



Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Schmidt, TL;Jasper, ME;Weeks, AR;Hoffmann, AA

Title:

Unbiased population heterozygosity estimates from genome-wide sequence data

Date:

2021-10-01

Citation:

Schmidt, T. L., Jasper, M. E., Weeks, A. R. & Hoffmann, A. A. (2021). Unbiased population heterozygosity estimates from genome-wide sequence data. *Methods in Ecology and Evolution*, 12 (10), pp.1888-1898. <https://doi.org/10.1111/2041-210X.13659>.

Persistent Link:

<https://hdl.handle.net/11343/337971>

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

MR TOM SCHMIDT (Orcid ID : 0000-0003-4695-075X)

DR ARY A. HOFFMANN (Orcid ID : 0000-0001-9497-7645)

Article type : Research Article

Unbiased population heterozygosity estimates from genome-wide sequence data

Thomas L Schmidt^{1*}, Moshe Jasper¹, Andrew R Weeks^{1,2}, Ary A Hoffmann¹

¹School of BioSciences, Bio21 Institute, University of Melbourne, Parkville, Victoria, Australia

²cesar Pty Ltd, Parkville, Victoria, Australia

*Correspondence: Tom Schmidt, Bio21 Institute, University of Melbourne, Parkville,
Victoria, Australia, toms@unimelb.edu.au

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/2041-210X.13659](https://doi.org/10.1111/2041-210X.13659)

This article is protected by copyright. All rights reserved

22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

Running title: Unbiased estimation of heterozygosity

Abstract

1. Heterozygosity is a metric of genetic variability frequently used to inform the management of threatened taxa. Estimating observed and expected heterozygosities from genome-wide sequence data has become increasingly common, and these estimates are often derived directly from genotypes at single nucleotide polymorphism (SNP) markers. While many SNP markers can provide precise estimates of genetic processes, the results of ‘downstream’ analysis with 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 observed and expected heterozygosity using two reduced-representation sequencing datasets, one from the mosquito *Aedes aegypti* (ddRADseq) and the other from a threatened grasshopper, *Keyacris scurra* (DArTseq).
3. We show that estimates based on polymorphic markers only (i.e. SNP heterozygosity) are always biased by global sample size (N), with smaller N producing larger estimates. By contrast, results are unbiased by sample size when calculations consider monomorphic as well as polymorphic sequence information (i.e. genome-wide or autosomal heterozygosity). SNP heterozygosity is also biased when differentiated populations are analysed together, while autosomal heterozygosity remains unbiased. We also show that when nucleotide sites with missing genotypes are included, observed and expected heterozygosity estimates diverge in proportion to the amount of missing data permitted at each site.
4. We make three recommendations for estimating genome-wide heterozygosity: (i) autosomal heterozygosity should be reported instead of (or in addition to) SNP heterozygosity; (ii) sites with any missing data should be omitted; (iii) populations

51 should be analysed in independent runs. This should facilitate comparisons within
52 and across studies and between observed and expected measures of heterozygosity.

53

54

55

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

58

59 **Introduction**

60 The power provided by single nucleotide polymorphism (SNP) markers detected using
61 genome-wide sequencing approaches is leading to their increased use in conservation
62 genetic studies (Garner et al., 2016). SNPs are popular for investigating levels of genetic
63 differentiation among remnant populations and for comparing levels and patterns of
64 genetic variation within populations (Campbell et al., 2019; Maroso et al., 2016) which
65 provides information on the adaptive potential of populations (Ørsted et al., 2019) as well
66 as patterns of inbreeding and relatedness (Mulvena et al., 2020). Results from SNP studies
67 are being interpreted for use in management decisions that include genetic rescue, genetic
68 mixing and founder selection in threatened species programs (Fitzpatrick et al., 2020). The
69 relative ease of generating SNP genotypes is leading to their increased use by non-
70 specialists, particularly through the availability of companies such as Diversity Arrays
71 Technology (<https://www.diversityarrays.com>), which provide SNP genotypes through
72 customised in-house processes (Gruber et al., 2017; Mulvena et al., 2020; Wright et al.,
73 2019).

74 Considering the popularity of SNP markers, it is important to be aware of any biases
75 inherent in their application to conservation genetics and elsewhere. While potential biases
76 have been considered for the detection of structure between populations (Linck & Battey,
77 2019; Wright et al., 2019), there has been less focus on the estimation of genetic variability
78 within populations. These estimates are important because they link to the evolutionary

79 potential of populations, which is typically higher in populations with greater genetic
80 variability (Hoffmann et al., 2017; Ørsted et al., 2019). Genetic variability of populations is
81 therefore crucial when making genetic management decisions for threatened species
82 (Hoffmann et al., 2020; Weeks et al., 2011).

83 Genetic variation in populations is measured in several ways, the most common of which
84 are heterozygosity (observed and expected) and the proportion of nucleotide sites that are
85 polymorphic. Heterozygosity is usually estimated from a substantial number of individuals
86 sampled from each population, but with large quantities of sequence data fewer individuals
87 may be needed (Nazareno et al., 2017). Accurate heterozygosity estimates also require that
88 the apparent diversity at a site is not related to errors introduced during sequencing or
89 genotyping, the latter of which requires adequate coverage to ensure that both strands of a
90 diploid individual are sequenced (Nielsen et al., 2011). While expected heterozygosity is
91 estimated from allele frequencies, observed heterozygosity is estimated from individual
92 genotypes directly and depends on both the amount of genetic variation in the population
93 and the level of inbreeding, which increases homozygosity (Ritland, 1996). Inbreeding can
94 thus be estimated by comparing observed heterozygosity to expected heterozygosity, with
95 the latter expected to be relatively higher when there is inbreeding ($F_{IS} > 0$).

96 For heterozygosity, h_i , the observed heterozygosity for an individual at site i can be averaged
97 across n sites as $\frac{\sum_{i=1}^n h_{i..n}}{n}$ and averaged across a sample of individuals to provide a population
98 estimate. This can be calculated from variation at polymorphic sites only (i.e. SNP
99 heterozygosity) or at both polymorphic and monomorphic sites (i.e. genome-wide or
100 autosomal heterozygosity). Both SNP heterozygosity and autosomal heterozygosity appear
101 in the literature; most population-focussed studies tend to report SNP heterozygosity (Bock
102 et al., 2018; Chen et al., 2016; Jones et al., 2012; Mathur et al., 2019; Surbakti et al., 2020)
103 although others use autosomal heterozygosity (Hohenlohe et al., 2010) which is the
104 parameter reported in studies comparing individual genomes (Gopalakrishnan et al., 2017;
105 Westbury et al., 2019). As SNP heterozygosities will be orders of magnitude larger than
106 autosomal heterozygosities, the two parameters cannot be directly compared, though for
107 studies of a single population autosomal heterozygosity can be converted to SNP
108 heterozygosity by dividing the estimate by the proportion of polymorphic sites. Fig. S1

109 provides a visualisation of how observed heterozygosity is calculated using all sequence
110 information (autosomal heterozygosity) and using polymorphic markers only (SNP
111 heterozygosity).

112 This paper investigates how SNP heterozygosity and autosomal heterozygosity perform
113 under variable conditions of sampling and filtering. These include local and global sample
114 size, rare allele filtering, missing data thresholds and the analysis of multiple differentiated
115 populations, all of which are common sources of variability within or between studies. We
116 explore these questions with a pair of genome-wide datasets of the sort frequently used for
117 assessing variation in wild populations. We focus initially on a ddRADseq dataset from one
118 population of a common species and then consider a DArTseq dataset from a threatened
119 species that covers multiple populations. We make some recommendations for assessing
120 heterozygosity when study aims include comparisons of genetic variability across
121 populations and with other studies.

122

123

124 **Materials and Methods**

125 **Sequence data from the same population**

126 We start by considering a single, well-mixed population. We use double digest restriction-
127 site associated (ddRAD) sequence data obtained from 100 female *Aedes aegypti* mosquitoes
128 sampled from a 0.125 km² area of Kuala Lumpur, Malaysia (Jasper et al., 2019). Note that as
129 this ddRADseq dataset contains only females, and as *Ae. aegypti* mosquitoes do not have
130 definable sex chromosomes but rather a small sex-determining region (Fontaine et al.,
131 2017), we did not need to filter out genotypes at sex chromosomes.

132 We took subsamples from this population as follows.

133 *Ten subsamples:* We tested the effect of five population sample sizes ($n = 10, 5, 4, 3, 2$) on
134 heterozygosity estimates by subsampling the 100 individuals, without replacement. We
135 repeated the subsampling 10 times for each sample size n .

136 *Nested subsamples:* We tested the effect of six larger sample sizes ($n = 50, 40, 30, 20, 10, 5$)
137 on heterozygosity estimates by subsampling the 100 individuals twice, without replacement.
138 This can help indicate whether filtering choices produce similar patterns at large n as at
139 small n . These subsamples were also used to test whether different filtering choices could
140 produce variable heterozygosity estimates from the same sample of individuals. To reduce
141 variation among subsamples of different size, we used a nested subsampling approach. The
142 100 individuals were randomly assigned to two groups, A and B, each of $n = 50$. Group A and
143 Group B were then subsampled once at each n , but where each subsample could only
144 include individuals that were included at the next highest n . For instance, the subsample at
145 $n = 30$ could only contain individuals that were present in the $n = 40$ subsample to allow for
146 a direct comparison between sample sizes.

147

148

149 **Sequence data from multiple populations**

150 We considered the issue of multiple populations being included in a comparison by
151 reanalysing a set of four populations of *Keyacris scurra* (Key's Matchstick Grasshopper)
152 taken from a larger set of sequencing data derived from a Diversity Arrays Technology
153 (DArT) approach. *Keyacris scurra* has recently been listed as endangered and is currently
154 restricted in range to refugia in south-eastern Australia. These four populations have
155 experienced very low gene flow and are highly differentiated (pairwise $F_{ST} = 0.14-0.28$). The
156 four populations were processed and sequenced together as part of the same project
157 (Hoffmann et al., 2020). Note that no reference assembly is available for this species so the
158 term "autosomal heterozygosity" here will also include sequence data from any
159 differentiated sex chromosomes (or regions of chromosomes).

160

161 **Sequence processing**

162 For the ddRADseq dataset, aligned sequences were built into a Stacks.v2 (Catchen et al.,
163 2013) catalog with the program `ref_map`. For the DArTSeq dataset, sequence data were
164 built into a *de novo* Stacks catalog using the program `denovo_map`, allowing for up to four

165 mismatches within and between individuals. We analysed both datasets with the Stacks
166 program “Populations”, which was used to estimate observed and expected heterozygosity
167 for a range of filtering settings described below.

168

169 **Results**

170 **Estimates based on polymorphic sites (SNP heterozygosity)**

171 Our first aim was to see how a dataset filtered with settings typically used for assessing
172 genetic structure (i.e. variation between populations) might perform when used to estimate
173 heterozygosity (i.e. variation within populations). Analysis of genetic structure will usually
174 consider only polymorphic sites (SNPs). When filtering SNPs, a common approach is to
175 combine the entire data set, remove sites not genotyped in a sufficient number of
176 individuals (typically 70-95%), and then filter out sites with a minor allele frequency (MAF)
177 or minor allele count (MAC) that is not met globally (Lemopoulos et al., 2019; Mathur et al.,
178 2019; Mulvena et al., 2020). Simulations suggest that a $MAC \geq 3$ may be optimal for
179 detecting population structure, as excluding rare alleles can lead to erroneous inferences of
180 admixture but including singletons and doubletons can confound model-based inferences of
181 structure (Linck & Battey, 2019).

182

183 *Population comparisons using polymorphic sites only: effects of sample size*

184 We start with a simple comparison of how global (N) and local (n) sample size affects SNP
185 heterozygosity estimates. For this we use the ten subsamples (n = 10, 5, 4, 3, 2), which we
186 analyse first individually (i.e. with each subsample run in a separate Stacks run) and then
187 together (i.e. where the ten subsamples are run in a single Stacks run). To investigate these
188 effects at $n \geq 5$, we use the nested subsamples from Groups A and B, first analysing each
189 subsample from Group A in individual runs, then analysing each pair of subsamples of equal
190 n from A and B together. Filtering followed a standard approach for assessing genetic
191 structure, retaining a single SNP from each RAD locus (--write-single-snp) which had no

192 more than 20% missing data and that had a $MAC \geq 3$. Observed and expected
193 heterozygosities were estimated from these filtered polymorphic sites.

194

Fig. 1: Boxplots showing effects of local and global sample size on heterozygosity estimates. Observed (blue; a-d,i-l,q-t) and expected (red; e-h,m-p,u-x) heterozygosities have been derived from three filtering treatments: polymorphic sites only, $\leq 20\%$ missing data, $MAC \geq 3$ (a-h); polymorphic sites only, 0% missing data, $MAC \geq 1$ (i-p); polymorphic and monomorphic sites, 0% missing data (q-x). Treatments have been applied to four *Ae. aegypti* datasets described in the main text: ten subsamples each of size n , analysed in individual runs (a,e,l,m,q,u); ten subsamples each of size n , analysed together in a single run (b,f,j,n,r,v); single nested subsamples from Group A, each undergoing jack-knife resampling (c,g,k,o,s,w); nested subsamples from Group A and Group B analysed together, each undergoing jack-knife resampling (d,h,l,p,t,x). All Y-axes use a log-10 scale.

195

196

197 SNP heterozygosity estimates are shown in Fig. 1a-h and indicate how this type of filtering
198 approach presents problems for comparing estimates across studies. In all cases, observed
199 and expected heterozygosities were larger when fewer samples were used for estimation.
200 Specifically, heterozygosities are biased by global N (total sample size in the analysis) rather
201 than local n (sample size of each population), as evident from comparisons of subsamples of
202 equal n analysed either in individual runs or together. Although these effects reduce as N
203 increases, they persist even with $n = 40$ and $N \geq 80$.

204 The source of this issue is that heterozygosity is generally lower for SNPs with rare alleles
205 (where most individuals are homozygous for the common allele) than for SNPs with
206 common alleles. For instance, for the $n = 3$ subsamples analysed in individual runs, all SNPs

207 have minor alleles at 0.5 frequency when $MAC \geq 3$ is applied (Fig. 1a, e), leading to expected
208 heterozygosity of 0.5. As additional samples are added, SNPs with rare alleles become more
209 likely to be detected, leading to lower heterozygosity estimates (c.f. Fig. S1). As MAC
210 filtering is applied globally, heterozygosity is lower when populations are analysed together
211 (Fig. 1b, f) as global sample size is ten times larger in these runs. However, even in these
212 runs there were clear differences between SNP heterozygosity estimates for $n = 10$ ($N =$
213 100) and $n \leq 5$ ($N \leq 50$).

214 Considering these inconsistencies in SNP heterozygosity estimates when filtering datasets
215 with 'typical' settings for genetic structure, we reran the above analyses with $MAC \geq 1$ and
216 selecting all SNPs rather than one per RAD locus. These analyses thus considered variation
217 at all polymorphic sites including those with singletons and doubletons. We used a
218 maximum missing data threshold of 0% (or as specified), to avoid potential artefacts caused
219 by including sites with missing data (see Fig. 2). Including all polymorphic sites reduced the
220 bias from sample size; however, similar patterns of strong bias were still observed (Fig. 1i-p).
221 Thus, while including singletons and doubletons reduces sample size biases because rare
222 alleles are then more likely to be detected in small samples, the biases will nevertheless
223 persist when sites are filtered on the basis of polymorphism.

224

225

226 *Population comparisons using polymorphic sites only: missing data thresholds*

227 We investigated effects of missing data thresholds on SNP heterozygosity using the nested
228 subsamples from Group A, filtered with thresholds of 0% (i.e. no missing data allowed), 10%,
229 20%, 30%, 40%, or 50%. Thus in each case variation in heterozygosity was assessed in a
230 single population of size n ($n = 50, 40, 30, 20, 10, 5$). We compared results from the two
231 filtering protocols described previously: a standard protocol for assessing genetic structure
232 (one SNP per RADtag with $MAC \geq 3$) and one that retains all polymorphic sites ($MAC \geq 1$).

233 We see a considerable effect of missing data thresholds on SNP heterozygosity (Fig. 2a-f).

234 Samples of larger n were more strongly affected by choice of missing data threshold, with

235 stringent filtering tending to produce higher estimates. When 50 individuals were used with

236 MAC ≥ 3 filtering (Fig. 2a), a 10% missing data threshold (a common parameter setting)
237 produced an estimate for observed heterozygosity 1.22 times higher than filtering with a
238 30% threshold (also a common parameter setting). This effect was stronger with MAC ≥ 1
239 filtering (1.36 times higher; Fig. 2d). Expected heterozygosities were less biased by missing
240 data thresholds but effects were still evident (Fig. 2b,e).

241 The higher observed heterozygosity estimates at 0% versus 50% thresholds might be
242 expected if there is a correlation between errors and the presence of missing data at a site.
243 In that case, errors at monomorphic sites with high missing data could be read as low
244 frequency polymorphisms, pushing down heterozygosity estimates when N is large. Also,
245 when singletons and doubletons are included more low-frequency errors would be included
246 in calculations, leading to the stronger effects seen when MAC ≥ 1 filtering was applied.

247 Finally, the large population sizes in *Ae. aegypti* indicate this Malaysian sample is unlikely to
248 contain inbred individuals, and thus we do not expect observed and expected
249 heterozygosities to differ substantially. We note that when estimates of observed and
250 expected heterozygosity are compared, these parameters are most similar when filtering
251 with a 0% missing data threshold and start to diverge as this threshold is increased. These
252 divergences are consistent across filtering types, from ~ 1.10 at 0% missing data to ~ 1.27 at
253 20% and ~ 1.50 at 50%. Less stringent missing data thresholds may thus introduce artefacts
254 of differential observed and expected heterozygosities, which may lead to incorrect
255 inferences of local breeding patterns.

256 In light of these inconsistencies, SNP heterozygosity appears prone to bias, regardless of
257 whether filtering follows a typical protocol used for genetic structure or when considering
258 every polymorphic site. This bias is demonstrated in Fig. S1, which shows how, when
259 calculating SNP heterozygosity, the numerator remains proportionate to the number of
260 heterozygous sites regardless of sample size, but the denominator is consistently biased
261 downwards. This downward bias should diminish for very large N but for rare populations or
262 small budgets this cannot be solved by sequencing more individuals. This limits the potential
263 for SNP heterozygosity estimates in one study to inform the results of other studies.

264

265

Fig. 2: Boxplots showing effects of missing data thresholds on heterozygosity estimates. Observed (blue; a,d,g) and expected (red; b,e,h) heterozygosities have been derived from the nested subsamples from Group A following three filtering treatments: polymorphic sites only, $MAC \geq 3$ (a-c); polymorphic sites only, $MAC \geq 1$ (d-f); polymorphic and monomorphic sites (g-i). For each n , the subsample has been filtered using a progression of missing data thresholds: from left to right, 0%, 10%, 20%, 30%, 40%, 50%. Each estimate has undergone jack-knife resampling. Subfigures c,f,i aggregate results across all subsamples to show how observed (left) and expected (right) heterozygosities diverge with less stringent missing data thresholds. All Y-axes use a log-10 scale.

267

268

269

270 **Estimates based on all polymorphic and monomorphic sites (autosomal**
271 **heterozygosity)**

272 Given the above challenges, we next explored how sample size and missing data affect
273 autosomal heterozygosity, which considers both monomorphic and polymorphic
274 nucleotides. We ran analyses on identical datasets to those used previously. For filtering, we
275 used no MAC cut-off, and estimated heterozygosity across every site rather than every
276 polymorphic site. In the output from the Stacks.v2 program "Populations", this corresponds

277 to the entries in the “# All positions (variant and fixed)” subsection. We used a maximum
278 missing data threshold of 0% (unless specified differently).

279

280 *Population comparisons using all sites: effects of sample size*

281 When considering variation at all nucleotide sites, observed and expected heterozygosity
282 estimates are far less affected by N than SNP heterozygosity estimates (Fig. 1q-x). Though
283 there was some variability among subsamples of smaller n, observed heterozygosities of
284 ~0.00039 and expected heterozygosities of ~0.00040 were consistently recorded. The
285 similar estimates for these two parameters match expectations for this sample of Malaysian
286 *Ae. aegypti*, where inbreeding is unlikely given the large size of mosquito populations and
287 the spatial distribution of sampling.

288 These consistent estimates are expected when all sites are taken into account because for
289 smaller samples the higher frequency of heterozygotes at polymorphic sites will be offset by
290 the lower number of polymorphic sites overall. Heterozygosity estimates from a set of
291 individuals thus correlate directly with population heterozygosity because sites are not first
292 filtered by polymorphism (Fig. S1). In this sense, autosomal heterozygosity is a parameter
293 that is both more robust to variation in study design and also a more accurate measure of
294 genetic variation which can be used in comparisons across studies and organisms (Westbury
295 et al., 2019).

296

297

298 *Population comparisons with all sites: effects of missing data thresholds*

299 Missing data thresholds had a smaller effect on autosomal heterozygosity than on SNP
300 heterozygosity (Fig. 2g,h). Nevertheless, the same problematic pattern is clear in the
301 divergence between observed and expected heterozygosities when sites with missing data
302 are included, which was of equivalent magnitude to divergences in SNP heterozygosity (Fig.
303 2 i). Considering these results, we propose that heterozygosity estimation should exclude
304 nucleotide sites that have any missing genotypes, as these may be more likely to contain
305 errors or otherwise skew parameter estimates. While this filtering might at first appear

306 overly stringent, autosomal heterozygosity is calculated using far more sites than SNP
307 heterozygosity, and should normally be based on sufficient sites even after strict filtering.
308 For example, when a 20% missing data threshold is used to estimate SNP heterozygosity
309 ($MAC \geq 1$) in 50 individuals, heterozygosity is estimated from 95,293 polymorphic sites.
310 When a 0% missing data threshold is used to estimate autosomal heterozygosity ($MAC \geq 0$),
311 heterozygosity is estimated from 7,813,360 sites, of which 17,968 are polymorphic. This
312 does not imply that the consistency in autosomal heterozygosity estimates is due to a larger
313 number of sites; an increase in the number of sites will not resolve biases in SNP
314 heterozygosity which reflect the sample size of individuals rather than sites (Fig. S1).
315 Similarly, the specific number of nucleotides used in autosomal heterozygosity calculations
316 may vary across studies, but should accurately reflect the degree of variation across the
317 genome.

318

319

320 **Multiple population considerations**

321 We first estimated heterozygosity for the four *K. scurra* populations using equal sample sizes
322 of 10. Fig. S2 compares results of a 0% missing data threshold against a 20% threshold,
323 showing that at 20%, the ratio of observed to expected heterozygosity is higher ($\bar{x} = 1.156$)
324 than at 0% ($\bar{x} = 1.091$), supporting our previous findings that missing data may bias the ratio
325 of heterozygosities. Accordingly, we used a 0% missing data threshold in the following
326 analyses.

327

328 *Population comparisons: local sample size variation*

329 Although global sample size analysed above had little effect on autosomal heterozygosity
330 when $n \geq 10$, we have yet to consider differences in sample size among populations. We
331 estimated heterozygosity for the four *K. scurra* populations with one population (Goulburn)
332 set at either half (5,10,10,10) or double (10,5,5,5) the size of the other populations
333 compared to an equal population size. We compared results for autosomal heterozygosity
334 and SNP heterozygosity following previous filtering settings ($MAC \geq 3$ and $MAC \geq 1$).

335 When 10 individuals are analysed from each population, the Goulburn, Hall and
336 Wallendbeen populations all have similar heterozygosities, while Cooma is much lower (Fig.
337 3). There was no strong effect from unequal sample size for either autosomal heterozygosity
338 (Fig. 3 e,f) or for SNP heterozygosity using all polymorphic sites (Fig. 3c,d), but filtering at
339 $MAC \geq 3$ revealed such an effect (Fig. 3a,b). The Goulburn population had either higher or
340 lower heterozygosity than the Hall and Wallendbeen populations, depending on whether
341 Goulburn had a greater or smaller n. This bias could lead to misinterpretation of relative SNP
342 heterozygosities within studies when populations with different sample sizes are analysed
343 together.

344

345

Fig. 3: Effects of differential local sample size on heterozygosity estimates. Observed (blue; a,c,e) and expected (red; b,d,f) heterozygosities have undergone three filtering treatments: polymorphic sites only, $MAC \geq 3$ (a,b); polymorphic sites only, $MAC \geq 1$ (c,d); polymorphic and monomorphic sites (e,f). Numbers in brackets indicate sample sizes for the four *K. scurra* populations sequentially (Goulburn to Wallendbeen). Shading reflects similarity among numbers.

346

347

348 *Population comparisons: impact of population structure*

349 Next, we investigate the effects of combining genetically differentiated populations in an
350 analysis. In terms of mtDNA variation and DArT SNPs, Cooma was separate from the other
351 populations and particularly Wallendbeen (Hoffmann et al., 2020). We estimated SNP and
352 autosomal heterozygosities for each of the four *K. scurra* populations analysed individually
353 (i.e. with each population in a separate Stacks run) and compared this to populations
354 analysed together (i.e. with all populations in a single Stacks run).

355 Autosomal heterozygosity is unaffected by whether differentiated populations are analysed
356 individually or together (Fig. 4b). However, strong biases on SNP heterozygosity are evident

357 (Fig. 4a). When the four populations are analysed individually, estimates are much higher
358 than when analysed together. Additionally, the population at Cooma, which otherwise
359 recorded the lowest heterozygosity of the four populations, has higher heterozygosity than
360 all other populations when analysed by itself.

361

Fig. 4: Effects of population genetic structure on heterozygosity estimates. Observed SNP heterozygosities (a) and autosomal heterozygosities (b) are presented for the four *K. scurra* populations which have either been analysed together in a single run or in separate runs individually. The number of sites (c) and number of locally polymorphic sites (d) retained after filtering are also presented. Shading reflects similarity among numbers.

362

363

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

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

374 These findings show how SNP heterozygosity estimates represent different parameters
375 when populations are analysed in individual runs compared to when they are analysed
376 alongside other populations. When analysed individually, the SNP heterozygosity of a
377 population is equal to autosomal heterozygosity multiplied by the proportion of sites that
378 are polymorphic (Fig. S1). When analysed with other populations, the SNP heterozygosity of
379 a population will be shaped by whichever other populations are included, as the structure
380 between these populations will determine which locally monomorphic sites are called as
381 SNPs. Accordingly, SNP heterozygosities, if they are to be reported, should probably be
382 calculated from populations analysed individually, and the total number of polymorphic
383 sites should also be reported for each population to provide further context to the
384 heterozygosity estimates. For autosomal heterozygosity, analysing multiple populations at
385 once introduces no biases while conferring no advantage, but does reduce the number of
386 retained sites (Fig. 4 c,d). It follows that calculations of autosomal heterozygosity should
387 analyse each population in individual runs. For observed autosomal heterozygosity, this
388 could be extended to analysing each individual in turn if needed.

389

390

391 **Comparing heterozygosity estimates**

392 A final consideration concerns how to interpret heterozygosity estimates across studies. We
393 have proposed several guidelines for filtering data to allow cross-study comparisons. The
394 most important of these is that heterozygosity estimates should be derived from variation
395 at both monomorphic and polymorphic sites. Table 1 compares variation in SNP
396 heterozygosity with variation in autosomal heterozygosity for the four *K. scurra* populations

397 analysed individually. We did not compare populations when analysed together due to the
398 confoundment of SNP heterozygosities in these analyses (Fig. 4).

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

404

405

406 Table 1. Comparison of observed and expected SNP heterozygosity and autosomal
407 heterozygosity in four *K. scurra* populations when estimated individually. Calculations are
408 based on the results from Fig. 4. MAX/MIN is the ratio between the largest score and the
409 smallest score.

	SNP heterozygosity (MAC \geq 1)		Autosomal heterozygosity	
	Observed	Expected	Observed	Expected
MAX/MIN	1.164	1.064	1.983	2.230
Coefficient of variance	0.063	0.024	0.236	0.267

410

411

412

413 Discussion

414 Comparisons of heterozygosity across populations and species are frequently used to inform
415 management decisions in conservation programs. An example of a relevant management
416 action involves selecting populations for genetic and evolutionary rescue, which aims to
417 decrease levels of inbreeding and increase levels of genetic variation in target populations

418 through targeted introductions of individuals from other populations (Hoffmann et al.,
419 2020; Whiteley et al., 2015). The usefulness of source populations for conservation
420 translocations is to a large extent determined by their genetic variability (Ørsted et al., 2019;
421 Reid et al., 2016). Tracking changes in heterozygosity across time can also be a worthwhile
422 means of tracking the genetic health of threatened populations (Mitrovski et al., 2008) and
423 is particularly useful for determining outcomes of management interventions (Weeks et al.,
424 2017). All of these objectives require that heterozygosity estimates are comparable across
425 populations within a study and across different studies.

426 In this paper, we have shown that filtering genome-wide sequence data using optimal
427 settings for detecting genetic structure will produce heterozygosity estimates that are
428 poorly-suited to these comparisons. Specifically we show that heterozygosity estimates that
429 consider only polymorphic sites (SNP heterozygosity) are always biased by global sample
430 size (N), with smaller sample sizes producing larger heterozygosity estimates.

431 Heterozygosity estimates that consider monomorphic and polymorphic sites (autosomal
432 heterozygosity) do not suffer from these biases. We also found that when sites with missing
433 data are included, observed and expected heterozygosity estimates diverge, with the
434 divergence proportional to the amount of missing data permitted. When multiple
435 populations were analysed together, SNP heterozygosity estimates were additionally biased
436 by allele frequency differences among populations. While analysing populations together
437 did not bias autosomal heterozygosity, it conferred no advantages over analysing
438 populations individually but reduced the number of available sites due to missing data
439 filtering.

440 Following this, we propose three general guidelines that should help meaningful
441 comparisons: (i) studies aiming to summarise population genetic variation should report
442 autosomal heterozygosity, either by itself or alongside SNP heterozygosity; (ii) sites with
443 missing data should be omitted from heterozygosity calculations; and (iii) populations
444 should be analysed in independent runs. Although we have not explicitly investigated the
445 importance of sequencing coverage, this is widely known to be critical for accurately
446 identifying heterozygotes (Nielsen et al., 2011). This being the case, our findings that
447 heterozygosity estimates can be consistent even at low n (Fig. 7a) point to the optimal
448 design for heterozygosity being deep sequencing of a small number of individuals (perhaps

449 5-10) from each population, rather than shallower sequencing of many individuals. Previous
450 work has found these numbers to be adequate (Nazareno, Bemmels, Dick, & Lohmann,
451 2017).

452 SNP heterozygosity is frequently the only measure of heterozygosity reported (Bock et al.,
453 2018; Chen et al., 2016; Jones et al., 2012; Mathur et al., 2019; Surbakti et al., 2020).
454 Although there may be cases where SNP heterozygosity is the appropriate parameter, it will
455 be subject to biases from sample size that do not affect autosomal heterozygosity, making
456 the latter a better 'default' choice for reporting variation in populations. As taxa of
457 conservation interest are frequently rare or difficult to sample in large numbers, large
458 sample sizes for all populations would be difficult to achieve. SNP heterozygosity also does
459 not capture all the variation in the genome. For instance, *K. scurra* from Cooma had a higher
460 SNP heterozygosity estimate (when calculated in an independent run) yet fewer
461 polymorphic sites than the other populations; this low level of polymorphism was evident in
462 the autosomal heterozygosity estimate.

463 While we advocate autosomal heterozygosity as a default choice for reporting genome-wide
464 averages, there may be circumstances where variation at a specific set of sites is of interest
465 (e.g. Chen et al., 2016). In this case, heterozygosity can be estimated free of bias provided
466 that this specific set of sites is not further filtered by polymorphism when new populations
467 are analysed, and thus sites should be retained even if the new population is locally
468 monomorphic. Returning to the work of Nazareno et al. (2017), the consistency of their
469 heterozygosity estimates across subsamples of different size is due to these subsamples not
470 being refiltered by polymorphism. While Nazareno et al. (2017) effectively demonstrates
471 how heterozygosity can be estimated with small samples, this approach would not be
472 applicable to comparisons across populations. Likewise, assessing variation at a specific set
473 of sites may evade sample size biases, but other biases may be introduced when assessing
474 variation in populations with different allele frequencies to those in populations used to
475 select the initial set of variable sites.

476 Whole-genome sequencing studies frequently report autosomal rather than SNP
477 heterozygosity (Gopalakrishnan et al., 2017; Westbury et al., 2019). However, this
478 methodology has been less commonly applied in reduced-representation sequencing
479 studies such as those using RADseq or DArTseq markers. One reason for this may be that

480 whole-genome studies frequently analyse single individuals rather than a sample, and as
481 SNP heterozygosity calculated from a single diploid individual will always be equal to 1 (Fig.
482 S1) this approach would not be considered. Another reason may be that reduced-
483 representation protocols are commonly applied to wild populations of understudied
484 organisms, where the first step is typically to analyse genetic structure among populations.
485 Monomorphic sites are uninformative for most analyses of genetic structure and thus these
486 are typically removed during filtering, and this remaining set of SNPs will be used to
487 estimate both genetic structure and heterozygosity. Here we propose that analysis of
488 genetic structure (variation between populations) and heterozygosity (variation within
489 populations) be treated as distinct targets and filtered using different parameter settings.

490 When expected heterozygosity is higher than observed heterozygosity, it is often treated as
491 evidence for local inbreeding (Hoffmann et al., 2020), an important parameter that can
492 indicate a need for genetic intervention in threatened species (Ralls et al., 2018). However,
493 Fig. 2 shows that inferences of inbreeding in the presence of missing data may be
494 confounded, as observed and expected heterozygosities quickly diverge as sites with
495 missing data are included, leading to $F_{IS} > 0$. These opposing effects of missing data filters on
496 observed and expected heterozygosities are consistent with sequencing error rates being
497 higher at nucleotide positions where there are some missing genotypes. A possible
498 explanation is that sequencing errors are more likely at monomorphic sites; these sites are
499 more common, and more likely to be retained than errors at polymorphic sites which may
500 introduce a third allele and be removed from the dataset (most filtering protocols retain
501 only biallelic sites). Monomorphic sites with errors may therefore be coded as low-
502 frequency SNPs. Including them could affect the observed and expected autosomal
503 heterozygosity estimates differently depending on whether errors are mostly coded as
504 homozygous or heterozygous.

505 Despite these issues, the *K. scurra* data also indicate clear instances where inbreeding levels
506 differ across populations. We note that for *K. scurra* from Cooma, observed and expected
507 heterozygosities were almost identical, while expected heterozygosities are 8-15% higher in
508 the other populations, suggesting a situation where inbreeding in Cooma is low but genetic
509 variability is also low. An accurate assessment of inbreeding in populations versus low
510 genetic variation is important when making recommendations around genetic mixing of

511 threatened populations, which can target both the masking of deleterious genes expressed
512 as a consequence of inbreeding as well as problems from low levels of genetic variability
513 (Hoffmann et al., 2020; Ralls et al., 2018; Weeks et al., 2011). It would be worth further
514 evaluating optimal filtering strategies for assessing the ratio of observed and expected
515 heterozygosities, using a set of populations where inbreeding is known to occur in some
516 populations but not others.

517 This study has highlighted issues with estimating population heterozygosities from SNP data
518 directly, and shown that autosomal heterozygosity estimates are more robust to the
519 influence of sample size and are likely to be more comparable across studies. We provide
520 general guidelines for estimating population heterozygosity from genome-wide sequence
521 data that are usually different from guidelines for estimating other population genetic
522 parameters such as gene flow, population structure, relatedness and effective population
523 size. As in previous assessments (Linck & Battey, 2019), our results demonstrate that SNP
524 datasets need to be carefully evaluated when they are used to obtain genetic parameters
525 for populations that inform management decisions.

526

527

528 **Acknowledgements**

529 This research was supported by the Australian Research Council (Discovery Grant
530 DP190100990), the Wellcome Trust (Grant no. 108508) and the National Health and Medical
531 Research Council (Program Grant no. 1132412; Fellowship Grant no. 1118640), and
532 facilitated by use of the Nectar Research Cloud.

533

534 **Data Archiving**

535 Aligned .bam files for 100 *Ae. aegypti* are available through the NCBI SRA, accession number
536 PRJNA735025. Sequence data for *K. scurra* is available from
537 <https://www.ncbi.nlm.nih.gov/bioproject/702007>.

538

539 Author Contributions

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

543

544

545 References

546

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

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

554 Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An
555 analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124–3140.
556 <https://doi.org/10.1111/mec.12354>

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

560 Fitzpatrick, S. W., Bradburd, G. S., Kremer, C. T., Salerno, P. E., Angeloni, L. M., & Funk, W. C.
561 (2020). Genomic and Fitness Consequences of Genetic Rescue in Wild Populations.
562 *Current Biology*, 30(3), 517-522.e5. <https://doi.org/10.1016/j.cub.2019.11.062>

563 Fontaine, A., Filipović, I., Fansiri, T., Hoffmann, A. A., Cheng, C., Kirkpatrick, M., Rašić, G., &
564 Lambrechts, L. (2017). Extensive genetic differentiation between homomorphic sex

565 chromosomes in the mosquito vector, *Aedes aegypti*. *Genome Biology and Evolution*,
566 9(9), 2322–2335. <https://doi.org/10.1093/gbe/evx171>

567 Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., Morin, P. A.,
568 Narum, S. R., O'Brien, S. J., Roffler, G., Templin, W. D., Sunnucks, P., Strait, J.,
569 Warheit, K. I., Seamons, T. R., Wenburg, J., Olsen, J., & Luikart, G. (2016). Genomics
570 in Conservation: Case Studies and Bridging the Gap between Data and Application.
571 *Trends in Ecology and Evolution*, 31(2), 81-83.
572 <https://doi.org/10.1016/j.tree.2015.10.009>

573 Gopalakrishnan, S., Samaniego Castruita, J. A., Sinding, M. H. S., Kuderna, L. F. K., Räikkönen,
574 J., Petersen, B., Sicheritz-Ponten, T., Larson, G., Orlando, L., Marques-Bonet, T.,
575 Hansen, A. J., Dalén, L., & Gilbert, M. T. P. (2017). The wolf reference genome
576 sequence (*Canis lupus lupus*) and its implications for *Canis* spp. Population genomics.
577 *BMC Genomics*, 18(1), 495. <https://doi.org/10.1186/s12864-017-3883-3>

578 Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2017). DARTR: An R package to
579 facilitate analysis of SNP data generated from reduced representation genome
580 sequencing. *Molecular Ecology Resources*, 18(3), 1–9. [https://doi.org/10.1111/1755-](https://doi.org/10.1111/1755-0998.12745)
581 [0998.12745](https://doi.org/10.1111/1755-0998.12745)

582 Hoffmann, A. A., Sgrò, C. M., & Kristensen, T. N. (2017). Revisiting Adaptive Potential,
583 Population Size, and Conservation. *Trends in Ecology and Evolution*, 32(7), 506-517.
584 <https://doi.org/10.1016/j.tree.2017.03.012>

585 Hoffmann, A. A., White, V., Jasper, M., Yagui, H., Sinclair, S., & Kearney, M. (2020). An
586 endangered flightless grasshopper with strong genetic structure maintains
587 population genetic variation despite extensive habitat loss. *Ecology and Evolution*,
588 11(10), 5364-5380. <https://doi.org/10.22541/AU.160403053.38828134/V1>

589 Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010).
590 Population Genomics of Parallel Adaptation in Threespine Stickleback using
591 Sequenced RAD Tags. *PLoS Genetics*, 6(2), e1000862.
592 <https://doi.org/10.1371/journal.pgen.1000862>

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

- 597 Jones, F. C., Chan, Y. F., Schmutz, J., Grimwood, J., Brady, S. D., Southwick, A. M., Absher, D.
598 M., Myers, R. M., Reimchen, T. E., Deagle, B. E., Schluter, D., & Kingsley, D. M. (2012).
599 A genome-wide SNP genotyping array reveals patterns of global and repeated
600 species-pair divergence in sticklebacks. *Current Biology*, *22*(1), 83–90.
601 <https://doi.org/10.1016/j.cub.2011.11.045>
- 602 Lemopoulos, A., Prokkola, J. M., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P.,
603 Koljonen, M. L., Koskiniemi, J., & Vainikka, A. (2019). Comparing RADseq and
604 microsatellites for estimating genetic diversity and relatedness—Implications for
605 brown trout conservation. *Ecology and Evolution*, *9*(4), 2106–2120.
606 <https://doi.org/10.1002/ece3.4905>
- 607 Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population
608 structure inference with genomic data sets. *Molecular Ecology Resources*, *19*(3),
609 639–647. <https://doi.org/10.1111/1755-0998.12995>
- 610 Maroso, F., Franch, R., Dalla Rovere, G., Arculeo, M., & Bargelloni, L. (2016). RAD SNP
611 markers as a tool for conservation of dolphinfish *Coryphaena hippurus* in the
612 Mediterranean Sea: Identification of subtle genetic structure and assessment of
613 populations sex-ratios. *Marine Genomics*, *28*, 57–62.
614 <https://doi.org/10.1016/j.margen.2016.07.003>
- 615 Mathur, S., Tomeček, J. M., Heniff, A., Luna, R., & DeWoody, J. A. (2019). Evidence of genetic
616 erosion in a peripheral population of a North American game bird: The Montezuma
617 quail (*Cyrtonyx montezumae*). *Conservation Genetics*, *20*(6), 1369–1381.
618 <https://doi.org/10.1007/s10592-019-01218-9>
- 619 Mitrovski, P., Hoffmann, A. A., Heinze, D. A., & Weeks, A. R. (2008). Rapid loss of genetic
620 variation in an endangered possum. *Biology Letters*, *4*(1), 134–138.
621 <https://doi.org/10.1098/rsbl.2007.0454>
- 622 Mulvena, S. R., Pierson, J. C., Farquharson, K. A., McLennan, E. A., Hogg, C. J., & Grueber, C.
623 E. (2020). Investigating inbreeding in a free-ranging, captive population of an
624 Australian marsupial. *Conservation Genetics*, *21*(4), 665–675.
625 <https://doi.org/10.1007/s10592-020-01278-2>
- 626 Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes
627 for population genomics: An empirical study from an Amazonian plant species.

- 628 *Molecular Ecology Resources*, 17(6), 1136–1147. <https://doi.org/10.1111/1755-0998.12654>
- 629
- 630 Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from
631 next-generation sequencing data. *Nature Reviews Genetics*, 12(6), 443–451.
632 <https://doi.org/10.1038/nrg2986>
- 633 Ørsted, M., Hoffmann, A. A., Sverrisdóttir, E., Nielsen, K. L., & Kristensen, T. N. (2019).
634 Genomic variation predicts adaptive evolutionary responses better than population
635 bottleneck history. *PLoS Genetics*, 15(6).
636 <https://doi.org/10.1371/journal.pgen.1008205>
- 637 Ralls, K., Ballou, J. D., Dudash, M. R., Eldridge, M. D. B., Fenster, C. B., Lacy, R. C., Sunnucks,
638 P., & Frankham, R. (2018). Call for a Paradigm Shift in the Genetic Management of
639 Fragmented Populations. *Conservation Letters*, 11(2), e12412.
640 <https://doi.org/10.1111/conl.12412>
- 641 Reid, N. M., Proestou, D. A., Clark, B. W., Warren, W. C., Colbourne, J. K., Shaw, J. R.,
642 Karchner, S. I., Hahn, M. E., Nacci, D., Oleksiak, M. F., Crawford, D. L., & Whitehead,
643 A. (2016). The genomic landscape of rapid repeated evolutionary adaptation to toxic
644 pollution in wild fish. *Science*, 354(6317), 1305–1308.
645 <https://doi.org/10.1126/science.aah4993>
- 646 Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients.
647 *Genetical Research*, 67(2), 175–185. <https://doi.org/10.1017/s0016672300033620>
- 648 Surbakti, S., Parker, H. G., McIntyre, J. K., Maury, H. K., Cairns, K. M., Selvig, M., Pangau-
649 Adam, M., Safonpo, A., Numberi, L., Runtuboi, D. Y. P., Davis, B. W., & Ostrander, E.
650 A. (2020). New Guinea highland wild dogs are the original New Guinea singing dogs.
651 *Proceedings of the National Academy of Sciences of the United States of America*,
652 117(39), 24369–24376. <https://doi.org/10.1073/pnas.2007242117>
- 653 Weeks, A. R., Heinze, D., Perrin, L., Stoklosa, J., Hoffmann, A. A., Van Rooyen, A., Kelly, T., &
654 Mansergh, I. (2017). Genetic rescue increases fitness and aids rapid recovery of an
655 endangered marsupial population. *Nature Communications*, 8(1).
656 <https://doi.org/10.1038/s41467-017-01182-3>
- 657 Weeks, A. R., Sgro, C. M., Young, A. G., Frankham, R., Mitchell, N. J., Miller, K. A., Byrne, M.,
658 Coates, D. J., Eldridge, M. D. B., Sunnucks, P., Breed, M. F., James, E. A., & Hoffmann,
659 A. A. (2011). Assessing the benefits and risks of translocations in changing

660 environments: A genetic perspective. *Evolutionary Applications*, 4(6), 709–725.
661 <https://doi.org/10.1111/j.1752-4571.2011.00192.x>

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

665 Whiteley, A. R., Fitzpatrick, S. W., Funk, W. C., & Tallmon, D. A. (2015). Genetic rescue to the
666 rescue. *Trends in Ecology and Evolution*, 30(1), 42-49.
667 <https://doi.org/10.1016/j.tree.2014.10.009>

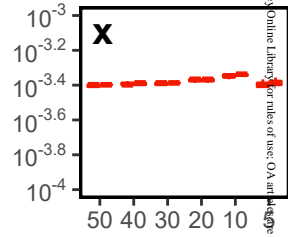
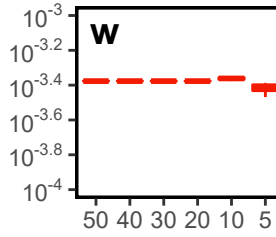
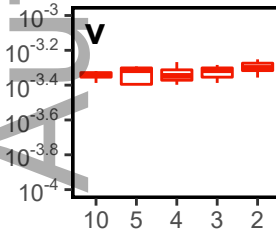
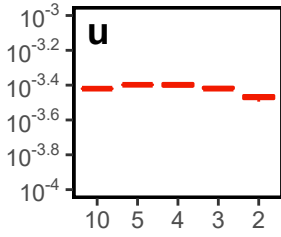
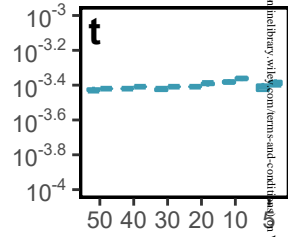
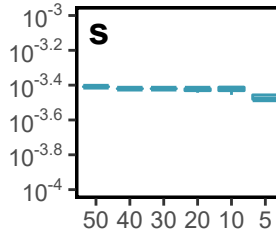
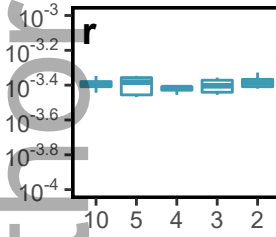
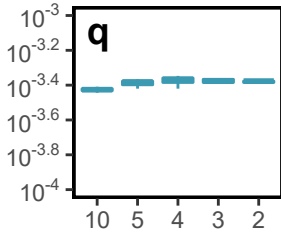
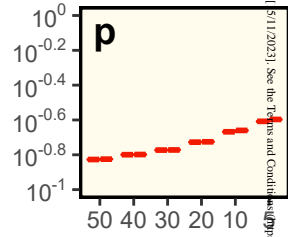
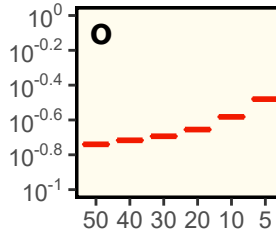
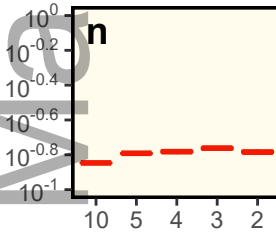
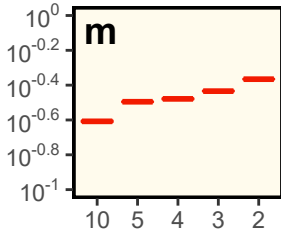
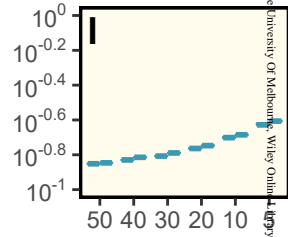
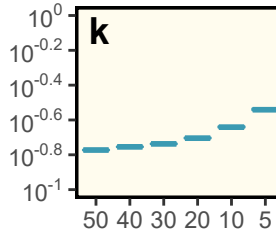
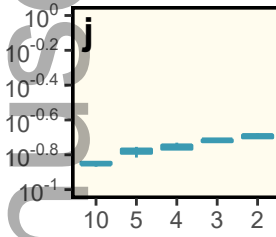
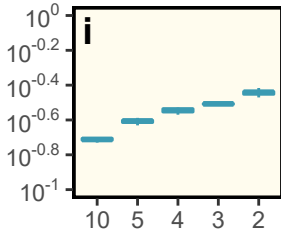
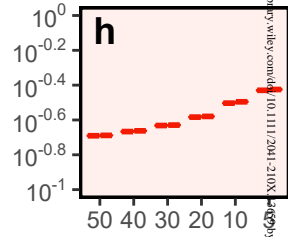
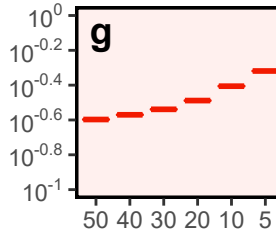
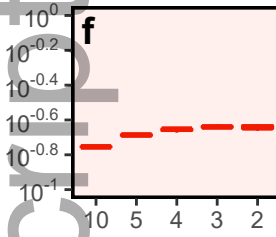
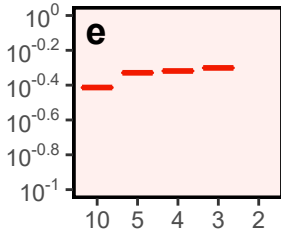
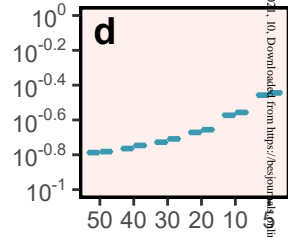
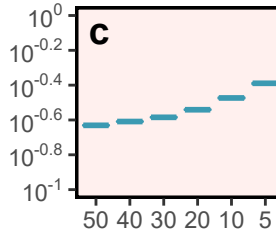
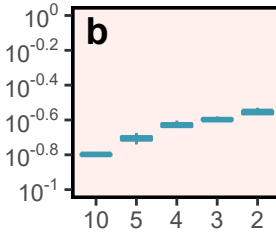
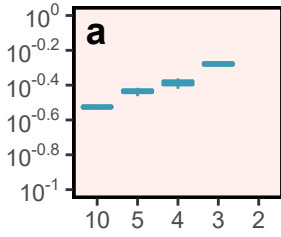
668 Wright, B. R., Grueber, C. E., Lott, M. J., Belov, K., Johnson, R. N., & Hogg, C. J. (2019). Impact
669 of reduced-representation sequencing protocols on detecting population structure in
670 a threatened marsupial. *Molecular Biology Reports*, 46(5), 5575–5580.
671 <https://doi.org/10.1007/s11033-019-04966-6>
672

10 subsamples
(individual runs)

10 subsamples
(run together)

Group A
(individually)

Groups A and B
(together)



Observed Heterozygosity

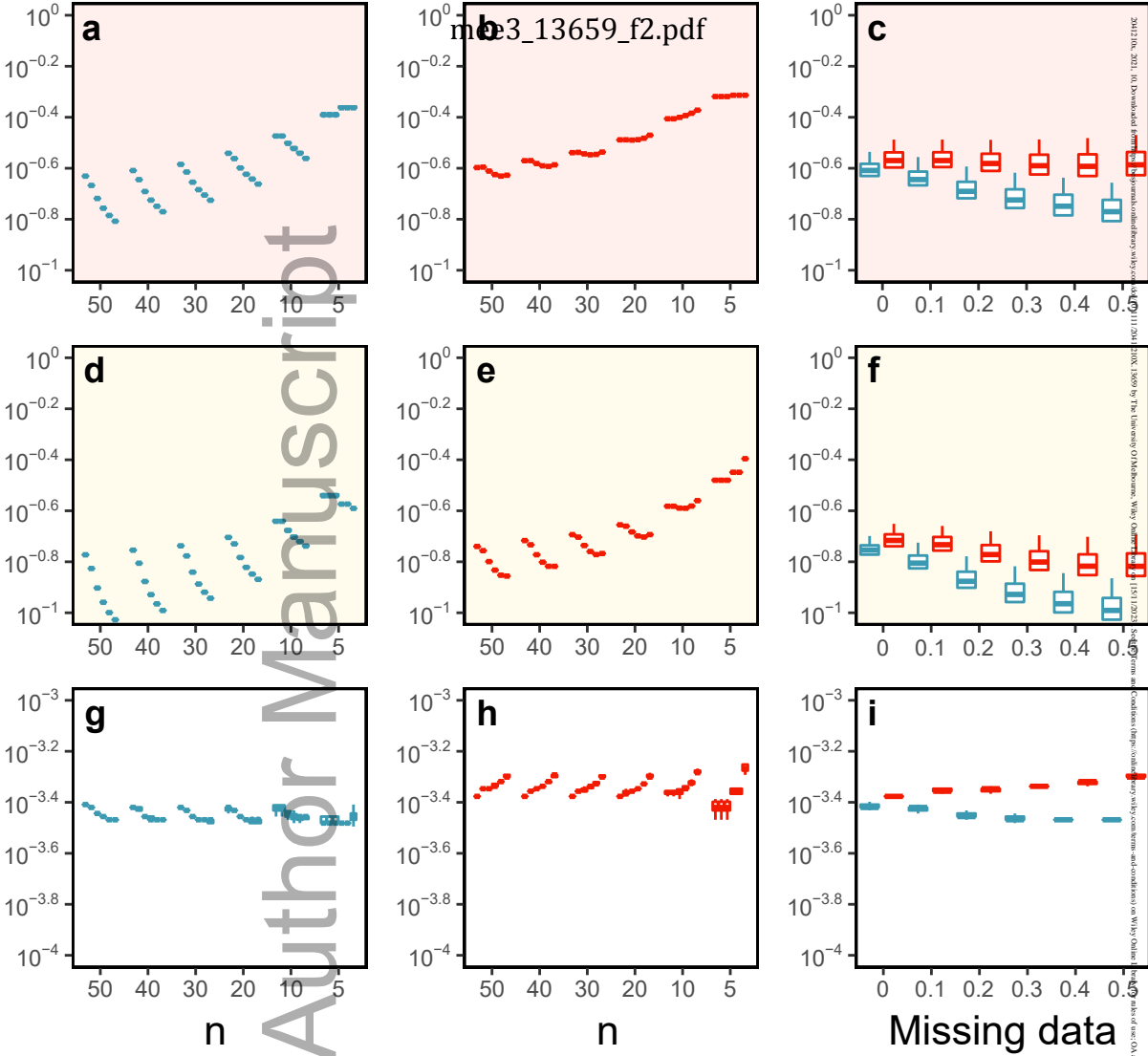
Expected Heterozygosity

SNP Heterozygosity, $MAC \geq 3$

SNP Heterozygosity, $MAC \geq 1$

Autosomal Heterozygosity

2024/10/20, 20:10:10 Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1365-3113.12108 by The University Of Melbourne, Wiley Online Library on [15/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



— Observed Heterozygosity

— Expected Heterozygosity

SNP Heterozygosity, $MAC \geq 3$

SNP Heterozygosity, $MAC \geq 1$

Autosomal Heterozygosity

This article is protected by copyright. All rights reserved.

a Observed (SNP MAC3) mee3_13659_f3.pdf

Goulburn	0.10434	0.09091	0.13365
Cooma	0.05185	0.05742	0.06084
Hall	0.11530	0.12599	0.13065
Wallendbeen	0.12204	0.13395	0.11562
	{10,10,10,10}	{5,10,10,10}	{10,5,5,5}

b Expected (SNP MAC3)

Goulburn	0.11182	0.09554	0.14469
Cooma	0.05466	0.06034	0.05766
Hall	0.12009	0.13212	0.13355
Wallendbeen	0.12499	0.13705	0.12227
	{10,10,10,10}	{5,10,10,10}	{10,5,5,5}

c Observed (SNP MAC1)

Goulburn	0.07515	0.07503	0.08664
Cooma	0.03790	0.04166	0.04506
Hall	0.08296	0.08721	0.10039
Wallendbeen	0.08938	0.09465	0.10076
	{10,10,10,10}	{5,10,10,10}	{10,5,5,5}

d Expected (SNP MAC1)

Goulburn	0.07935	0.07768	0.09399
Cooma	0.03704	0.03985	0.04264
Hall	0.08650	0.09345	0.09785
Wallendbeen	0.09530	0.09972	0.10125
	{10,10,10,10}	{5,10,10,10}	{10,5,5,5}

e Observed (Autosomal)

Goulburn	0.00312	0.00293	0.00331
Cooma	0.00153	0.00163	0.00172
Hall	0.00345	0.00341	0.00384
Wallendbeen	0.00371	0.00370	0.00386
	{10,10,10,10}	{5,10,10,10}	{10,5,5,5}

f Expected (Autosomal)

Goulburn	0.00330	0.00304	0.00360
Cooma	0.00154	0.00156	0.00163
Hall	0.00359	0.00365	0.00374
Wallendbeen	0.00396	0.00390	0.00388
	{10,10,10,10}	{5,10,10,10}	{10,5,5,5}

This article is protected by copyright. All rights reserved

2012-06-20 10:10 Downloaded from https://academic.oup.com/iagc/advance-article-abstract/doi/10.1093/iagc/iaab001/636891 by The University Of Melbourne user on 11 January 2021. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

a Observed (SNP MAG1) me3_13659_f4.pdf

Goulburn	0.07515	0.23314
Cooma	0.03790	0.27138
Hall	0.08296	0.23740
Wallendbeen	0.08938	0.23777

{10,10,10,10} 10 individually

b Observed (Autosomal)

Goulburn	0.00312	0.00339
Cooma	0.00158	0.00179
Hall	0.00345	0.00331
Wallendbeen	0.00371	0.00355

{10,10,10,10} 10 individually

c No. Sites

Goulburn	101968	365368
Cooma	101968	548373
Hall	101972	329382
Wallendbeen	101971	366453

{10,10,10,10} 10 individually

d No. Polymorphic sites

Goulburn	1356	5200
Cooma	562	3657
Hall	1408	4567
Wallendbeen	1557	5392

{10,10,10,10} 10 individually

This article is protected by copyright. All rights reserved

bioRxiv preprint doi: <https://doi.org/10.1101/2023.03.15.531111>; this version posted March 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.