Research Article **Open Access**

Jarrett D. Phillips, Rodger A. Gwiazdowski, Daniel Ashlock, Robert Hanner*

An exploration of sufficient sampling effort to describe intraspecific DNA barcode haplotype diversity: examples from the ray-finned fishes (Chordata: Actinopterygii)

DOI 10.1515/dna-2015-0008

Received February 26, 2015; accepted June 9, 2015

Abstract: Estimating appropriate sample sizes to measure species abundance and richness is a fundamental problem for most biodiversity research. In this study, we explore a method to measure sampling sufficiency based on haplotype diversity in the ray-finned fishes (Animalia: Chordata: Actinopterygii). To do this, we use linear regression and hypothesis testing methods on haplotype accumulation curves from DNA barcodes for 18 species of fishes, in the statistics platform R. We use a simple mathematical model to estimate sampling sufficiency from a sample-number based prediction of intraspecific haplotype diversity, given an assumption of equal haplotype frequencies. Our model finds that haplotype diversity for most of the 18 fish species remains largely unsampled, and this appears to be a result of small sample sizes. Lastly, we discuss how our overly simple model may be a useful starting point to develop future estimators for intraspecific sampling sufficiency in studies using DNA barcodes.

Keywords: Chao1 abundance estimator; DNA barcoding; haplotype accumulation curve; method of moments

***Corresponding author: Robert Hanner,** Centre for Biodiversity Genomics, Department of Integrative Biology, University of Guelph, Ontario, N1G 2W1 Canada, Email: rhanner@uoguelph.ca

Jarrett D. Phillips, Centre for Biodiversity Genomics, Department of Integrative Biology, University of Guelph, Ontario, N1G 2W1 Canada

Rodger A. Gwiazdowski, Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, N1G 2W1 Canada

Daniel Ashlock, Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, N1G 2W1 Canada

1 Introduction

Most biodiversity research requires an estimate of adequate sample sizes to achieve a study's objective [1]. Sample sizes that are sufficient to address research questions often depend on sampling methodologies and the organism being considered [2]. Adequate sample sizes involving molecular genetic measurements are directly related to a species' genetic variation. A common measurement of intraspecific variation is mitochondrial DNA (mtDNA) haplotype diversity, which is largely affected by underlying evolutionary biological processes such as gene flow and random genetic drift. As such, sample sizes sufficient to observe within-species mtDNA variation will vary widely across taxa. Haplotype diversity represents the prevalence of haplotypes at the population level and is analogous to the concept of heterozygosity at the locus level, except that the former pertains only to haploid data. A simple measure of haplotype diversity was first provided by [3] and is calculated as

$$h = \frac{N}{N-1}\sum_i \left(1 - p_i^2\right)$$

where N is the sample size and p_i represents the frequency of each haplotype in a given sample. Estimates of h (which range from 0-1) are greatly affected by sampling intensity, particularly undersampling, which has been observed especially for mtDNA markers [4]. Another widely used metric of haplotype variation is the absolute number of (unique) species haplotypes (used here throughout) which is comparable in magnitude to actual specimen sample sizes.

A standardized tool for genetic biodiversity assessment is DNA barcoding [5], because this method uses easily obtainable mtDNA diversity from the 5' cytochrome *c* oxidase subunit I (COI) gene to identify species. However, methods to describe a sample set required to observe a

full range of DNA barcode haplotypes within a species have not been well developed. A general consensus for adequate sample sizes for DNA barcode studies appears to be ~ 5-10 specimens per species [6]; however, this range is highly variable within the Barcode of Life Data Systems (BOLD) [7], owing to both the relative difficulty and cost of sample collection and mtDNA sequencing [4]. As such, previous studies incorporating DNA barcodes across various taxonomic groups have resulted in a wide range of intraspecific sampling effort: very few specimens in the case of rare species, to upwards of 500 individuals for some species of insects within BOLD.

Here, we share a brief exploration in estimating sampling sufficiency by observing intraspecific haplotype diversity in the ray-finned fishes (Animalia: Chordata: Actinopterygii), a group that is among the largest of all vertebrates, and also has a large number of DNA barcodes available within BOLD. In the present study, we define *sampling sufficiency* to be the sample size at which sampling accuracy is maximized and above which no new sampling information is likely to be gained. We recognize that estimating a sample size necessary to observe the range of mtDNA haplotype diversity within a species involves at least three measures: sample number, genetic dispersion and geographic dispersion [8]. Because geographic dispersion is multidimensional and because spatiotemporal metadata (e.g. GPS coordinates) are lacking for many fish species within BOLD, we focus only on exploring the dynamics of estimating intraspecific sample sufficiency based on sample number and genetic dispersion (as predicted haplotype diversity). To do this, we use haplotype accumulation curves calibrated by a simple variant of the statistical method of moments, which is a method of parameter estimation based on the law of large numbers [9]. Such a method provides a useful stopping criterion for specimen sampling above which no new haplotypes are likely to be observed. Haplotype accumulation curves provide a graphical way to assess the extent of haplotype sampling similar to the use of rarefaction curves to assess species richness [10]. Such curves depict the extent of saturation as a function of the number of specimens sampled and the number of haplotypes accumulated. Those species whose curves show rapid saturation indicate that much of the intraspecific haplotype diversity may have been sampled. Species curves showing little to no indication of asymptotic behavior suggest further sampling is necessary to document the extent of standing genetic variation present.

The issue of sampling intensity is rarely raised in relation to barcode studies, which often focus on maximizing the number of species sampled rather than exhaustively sampling any one species [6,11]. Thus, there are few prior studies exploring haplotype accumulation curves in relation to sample size estimation using DNA barcode data (e.g., fungi: [2]; butterflies: [6]; aphids: [12]). Of potential relevance to estimating sampling sufficiency for fishes is an analysis of mtDNA haplotype variation in Lake trout (*Salvelinus namaycush*) stocked into Lake Ontario [13]. Here, [13] found that a minimum of $n \approx 60$ individuals needed to be randomly sampled in order to observe with $\beta = 95\%$ confidence any one individual having a particular haplotype present at a frequency of at least $P = 5\%$ according to the equation

$$n = \frac{\ln(1-\beta)}{\ln(1-P)}.$$

Estimating sample sizes necessary for describing the genetic diversity of a species is also dependent on underlying biological processes, population structure as well as lineage history. Therefore estimates based on rigorous statistical considerations alone may not be adequate.

In this paper, we develop our ideas as an R-based workflow that uses DNA barcodes of actinopterygians, identified to species and retrieved from BOLD, to estimate intraspecific sample sizes that should adequately represent haplotype variation within a species.

2 Methods

2.1 Species retrieval from BOLD

All publicly accessible sequences from Actinopterygii were first retrieved from BOLD on May 30, 2014 using the keyword 'Actinopterygii'. Records were then searched manually for all species represented by at least 60 specimens, chosen as an *a priori* minimum inspired by [13]. This minimum sample size criterion was used in all subsequent steps of our pipeline to ensure quality control and integrity of selected species. A total of 12,210 specimens covering 107 species (mean: 115 specimens/species) from 16 orders, 46 families and 75 genera were found. All but three species had formal taxonomic names; the remaining were interim.

2.2 Sequence cleaning and processing

DNA barcode sequences were directly read from BOLD into R using the package 'SPIDER' (SPecies IDentity and

Evolution in R; [14]) using the functions search.BOLD(), to find specimens, and read.BOLD() which downloads sequences found by search.BOLD(). Sequences were written to FASTA files using the function write.dna() from the R package 'APE' (Analysis of Phylogenetics and Evolution; [15]). FASTA files were then read into MEGA6 (Molecular Evolutionary Genetics Analysis; [16]) as the start of a conservative sequence quality check and alignment workflow. Haplotype functions (described below) in both SPIDER and 'PEGAS' (Population and Evolutionary Genetics Analysis System; [17]) will overestimate haplotype counts if missing or ambiguous sequence data are present.

The first step of data curation involved removing all GenBank specimens, as these often lack metadata requirements sufficient for compliance with BARCODE data [7,18]. In this dataset, GenBank specimens corresponded to the identifiers ANGBF, CYTC, GBGCA and GBGC. Next, sequences were aligned using MUSCLE (MUltiple Sequence Comparison by Log Expectation; [19]) with default parameters for all species and then trimmed to 652 bp. The presence of ambiguous bases was handled using the functions checkDNA() in SPIDER and base.freq() in APE. The function checkDNA() gives the number of nucleotide base positions that consist of missing or ambiguous data for each specimen; whereas, the function base.freq() outputs average nucleotide (A, C, G and T) and ambiguous/missing base frequencies (R, M, W, S, K, Y, V, H, D, B, N, - and ?). The latter function was used as a criterion to ensure no missing or ambiguous data were present within alignments (i.e., it was ensured that these frequencies were all equal to 0). Lastly, alignments were translated to amino acids using the vertebrate mitochondrial code table in MEGA6 and all sequences with stop codons were removed. Species not meeting our minimum sample size criterion were discarded.

Table 1. Intraspecific haplotype and specimen sample sizes for the 18 Actinopterygii species calculated from the proposed sampling model. All values are rounded up to the nearest whole number.

Species	In BOLD	N	H	H*	N*	N* – N	H* – H	% sampled	% missing
Siamese fighting fish (*Betta splendens*)	145	76	4	10	190	114	6	40	60
Brook stickleback (*Culaea inconstans*)	119	87	19	190	870	783	171	10	90
Johnny darter (*Etheostoma nigrum*)	226	174	24	300	2175	2001	276	8	92
Tessellated darter (*Etheostoma olmstedi*)	159	127	19	190	1270	1143	171	10	90
Orangebelly darter (*Etheostoma radiosum*)	118	88	32	528	1452	1364	496	6	94
Golden shiner (*Notemigonus crysoleucas*)	332	262	20	210	2751	2489	190	10	90
Chum salmon (*Oncorhynchus keta*)	106	75	8	36	338	263	28	22	78
Coho salmon (*Oncorhynchus kisutch*)	166	145	11	66	870	725	55	17	83
Rainbow trout (*Oncorhynchus mykiss*)	284	224	18	171	2128	1904	153	11	89
Sockeye salmon (*Oncorhynchus nerka*)	78	68	9	45	340	272	36	20	80
Chinook salmon (*Oncorhynchus tshawytscha*)	236	213	12	78	1385	1172	66	15	85
Fathead minnow (*Pimephales promelas*)	206	175	13	91	1225	1050	78	14	86
Barred sorubim (*Pseudoplatystoma fasciatum*)	145	126	20	210	1323	1197	190	10	90
Western blacknose dace (*Rhinichthys obtusus*)	125	94	24	300	1175	1081	276	8	92
Rockfish (*Sebastes* sp.)	198	98	2	3	147	49	1	67	33
Longfin damselfish (*Stegastes diencaeus*)	379	347	30	465	5379	5032	435	6	94
Beau Gregory (*Stegastes leucostictus*)	293	266	13	91	1862	1596	78	14	86
Blue-striped cave goby (*Trimma tevegae*)	78	70	20	210	735	665	190	10	90

After sequence processing, the useable dataset was considerably reduced (Table 1) consisting of 18 species (one interim) (2715 specimens; 68-347 specimens/species; mean: 151 specimens/species) from 6 orders, 9 families and 11 genera. Because sequences clustered according to Barcode Index Numbers (BINs) [20] closely mirror actual species, the one unnamed species, *Sebastes* sp., was tentatively considered as a single species due to being associated with only a single BIN (i.e., no other specimens or species shared that BIN). Cleaned alignments were exported as FASTA files from MEGA6 and imported into R using the APE function read.FASTA().

2.3 Haplotype accumulation curves

The number of haplotypes and their corresponding frequencies were calculated using PEGAS with the function haplotype(). Haplotype accumulation curves were generated using the functions haploAccum() and plot. haploAccum() from SPIDER. The function haploAccum() carries out haplotype accumulation without replacement through random permutation subsampling using the function argument 'random'. Specimen and haplotype counts from haploAccum() were then plotted with plot. haploAccum(). A total of 1000 permutations were used in generating haplotype accumulation curves for all 18 species. 1000 permutations was selected in order to reduce noisiness and increase smoothness of generated curves as the use of too few permutations (e.g. 100) resulted in very stochastic-looking accumulation curves. Permutation sizes larger than 1000 typically resulted in significantly increased computation time, but overall differed little in terms of smoothness from curves generated using our chosen permutation size. 95% confidence intervals were also computed for all curves and displayed as error bars. Since haplotype accumulation performed by haploAccum() is done in a random fashion, resulting haplotype accumulation curves will vary slightly between runs.

2.4 Statistical analyses

Haplotype diversity and sampling sufficiency for all 18 species were assessed in two ways: (1) linear regression analyses to evaluate the magnitude of calculated slopes and formal hypothesis tests on slope estimates; and (2) estimation of sample sizes required to represent intraspecific haplotype diversity. Linear regression analyses, based on the last 10 points occurring on haplotype accumulation curves, were carried out using the R functions lm() and summary() [21]. Species whose curve slopes ranged from 0.01 and above were considered to be undersampled; whereas, those with curve slopes below 0.01 were deemed to be almost fully sampled [22]. One-sided hypothesis tests carried out on slope estimate outputs from summary() were as follows: H_0: $\beta_1 = 0$ versus H_1: $\beta_1 > 0$. In all cases, the null hypothesis for evidence of no additional haplotypes was tested against the alternative hypothesis of additional haplotypes at the 5% significance level.

2.5 Estimating haplotype diversity

A nonparametric estimate of the sample size needed to account for all haplotype diversity for each of the 18 species was determined using information on the observed number of specimens and haplotypes. We used the Chao1 estimate of abundance [23] that uses the observed sample size and observed haplotype number to determine appropriate minimum sample size estimates for both haplotype diversity and intraspecific sampling sufficiency. The mathematical approach we used is analogous to a simple mark-recapture technique used widely in ecological settings to estimate population sizes of mobile animals collected from multiple sites [24]. A key assumption of our model is that all haplotypes occur with equal frequency in the sampling for a species. That is, haplotypes are assumed to follow a discrete uniform distribution. This is analogous to the assumption of equal catchability of animals in the mark-recapture model [24, 25]. For example, if $N = 100$ specimens of a given species are randomly sampled without replacement and $H = 10$ haplotypes are observed, then we should expect each haplotype to be represented by $\frac{N}{H} = 10$ specimens. Unlike conventional mark-recapture methods, which assume a single population with finite but constant size, our model further assumes that sampling is done from a single *infinitely large* panmitic population with constant size (i.e., as if all diversity for a species were represented within BOLD), where geographic and population structure are ignored. We recognize such assumptions may be biologically unrealistic, but are necessary here to maintain the simplicity of the model. The total number of intraspecific haplotypes was estimated using the function chaoHaplo() in SPIDER. The Chao1 estimate takes into account the total observed number of haplotypes as well as the number of singleton and doubleton sequences (those occurring only once and those appearing exactly twice) in a dataset given that a large number of individual specimens have been sampled [14]. The idea behind such an estimator lies in the expectation that the majority of unique haplotypes are rare (singletons), being represented

by only a single individual. Once all haplotypes have been observed at least twice (doubletons), it is considered unlikely that any new haplotypes will be found. Thus, observed samples with many singletons should be estimated to require larger sample sizes.

An estimate of the number of specimens that should be randomly sampled to recover all haplotypes for a given species was calculated by developing a simple equation

$$N^* = \frac{NH^*}{H} = \frac{N(H+1)}{2}$$

(derived below) where N and H are the observed number of specimens and haplotypes respectively in a given species sample and H^* is the Chao1 abundance estimator

$$H^* = \frac{H(H+1)}{2}.$$

Thus, given that N specimens have already been sampled, this leaves N^*–N individuals left to be sampled (and therefore H^*–H remaining haplotypes). Sampling sufficiencies (as a percentage of the observed number of specimens or haplotypes sampled or missing) were calculated for each of the 18 fish species as follows: $\frac{H}{H^*}\times 100\%$ (or equivalently, $\frac{N}{N^*}\times 100\%$) and $\left(1-\frac{H}{H^*}\right)\times 100\%$ (equivalently, $\left(1-\frac{N}{N^*}\right)\times 100\%$). These approaches give simple measures of 'closeness' between observed and estimated sample sizes. Ideally, N should be as close to N^* (and thus H as close to H^*) as possible (where $N^* - N$ and $H^* - H$ are minimized). This ensures also that $\frac{H}{H^*}$ (and therefore $\frac{N}{N^*}$) is maximized and $1-\frac{H}{H^*}$ (and thus $1-\frac{N}{N^*}$) is minimized.

Suppose N specimens are randomly sampled without replacement from a particular species and H haplotypes are observed. The number of haplotypes (H^*) for a species can be approximated using the Chao1 abundance estimator. The number of specimens (N^*) required to recover H^* haplotypes can then be easily found. The derivation of our model along with sample calculations follows. If we assume that species haplotypes occur at equal (uniform) frequency, then:

$$\frac{H}{N} = \frac{H^*}{N^*} \tag{1}$$

and after some algebra,

$$N^* = \frac{NH^*}{H} \tag{2}$$

The Chao1 abundance estimator H^* is:

$$H^* = \frac{H(H+1)}{2} \tag{3}$$

N^* can be simplified by substituting (3) into (2):

$$N^* = \frac{N(H+1)}{2} \tag{4}$$

We illustrate calculations of H^* and N^* for the Siamese fighting fish (*Betta splendens*) with $H = 4$ and $N = 76$:

$$H^* = \frac{4(4+1)}{2} = 10$$

$$N^* = \frac{76(10)}{4} = \frac{76(4+1)}{2} = 190$$

Given the sample size and haplotype number observed for *Betta splendens*, this method estimates a total of 190. Specimens would need to be randomly sampled from this species to recover all 10 estimated haplotypes.

3 Results

Our analyses suggest that the haplotype diversity for all 18 species examined here remains largely unsampled. Table 1 summarizes our findings for all species, including observed sample numbers and estimated total specimen/haplotype counts and sampling coverage. Haplotype accumulation curves and their corresponding slope values are also shown along with haplotype frequency barplots for several species showing patterns representative of the 18 species dataset (Figure 1). This information is also available for all 18 species as supplemental material. All slope estimates were found to be statistically significant ($p \approx 0$).

Haplotype accumulation curves failed to reach an asymptote for all 18 species; however, three of 18 species appear to approach an asymptote, i.e.: Chinook salmon (*Oncorhynchus tshawytscha*), Siamese fighting fish (*Betta splendens*) and Rockfish (*Sebastes* sp.) (Figure 1). Haplotype variation across all 18 species varies widely (Table 2). For example, relatively wide-ranging haplotype numbers (8-18) were observed for salmonids (Table 1, Figure 1), whereas darters show consistently high haplotype numbers (19-32) (Table 1). Among all species, the extreme cases are the high H^* estimate for the Orangebelly darter (*Etheostoma radiosum*) ($H = 32$, $H^* = 528$, % sampled = 6, % missing = 94) and the low H^* estimate for the Rockfish (*Sebastes* sp.) ($H = 2$, $H^* = 3$, % sampled = 67, % missing = 33). The haplotype accumulation curve for Chinook salmon appears to be approaching saturation despite a large number of haplotypes still unaccounted for ($H = 12$, $H^* = 78$).

4 Discussion

Here, we briefly explored a method to measure barcode haplotype sampling sufficiency based on actual sample sizes and observed intraspecific haplotype diversity as found among densely sampled actinopterygian fishes

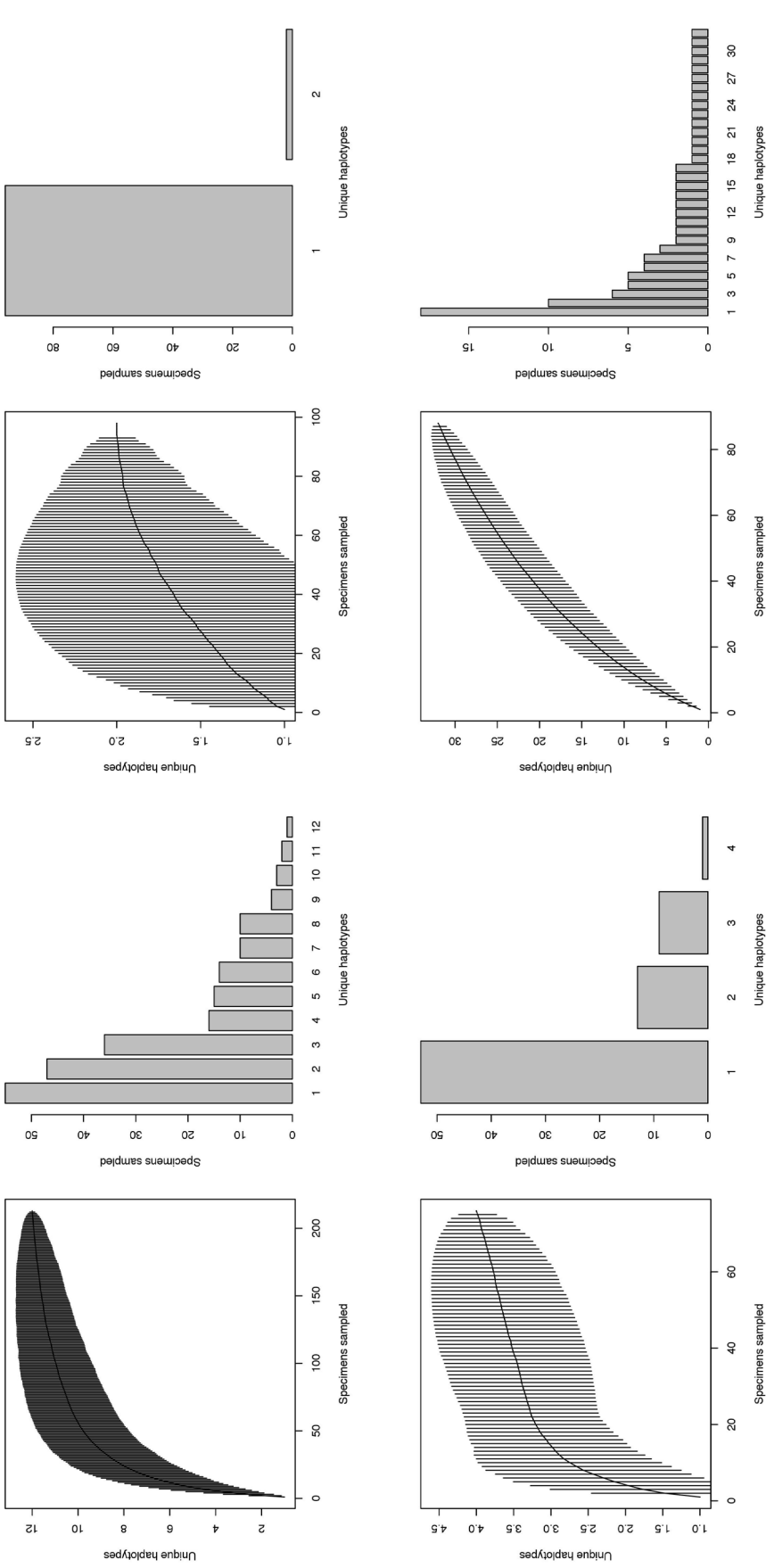

Figure 1. Haplotype accumulation curves and frequency histograms for four species: Chinook salmon (*Oncorhynchus tshawytscha*; top-left), Rockfish (*Sebastes* sp.; top-right), Siamese fighting fish (*Betta splendens*; bottom-left) and Orangebelly darter (*Etheostoma radiosum*; bottom-right) selected to show a range of sample sizes and haplotype diversity. Calculated slope estimates for the above-listed species based on the last ten points on the curve are respectively 0.006, 0.001, 0.013 and 0.180 and are intended to illustrate varying levels of sampling sufficiency observed for these species.

catalogued within BOLD. This was achieved using a simple mathematical model that is similar in practice to mark-recapture methods. Our results (available as supplemental material in the form of R-scripted code, sequence files and all accompanying figures/tables) suggest that the barcode sample sizes available for this study appear insufficient to predict haplotype diversity within species. The results show a wide range of sampling sufficiency across the 18 species; it appears much of the haplotype diversity within most of those species remains unsampled, including those species with relatively large sample sizes (e.g., ≥ 200). These findings could be due to at least two issues: (1) that a small number of points (10) were used in the calculation of curve slopes to assess sampling sufficiency; and/or (2) that the true number of species haplotypes is being overestimated. These issues seem to be most apparent for *O. tshawytscha*, where the discrepancies of premature curve saturation and missing haplotypes noted previously between calculated model estimates and the corresponding haplotype accumulation curve for this species were found (Table 1, Figure 1). Clearly, for issue (1) above, the use of an appropriate number of points is necessary, as using too few samples can lead to biases in haplotype diversity (over or under) estimation [4]. Consistent results require the use of comparable data across species. The relatively small sample size available for many species as well as the computational costs for this exploratory study were the primary driving factors behind choosing ten points. Alternately, the use of a fixed proportion, rather than a fixed number of points, may be a viable future option, as is a case where proportions are allowed to vary between species. The use of a fractional range of points falling on the last 20-15% and 15-10% as well as the last 10% of the curves in the calculation of slope estimates to observe a change in statistical significance of slope values is one possible solution to avoid potential bias. Such a statistical test has the advantage of localizing the point of saturation; whereas, single tests merely show that saturation of species haplotype accumulation curves is evident.

Issue (2), an inflated estimate of haplotypes, may be the result of the assumption of equal species haplotype frequencies in our model. An example of this may be our haplotype estimate for the Orangebelly darter (*Etheostoma radiosum*). *Etheostoma* is known to have high haplotype diversity [26]; however, we think our estimated total of 528 haplotypes for *E. radiosum* seems unrealistic. As a null method for this exploratory study, the assumption of equal haplotype frequencies has the advantage of greatly simplifying calculations. For instance, the equations for sampling sufficiencies outlined earlier can be expressed in terms of the number of specimens (N and N^*) or the number of haplotypes (H and H^*). Both methods give the same calculated value. Such a feature would not be apparent for an assumption of unequal haplotype frequencies as identifying the distribution of haplotypes would be difficult and would likely be species-specific.

We recognize that estimates of N^* calculated from our model likely represent underestimates of the true number of individuals of a given species which should be sampled. Many more specimens should therefore be sampled in order to ensure a sufficient number of haplotypes have been recovered. Equal haplotype frequencies are rarely observed in natural populations, and we suggest the development of more sophisticated models should explore the use of data simulations to evolve models that explicitly account for variation in species haplotype frequencies.

Conflict of interest: Authors declare nothing to disclose.

References

[1] Lenth R.V., Some practical guidelines for effective sample size determination, Am. Stat., 2001, 55, 187-193
[2] Lindblom L., Sample size and haplotype richness in population samples of the lichen-forming ascomycete *Xanthoria parietina*, The Lichenologist, 2009, 41, 529-535
[3] Nei M., Molecular Evolutionary Genetics, Columbia University Press, New York, 1987
[4] Goodall-Copestake W.P., Tarling G.A., Murphy E.J., On the comparison of population-level estimates of haplotype and nucleotide diversity: a case study using the gene *cox1* in animals, Heredity, 2012, 109, 50-56
[5] Hebert P.D.N., Cywinska A., Ball S.A., deWaard J.R., Biological identifications through DNA barcodes, Phil. Trans. Soc. Lond. B., 2003, 270, 313-321
[6] Zhang A.B., He L.J., Crozier R.H., Muster C., Zhu, C.-D., Estimating sample sizes for DNA barcoding, Mol. Phylogenet. Evol., 2010, 54, 1035-1039
[7] Ratnasingham S., Hebert P.D.N., BOLD: The Barcode of Life Data System (http://www.barcodingoflife.org), Mol. Ecol. Notes, 2007, 7, 355-364
[8] Muirhead J.R., Gray D.K., Kelly D.W., Ellis S.M., Heath D.D., MacIssac H.J., Identifying the source of species invasions: sampling intensity vs. genetic diversity, Mol. Ecol., 2008, 17, 1020-1035
[9] Pearson K., Method of moments and method of maximum likelihood, Biometrika, 1936, 28, 34-59.
[10] Gotelli N.J., Colwell R.K. Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness, Ecol. Lett., 2001, 4, 379-391
[11] Matz M.V., Nielsen R., A likelihood ratio test for species membership based on DNA sequence data, Phil. Trans. R. Soc. B., 2005, 360: 1969-1974
[12] Coeur d'acier A., Cruaud A., Artige E., Genson G., Clamens A-L., Pierre E., *et al.*, DNA barcoding and the associated

PhylAphidB@se website for the identification of European aphids (Insecta: Hemiptera: Aphididae), PLOS ONE, 2014, 9(6)
[13] Grewe P.M., Krueger C.C., Aquadro C.F., Bermingham E., Kincaid H.L., May B., Mitochondrial DNA variation among Lake Trout (*Salvelinus namaycush*) strains stocked into Lake Ontario, Can. J. Fish Aquat. Sci., 1993, 50, 2397-2403
[14] Brown S. D. J., Collins R. A., Boyer S., Lefort M.-C., Malumbres-Olarte J., Vink C. J. *et al.*, SPIDER: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding, Mol. Ecol. Resour., 2012, 12, 562-565
[15] Paradis E., Claude J., Strimmer K., APE: analyses of phylogenetics and evolution in R language, Bioinformatics, 2004, 20, 289-290
[16] Tamura K., Stecher G., Peterson D., Filipski A., Kumar S., MEGA6: Molecular Evolutionary Genetics Analysis version 6.0, Mol. Biol. Evol., 2013, 30, 2725-2729
[17] Paradis E., pegas: an R package for population genetics with an integrated-modular approach, Bioinformatics, 2010, 26, 419-420
[18] Hanner R., Data standards for BARCODE records in INSDC (BRIs), 2009
[19] Edgar R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res., 2004, 32: 1792-1797
[20] Ratnasingham S., Hebert P.D.N., A DNA-based registry for all animal species: The Barcode Index Number (BIN) system, PLOS ONE, 2013, 8
[21] R Core Team, *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013
[22] Hortal J., Lobo J.M., An ED-based protocol for optimal sampling of biodiversity, Biodiversity and Conservation, 2005, 14, 2913-2947
[23] Chao A., Nonparametric estimation of the number of classes in a population, Scand. J. Statist., 1984, 11, 265-270
[24] Chao A., Estimating the population size for capture-recapture data with unequal catchability, Biometrics, 1987, 43, 783-791
[25] Chao A., Estimating population size for sparse data in capture-recapture experiments, Biometrics, 1989, 45, 427-438
[26] Haponski A.E., Bollin T.L., Jedlicka M.A., Stepien C.A., Landscape genetic patterns of the rainbow darter *Etheostoma caeruleum*: a catchment analysis of mitochondrial DNA sequences and nuclear microsatellites, J. Fish Biol., 2009, 75, 2244-2268

Supplemental Material: The online version of this article (DOI: 10.1515/dna-2015-0008) offers supplementary material.