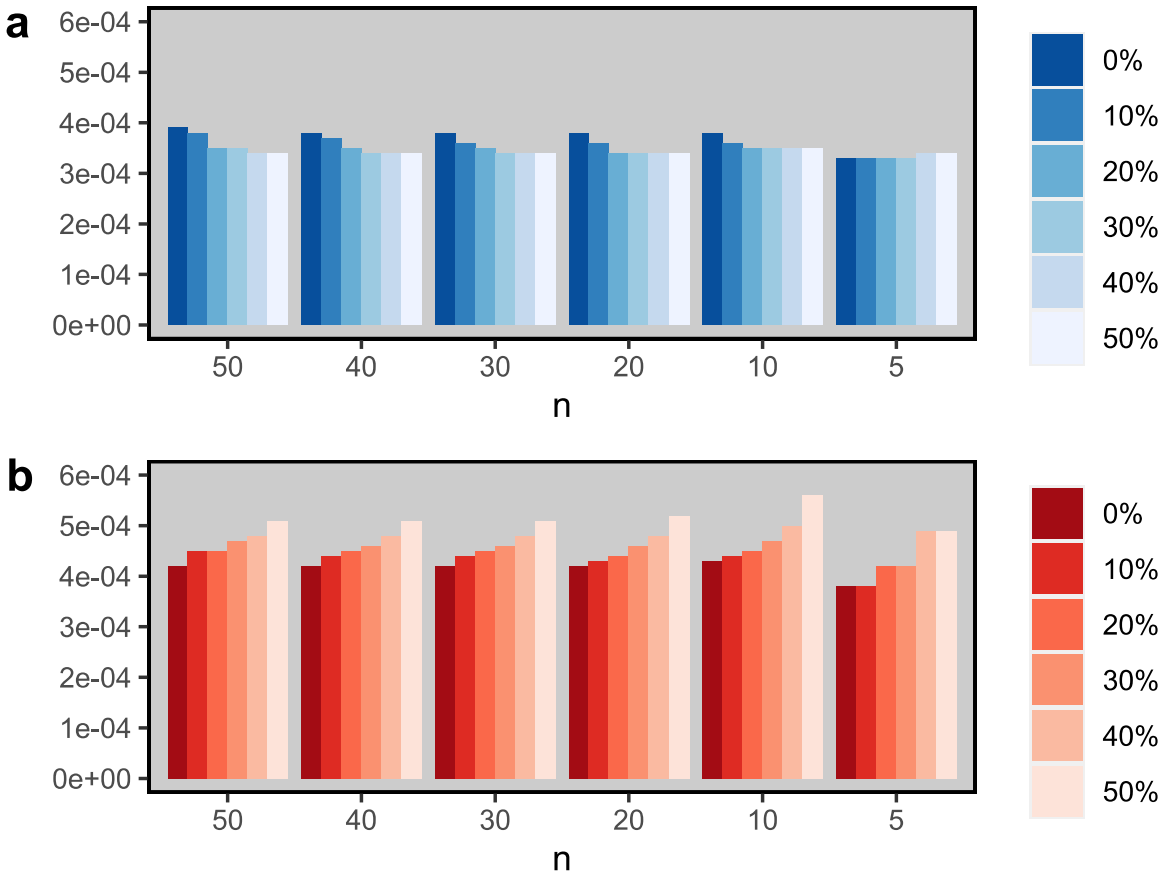

Fig. 6: Effect of missing data thresholds on observed (a) and expected (b) autosomal heterozygosity estimates from the Group A mosquito samples.

Considering these results, we propose that heterozygosity estimation should exclude loci that have any missing genotypes, as these may be more likely to contain errors or otherwise skew parameter estimates. While this filtering might at first appear overly stringent, autosomal heterozygosity is calculated using far more nucleotides than SNP heterozygosity, and should normally be based on sufficient nucleotides even after strict filtering. For example, when a 20% missing data threshold is used to estimate SNP heterozygosity (MAC ≥ 1), heterozygosity is estimated from 95,293 polymorphic loci. When a 0% missing data threshold is used to estimate autosomal heterozygosity (MAC ≥ 0), heterozygosity is estimated from 7,813,360 nucleotides, of which 17,968 are polymorphic. The specific number of nucleotides used in autosomal heterozygosity calculations may vary across studies, but appear more likely to accurately reflect the degree of genetic variation across the genome.

Multiple population considerations

We first estimated heterozygosity for the four *K. scurra* populations using equal sample sizes of 10. Fig. S4 compares results of a 0% missing data threshold against a 20% threshold, showing that at 20%, the ratio of observed to expected heterozygosity is higher ($\bar{x}$ = 1.156) than at 0% ($\bar{x}$ = 1.091), supporting our previous findings that missing data biases the ratio of heterozygosities. Accordingly, we used a 0% missing data threshold in the following analyses.

Population comparisons: local sample size variation

Although global sample size analysed above had little effect on autosomal heterozygosity when n ≥ 10, we have yet to consider differences in local sample size. We estimated heterozygosity for the four *K. scurra* populations with one population (Goulburn) set at either half (5,10,10,10) or double (10,5,5,5) the size of the other populations compared to an equal population size. We compared results for autosomal heterozygosity (Fig. 7a) and SNP heterozygosity (Fig. 7b).

When 10 individuals are analysed from each population, the Goulburn, Hall and Wallendbeen populations all have similar autosomal heterozygosities, while Cooma is much lower (Fig. 7a). There was no noticeable effect from unequal sample size. However, unequal sample size had clear effects on SNP heterozygosity (Fig. 7b). Our previous observation that variation in N had a global effect on SNP heterozygosity (Fig 1) was again evident, and here revealed that variation in n affected local estimates. This bias could lead to misinterpretation of relative SNP heterozygosities within studies when populations with different sample sizes are included in analyses.

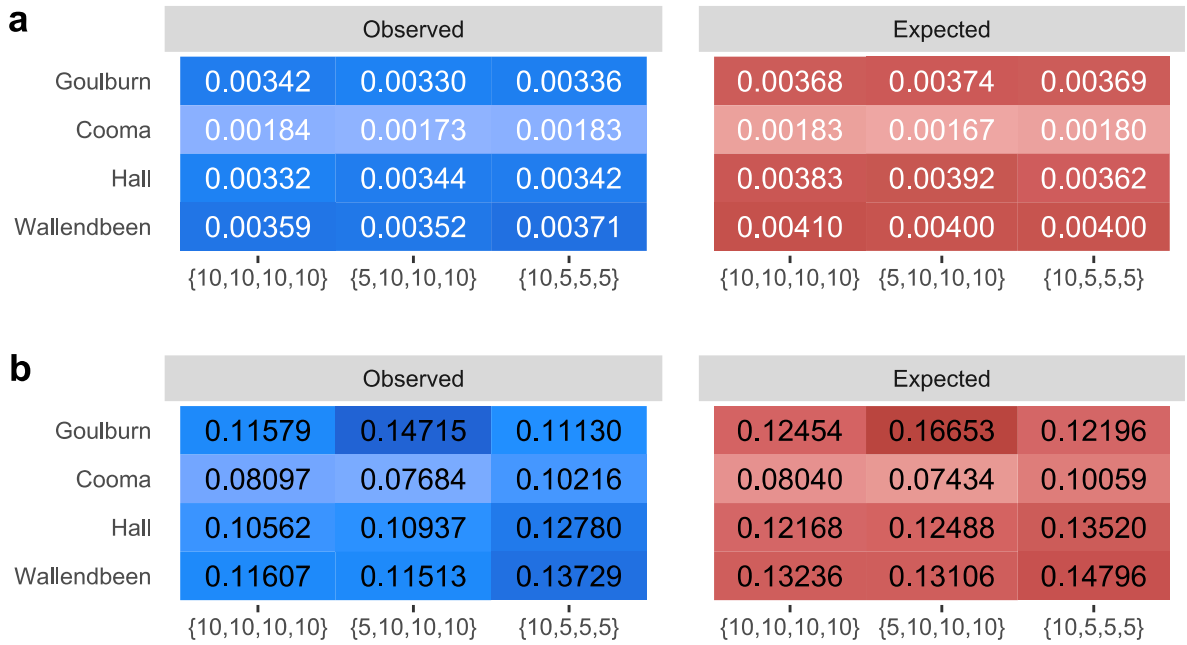

Fig. 7: Estimates of observed (blue) and expected (red) autosomal heterozygosity (a, white text) and SNP heterozygosity (b, black text) in four *K. scurra* populations of equal or unequal sample size. Numbers in brackets indicate sample sizes for the four populations sequentially (Goulburn to Wallendbeen). Shading reflects similarity among numbers.

Population comparisons: impact of population structure

Next, we investigate the effects of combining genetically differentiated populations in an analysis. In terms of mtDNA variation and DArT SNPs, Cooma was separate from the other populations and particularly Wallendbeen (Hoffmann et al., 2020). We estimated heterozygosity for each of the four *K. scurra* populations individually and compared this to populations analysed together. We used the same 0% missing data threshold as before, and compared results for autosomal heterozygosity (Fig. 8a) and SNP heterozygosity (Fig. 8b).

Autosomal heterozygosity is unaffected when analysing differentiated populations together or individually (Fig. 8a). However, strong biases on SNP heterozygosity are evident (Fig. 8b). When the four populations are analysed individually, estimates are much higher than when analysed together. Additionally, the population at Cooma, which otherwise recorded the lowest heterozygosity of the four populations, has higher heterozygosity than all other populations when analysed by itself.

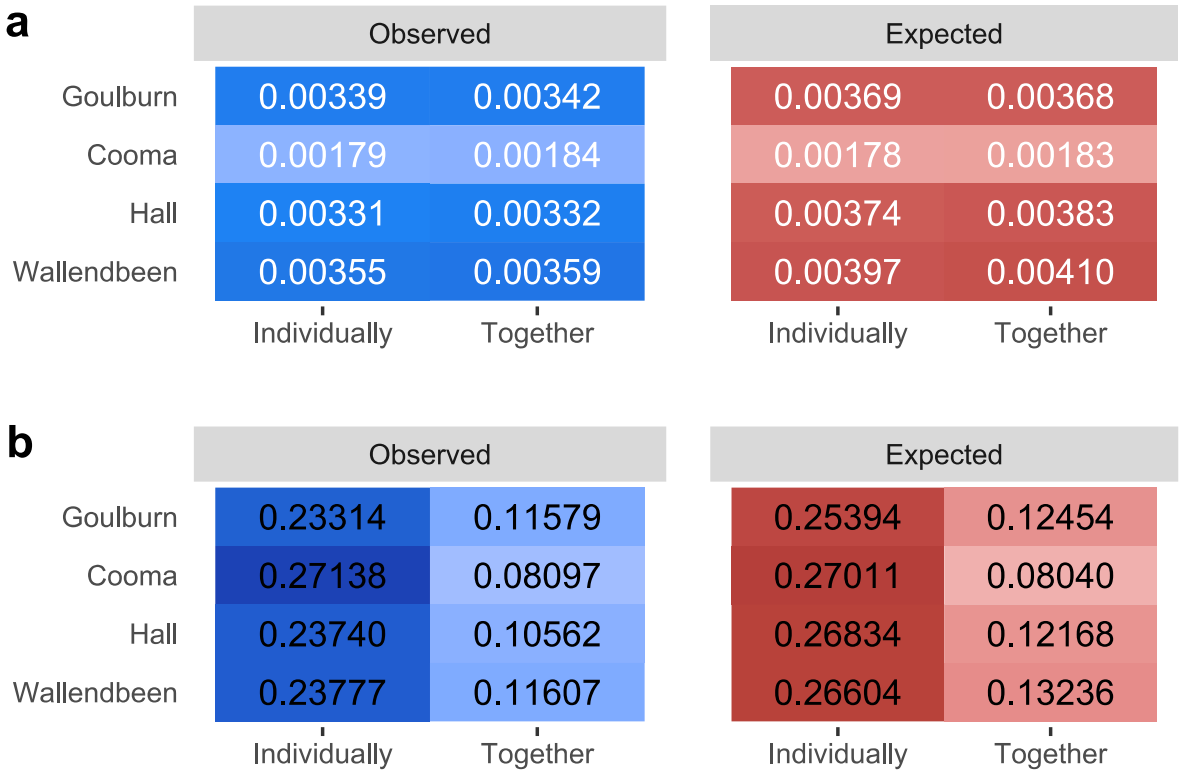

Fig. 8: Estimates of observed (blue) and expected (red) autosomal heterozygosity (a, white text) and SNP heterozygosity (b, black text) in four *K. scurra* populations when estimated individually or when the four populations were combined. Shading reflects similarity among numbers.

As allele frequencies vary among these populations, many variant sites will only be polymorphic in one or two populations and are monomorphic in the others. Thus when populations are analysed together, estimates for each population will include these variant sites that are locally monomorphic, leading to lower heterozygosity estimates than when analysed individually.

A similar explanation accounts for the sharp variation in estimates at Cooma. When analysed individually, Cooma recorded fewer polymorphic sites (3235) than the other populations (4693, 4734, 5182); this pattern was also observed when populations were analysed together. However, heterozygosity at these 3235 polymorphic sites was higher than at the other populations.

These findings highlight that SNP heterozygosity estimates reflect different parameters when they are analysed individually compared to when they are analysed alongside other

populations. When analysed individually, they provide a true estimate of SNP heterozygosity. But when analysed with other populations, they produce an estimate that will be shaped by whichever other populations are included. Accordingly, SNP heterozygosities, if they are to be included, should probably be calculated from populations analysed individually, and the total number of variant sites should also be reported for each population to provide further context to the heterozygosity estimates.

Comparing heterozygosity estimates

A final consideration concerns how to interpret heterozygosity estimates across studies. We have proposed several guidelines for filtering data to allow cross-study comparisons. The most important of these is that heterozygosity estimates should be derived from variation at both monomorphic and polymorphic sites. Table 1 compares variation in SNP heterozygosity with variation in autosomal heterozygosity for the four *K. scurra* populations analysed individually. We did not compare populations when analysed together due to the confoundment of SNP heterozygosities in these analyses (Fig 8b).

For *K. scurra*, variation in autosomal heterozygosity is approximately twice as large as variation in SNP heterozygosity (Table 1). A large difference in SNP heterozygosity might not be detected even when comparing populations with very low and very high levels of genetic variation because the exclusion of monomorphic sites in each population will reduce differences in genetic variability among the populations.

Table 1. Comparison of observed and expected SNP heterozygosity and autosomal heterozygosity in four *K. scurra* populations when estimated individually. MAX/MIN is the ratio between the largest score and the smallest score, Relative s.d. (%) is the standard deviation divided by the mean as a percentage.

	SNP heterozygosity		**Autosomal heterozygosity**	
	Observed	Expected	Observed	Expected
Goulburn	0.23314	0.25394	0.00339	0.00369
Cooma	0.27138	0.27011	0.00179	0.00178
Hall	0.23740	0.26834	0.00331	0.00374
Wallendbeen	0.23777	0.26604	0.00355	0.00397
MAX/MIN	1.16	1.06	1.98	2.23
Relative s.d. (%)	6.28	2.39	23.58	26.74

Discussion

Comparisons of heterozygosity across populations and species are frequently used to inform management decisions in conservation programs. Management actions will normally require heterozygosity estimates that can be used to make comparisons across populations and related species, and they may involve comparisons across different studies. An example of a relevant management action involves selecting populations for genetic and evolutionary rescue which aims to decrease levels of inbreeding but increase levels of genetic variation in target populations based on targeted introductions of individuals from other populations (Whiteley, Fitzpatrick, Funk, & Tallmon, 2015; Hoffmann et al., 2020). The usefulness of source populations for conservation translocations is to a large extent determined by their genetic variability (Reid et al., 2016; Ørsted et al., 2019). Moreover, changes in heterozygosity across time can be a worthwhile means of tracking the genetic health of threatened populations (Mitrovski, Hoffmann, Heinze, & Weeks, 2008) and is particularly useful for determining outcomes of management interventions (Weeks et al., 2017).

By applying different filtering settings to a pair of reduced-representation sequencing datasets, we have shown that filtering using optimal settings for detecting genetic structure will produce heterozygosity estimates that can be difficult to compare across studies, and specifically we show that heterozygosity estimates that consider only polymorphic sites are biased by local sample size. Estimates of SNP heterozygosity are therefore poorly-suited to comparisons across studies. Autosomal heterozygosity estimates do not suffer from these biases, and appear to be a more useful parameter to report. We also found that when sites with missing data are included, observed and expected heterozygosity estimates diverge, with the divergence proportional to the amount of missing data permitted. Following this, we propose two general guidelines that should help meaningful comparisons: (i) studies aiming to summarise population genetic variation should report autosomal heterozygosity, either by itself or alongside SNP heterozygosity; (ii) sites with missing data should be omitted from heterozygosity calculations. Although we have not explicitly investigated the importance of sequencing coverage, this is widely known to be critical for accurately identifying heterozygotes (Nielsen et al., 2011). This being the case, our findings that heterozygosity estimates can be consistent even at low n (Fig. 7a) point to the optimal design for heterozygosity being deep sequencing of a small number of individuals (perhaps 5-10) from each population, rather than shallower sequencing of many individuals.

SNP heterozygosity is frequently the only measure of heterozygosity reported (Jones et al., 2012; Chen et al., 2016; Bock et al., 2018; Mathur et al., 2019; Surbakti et al., 2020). Although there may be cases where SNP heterozygosity is the appropriate parameter, it will be subject to biases that do not affect autosomal heterozygosity, making the latter a better 'default' choice for reporting variation in populations. A key advantage is the consistency of heterozygosity estimates across different sample sizes. Polymorphic sites considered alone

produced biased estimates that failed to reach an asymptote even at n ≥ 40. As taxa of conservation interest are frequently rare or difficult to sample in large numbers, large sample sizes would in any case be rare. SNP heterozygosity also does not capture the actual level of overall variation in the genome. For instance, *K. scurra* from Cooma had a higher SNP heterozygosity estimate (when calculated at the population level independently) yet fewer polymorphic sites than the other populations; this low level of polymorphism was evident in the autosomal heterozygosity estimate.

Whole-genome sequencing studies frequently report heterozygosity across all sites and with stringent missing data thresholds as proposed here (e.g. Gopalakrishnan et al., 2017; Westbury, Petersen, Garde, Heide-Jørgensen, & Lorenzen, 2019). However, this methodology has been less commonly applied in reduced-representation sequencing studies such as those using RADseq or DArTseq markers. One possible reason may be that reduced-representation protocols are commonly applied to wild populations of understudied organisms, where the first step is typically to analyse genetic structure among populations. Monomorphic sites are uninformative for most analyses of genetic structure and thus these will typically be removed during filtering, and this remaining set of SNPs will be used to estimate both genetic structure and heterozygosity. Here we propose that filtering for genetic structure (variation between populations) and heterozygosity (variation within populations) be treated as distinct targets for processing with different parameter settings.

When expected heterozygosity is higher than observed heterozygosity, it is often treated as evidence for local inbreeding (e.g. Hoffmann et al., 2020), an important parameter that can indicate a need for genetic intervention in threatened species (Ralls et al., 2018). However, we observed frequent changes in the ratio of these two parameters dependent on factors including missing data and sample size. Figs 3, 6, and S3 indicate that inferences of inbreeding in the presence of missing data will be confounded, as the heterozygosity parameters quickly diverge as sites with missing data are included. These opposing effects of missing data filters on observed and expected heterozygosity are consistent with sequencing error rates being higher at nucleotide positions where there are some missing genotypes.

A possible explanation is that sequencing errors are more likely at monomorphic sites; these sites are more common, and more likely to be retained than errors at polymorphic sites which may introduce a third allele and be removed from the dataset (most filtering protocols retain only biallelic sites). Monomorphic sites with errors may therefore be coded as low-frequency SNPs. Including them could affect the observed and expected autosomal heterozygosity estimates differently depending on whether errors are mostly coded as homozygous or heterozygous. Sequencing errors affecting monomorphic sites could also account for observations on SNP heterozygosity (Figs 3, S3), where higher missing data thresholds result in a decrease in observed and expected heterozygosity (due to monomorphic loci being erroneous included as low-frequency SNPs).

Despite these issues, the *K. scurra* data also indicate clear instances where inbreeding levels differ across populations. We note that for *K. scurra* from Cooma, observed and expected heterozygosities were almost identical, while expected heterozygosities are 8-15% higher in the other populations, suggesting a situation where inbreeding in Cooma is low but genetic variability is also low. An accurate assessment of inbreeding in populations versus low genetic variation is important when making recommendations around genetic mixing of threatened populations, which can target both the masking of deleterious genes expressed as a consequence of inbreeding as well as problems from low levels of genetic variability (Weeks et al., 2011; Ralls et al., 2018; Hoffmann et al., 2020). It would be worth further evaluating optimal filtering strategies for assessing the ratio of observed and expected heterozygosities, using a set of populations where inbreeding is known to occur in some populations but not others.

Here we highlight issues with estimating population heterozygosities from SNP datasets and indicate that autosomal heterozygosity estimates are more robust to the influence of sample size and missing data, and are likely to be more comparable across studies. We provide general guidelines for estimating population heterozygosity from reduced representation sequencing datasets that are likely to be different from guidelines for estimating other population genetic parameters such as gene flow, population structure, relatedness and effective population size. As in previous assessments (Linck & Battey, 2019), our results indicate that SNP datasets need to be carefully evaluated when they are used to obtain genetic parameters for populations that inform management decisions.

Acknowledgements

This research was supported by the Australian Research Council (Discovery Grant DP190100990) and the National Health and Medical Research Council (Program Grant no. 1132412; Fellowship Grant no. 1118640). This research was facilitated by use of the Nectar Research Cloud.

Data Archiving

Aligned .bam files for 100 *Ae. aegypti* and 40 *K. scurra* will be made available through the NCBI SRA by time of publication.

Author Contributions

AAH and ARW conceived the study; TLS, MJ, AAH and ARW designed the methodology; TLS and MJ analysed the data; TLS, AAH and ARW led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

References

Bock, D. G., Kantar, M. B., Caseys, C., Matthey-Doret, R., & Rieseberg, L. H. (2018). Evolution of invasiveness by genetic accommodation. *Nature Ecology and Evolution*, *2*(6), 991–999. doi:10.1038/s41559-018-0553-z

Campbell, M. R., Vu, N. V., LaGrange, A. P., Hardy, R. S., Ross, T. J., & Narum, S. R. (2019). Development and Application of Single-Nucleotide Polymorphism (<scp>SNP</scp>) Genetic Markers for Conservation Monitoring of Burbot Populations. *Transactions of the American Fisheries Society*, *148*(3), 661–670. doi:10.1002/tafs.10157

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–40. doi:10.1111/mec.12354

Chen, N., Cosgrove, E. J., Bowman, R., Fitzpatrick, J. W., & Clark, A. G. (2016). Genomic Consequences of Population Decline in the Endangered Florida Scrub-Jay. *Current Biology*, *26*(21), 2974–2979. doi:10.1016/j.cub.2016.08.062

Coltman, D. W., Pilkington, J. G., Smith, J. A., & Pemberton, J. M. (1999). Parasite-mediated selection against inbred Soay sheep in a free-living, island population. *Evolution*, *53*(4), 1259–1267. doi:10.1111/j.1558-5646.1999.tb04538.x

De Meeûs, T. (2018). Revisiting FIS, FST, Wahlund Effects, and Null Alleles. *Journal of Heredity*, *109*(4), 446–456. doi:10.1093/jhered/esx106

Fitzpatrick, S. W., Bradburd, G. S., Kremer, C. T., Salerno, P. E., Angeloni, L. M., & Funk, W. C. (2020). Genomic and Fitness Consequences of Genetic Rescue in Wild Populations. *Current Biology*, *30*(3), 517-522.e5. doi:10.1016/j.cub.2019.11.062

Fontaine, A., Filipović, I., Fansiri, T., Hoffmann, A. A., Cheng, C., Kirkpatrick, M., … Lambrechts, L. (2017). Extensive genetic differentiation between homomorphic sex chromosomes in the mosquito vector, *Aedes aegypti*. *Genome Biology and Evolution*, *9*(9), 2322–2335. doi:10.1093/gbe/evx171

Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., … Luikart, G. (2016, February 1). Genomics in Conservation: Case Studies and Bridging the Gap between Data and Application. *Trends in Ecology and Evolution*. Elsevier Ltd. doi:10.1016/j.tree.2015.10.009

Gopalakrishnan, S., Samaniego Castruita, J. A., Sinding, M. H. S., Kuderna, L. F. K., Räikkönen, J., Petersen, B., … Gilbert, M. T. P. (2017). The wolf reference genome sequence (Canis lupus lupus) and its implications for Canis spp. population genomics. *BMC Genomics*, *18*(1), 495. doi:10.1186/s12864-017-3883-3

Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2017). DARTR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*, *18*(3), 1–9. doi:10.1111/1755-0998.12745

Hoffmann, A. A., Sgrò, C. M., & Kristensen, T. N. (2017, July 1). Revisiting Adaptive Potential, Population Size, and Conservation. *Trends in Ecology and Evolution*. Elsevier Ltd. doi:10.1016/j.tree.2017.03.012

Hoffmann, A. A., White, V., Jasper, M., Yagui, H., Sinclair, S., & Kearney, M. (2020). An endangered flightless grasshopper with strong genetic structure maintains population genetic variation despite extensive habitat loss. *Authorea Preprints*. doi:10.22541/AU.160403053.38828134/V1

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genetics*,

6(2), e1000862. doi:10.1371/journal.pgen.1000862

Jasper, M., Schmidt, T. L., Ahmad, N. W., Sinkins, S. P., & Hoffmann, A. A. (2019). A genomic approach to inferring kinship reveals limited intergenerational dispersal in the yellow fever mosquito. *Molecular Ecology Resources*, *19*(5), 1254–1264. doi:10.1111/1755-0998.13043

Jones, F. C., Chan, Y. F., Schmutz, J., Grimwood, J., Brady, S. D., Southwick, A. M., … Kingsley, D. M. (2012). A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current Biology*, *22*(1), 83–90. doi:10.1016/j.cub.2011.11.045

Lemopoulos, A., Prokkola, J. M., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P., … Vainikka, A. (2019). Comparing RADseq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. *Ecology and Evolution*, *9*(4), 2106–2120. doi:10.1002/ece3.4905

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, *19*(3), 639–647. doi:10.1111/1755-0998.12995

Maroso, F., Franch, R., Dalla Rovere, G., Arculeo, M., & Bargelloni, L. (2016). RAD SNP markers as a tool for conservation of dolphinfish *Coryphaena hippurus* in the Mediterranean Sea: Identification of subtle genetic structure and assessment of populations sex-ratios. *Marine Genomics*, *28*, 57–62. doi:10.1016/j.margen.2016.07.003

Mathur, S., Tomeček, J. M., Heniff, A., Luna, R., & DeWoody, J. A. (2019). Evidence of genetic erosion in a peripheral population of a North American game bird: the Montezuma quail (*Cyrtonyx montezumae*). *Conservation Genetics*, *20*(6), 1369–1381. doi:10.1007/s10592-019-01218-9

Mitrovski, P., Hoffmann, A. A., Heinze, D. A., & Weeks, A. R. (2008). Rapid loss of genetic variation in an endangered possum. *Biology Letters*, *4*(1), 134–138. doi:10.1098/rsbl.2007.0454

Mulvena, S. R., Pierson, J. C., Farquharson, K. A., McLennan, E. A., Hogg, C. J., & Grueber, C. E. (2020). Investigating inbreeding in a free-ranging, captive population of an Australian marsupial. *Conservation Genetics*, *21*(4), 665–675. doi:10.1007/s10592-020-01278-2

Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Molecular Ecology Resources*, *17*(6), 1136–1147. doi:10.1111/1755-0998.12654

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443–451. doi:10.1038/nrg2986

Ørsted, M., Hoffmann, A. A., Sverrisdóttir, E., Nielsen, K. L., & Kristensen, T. N. (2019). Genomic variation predicts adaptive evolutionary responses better than population bottleneck history. *PLoS Genetics*, *15*(6). doi:10.1371/journal.pgen.1008205

Ralls, K., Ballou, J. D., Dudash, M. R., Eldridge, M. D. B., Fenster, C. B., Lacy, R. C., … Frankham, R. (2018). Call for a Paradigm Shift in the Genetic Management of Fragmented Populations. *Conservation Letters*, *11*(2), e12412. doi:10.1111/conl.12412

Reid, N. M., Proestou, D. A., Clark, B. W., Warren, W. C., Colbourne, J. K., Shaw, J. R., … Whitehead, A. (2016). The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science*, *354*(6317), 1305–1308. doi:10.1126/science.aah4993

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical*

Research, *67*(2), 175–185. doi:10.1017/s0016672300033620

Surbakti, S., Parker, H. G., McIntyre, J. K., Maury, H. K., Cairns, K. M., Selvig, M., … Ostrander, E. A. (2020). New Guinea highland wild dogs are the original New Guinea singing dogs. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(39), 24369–24376. doi:10.1073/pnas.2007242117

Waples, R. S., & Allendorf, F. (2015). Testing for hardy-weinberg proportions: Have we lost the plot? *Journal of Heredity*, *106*(1), 1–19. doi:10.1093/jhered/esu062

Weeks, A. R., Heinze, D., Perrin, L., Stoklosa, J., Hoffmann, A. A., Van Rooyen, A., … Mansergh, I. (2017). Genetic rescue increases fitness and aids rapid recovery of an endangered marsupial population. *Nature Communications*, *8*(1). doi:10.1038/s41467-017-01182-3

Weeks, A. R., Sgro, C. M., Young, A. G., Frankham, R., Mitchell, N. J., Miller, K. A., … Hoffmann, A. A. (2011). Assessing the benefits and risks of translocations in changing environments: A genetic perspective. *Evolutionary Applications*, *4*(6), 709–725. doi:10.1111/j.1752-4571.2011.00192.x

Westbury, M. V., Petersen, B., Garde, E., Heide-Jørgensen, M. P., & Lorenzen, E. D. (2019). Narwhal Genome Reveals Long-Term Low Genetic Diversity despite Current Large Abundance Size. *IScience*, *15*, 592–599. doi:10.1016/j.isci.2019.03.023

Whiteley, A. R., Fitzpatrick, S. W., Funk, W. C., & Tallmon, D. A. (2015, January 1). Genetic rescue to the rescue. *Trends in Ecology and Evolution*. Elsevier Ltd. doi:10.1016/j.tree.2014.10.009

Wright, B. R., Grueber, C. E., Lott, M. J., Belov, K., Johnson, R. N., & Hogg, C. J. (2019). Impact of reduced-representation sequencing protocols on detecting population structure in a threatened marsupial. *Molecular Biology Reports*, *46*(5), 5575–5580. doi:10.1007/s11033-019-04966-6