A Novel Statistical Framework for Assessment of Intraspecific Haplotype Sampling Completeness

by

Jarrett D. Phillips

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Computational Sciences

Guelph, Ontario, Canada

©Jarrett D. Phillips, April, 2021

ABSTRACT

A NOVEL STATISTICAL FRAMEWORK FOR ASSESSMENT INTRASPECIFIC HAPLOTYPE SAMPLING COMPLETENESS

Jarrett Daniel Phillips
University of Guelph, 2019

Co-Advisors:
Dr. Daniel Gillis and Dr. Robert Hanner

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge the support and guidance of my coadvisors, Drs. Dan Gillis and Bob Hanner for their encouragement over the years. In addition, thanks go out to all current and past members of the Gillis and Hanner Lab groups for all the fun times that were had throughout the journey.

Secondly, I wish to thank Dr. Deb Stacey and Dr. Graham Taylor for serving on my advisory committee, graciously reading my individual manuscripts and the current thesis, as well as engaging in stimulating discussion during my many committee meetings and graduate seminars over the past four and a half or so years.

This work was supported in part by a Graduate Excellence Entrance (GEE) Scholarship. Other sources of funding came from various travel grants which greatly aided the ability to travel and present my research at the 7th and 8th International Barcode of Life (iBOL) Conferences held in Kruger National Park, South Africa and Trondheim, Norway in 2017 and 2019 respectively.

Lastly, I would like to thank my family for their love and encouragement throughout my journey into the vast unknown that is doctoral study. Now they can finally stop constantly asking: "When are you going to get your PhD.?!"

Table of Contents

References 143

List of Tables

List of Figures

Chapter 1

Thesis Overview

1.0.1 Thesis Outline

This thesis outlines a novel statistical framework for assessment of COI DNA barcode haplotype sampling completeness. Chapter 2 consists of a literature review of studies conducted to date pertaining to sample size estimation for DNA barcoding, as well as the need for further research in this area. Chapter 2 details a stochastic optimization algorithm caled `HACSim` that is programmed in the R Statistical Environment. `HACSim` can be employed to calculate likely required specimen sample sizes needed to capture a given fraction of genetic diversity within species. Chapter 3 is devoted to a small simulation study assessing both the validity and statistical performance of `HACSim`, as well as its overall utility for assessing sampling completeness within DNA barcoding studies. Finally, in Chapter 4, it is argued that DNA barcoding is currently lacking in
statistical rigor and that better statistical methods are necessary to more accurately assess standing genetic variation at the species level.

1.0.2 Thesis Objectives

1.0.3 Thesis Statement

Through the development of a novel stochastic simulation algorithm for the generation of haplotype

accumulation curves, the current research will provide a framework that can be employed to determine plausible specimen sample sizes sufficient to quantify levels of haplotypic sampling completeness within species under both uniform and non-uniform haplotype frequency distributions.

Chapter 2

Incomplete estimates of genetic diversity within species: Implications for DNA barcoding

Jarrett D. Phillips[1,2], Daniel J. Gillis[1] and Robert H. Hanner[2,3]

[1]School of Computer Science

[2]Centre for Biodiversity Genomics, Biodiversity Institute of Ontario

[3]Department of Integrative Biology

Published in *Ecology and Evolution*

ABSTRACT

DNA barcoding has greatly accelerated the pace of specimen identification to the species level, as well as species delineation. Whereas the application of DNA barcoding to the matching of unknown specimens to known species is straightforward, its use for species delimitation is more controversial, as species discovery hinges critically on present levels of haplotype diversity, as well as patterning of standing genetic variation that exists within and between species. Typical sample sizes for molecular biodiversity assessment using DNA barcodes range from 5-10 individuals per species. However, required levels that are necessary to fully gauge haplotype variation at the species level are presumed to be strongly taxon-specific. Importantly, little attention has been paid to determining appropriate specimen sample sizes that are necessary to reveal the majority of intraspecific haplotype variation within any one species.

In this paper, we present a brief outline of the current literature and methods on intraspecific sample size estimation for the assessment of COI DNA barcode haplotype sampling completeness. The importance of adequate sample sizes for studies of molecular biodiversity is stressed, with application to a variety of metazoan taxa, through reviewing foundational statistical and population genetic models, with specific application to ray-finned fishes (Chordata: Actinopterygii). Finally, promising avenues for further research in this area are highlighted.

2.1 Introduction

One of the most fundamental problems underpinning much of modern molecular biodiversity research is the issue of determining optimal levels of sampling effort that are required in order to adequately characterize biological sequence variation at the species level. Molecular genetic studies of biodiversity that utilize mitochondrial DNA (mtDNA) marker variation for the purpose of characterizing existing species genetic diversity are particularly sensitive to sample sizes. Four fundamental evolutionary forces act to alter the genetic composition of species populations: migration/gene flow, mutation, natural selection and random genetic drift. The effect of genetic drift on species populations is most evident when population sizes are small, as in the case of a recent bottleneck or founder event, resulting in the rapid loss of genetic diversity. Species differ in both their evolutionary histories and in their geographic distributions; therefore, the question of accurately determining how many samples to include in order to observe a wide range of species genetic variation has been an ongoing area of interest and research. This is an important question deserving of more attention. Accurate determination of within-species (intraspecific) sample sizes for mtDNA diversity estimation permits detailed analyses to be undertaken at the phylogenetic and phylogeographic levels in order to infer key biological processes such as isolation, dispersal and speciation [10, 41, 58]. Aside from addressing purely biological questions, the issue of determining optimal sampling strategies and sample sizes for genetic variation assessment at the species level also manifests at applied socioeconomic scales, particularly in the detection of food or natural health product fraud

and in the monitoring of aquatic and terrestrial ecosystems [89].

Within the field of biodiversity science, researchers have long recognized the importance of sampling design in order to achieve a study's objectives. According to Lindblom [105], well-developed sampling designs within the field of molecular biodiversity science should be formulated around three basic areas: research study questions, research study aims and taxonomic focus. In addition to these three areas, Costa *et al.* [36] point to further considerations: planning the number and geographical distribution of specimens to be sampled, the category and number of genetic loci to be examined, and the spatial distribution and number of individuals to be sampled within each species' population. While there is a lack of clear sampling guidelines currently in place for optimal spatio-temporal assessment of species populations, Pante *et al.* [133] argues that such schemes should be guided by adequate coverage of both the putative geographic/ecologic range of the species under study, as well as potentially closely related species over its entire range. Given that much of species spatio-temporal metadata is not reported alongside genetic data, such assessments become problematic unless community standards and practices are improved [68, 125, 175]. Where this becomes particularly important is in the development and design of species-specific real-time Polymerase Chain Reaction (qPCR) primers and probes, for integration within environmental DNA (eDNA) assays for instance. This is especially the case if such tools are to be continuously implemented within regulatory or forensic settings such as the Canadian Food Inspection Agency (CFIA) [162] and the United States Food and Drug Administration (USFDA), as the success of such

methods depends greatly on the extent of geographic coverage of species genetic diversity.

The overall goal of sampling is to make inferences concerning a population of interest based only on information contained within finite samples drawn from the larger population. This is done though estimating population parameters such as the population mean (μ) using the sample mean ($\bar{x}$). One example, relevant to molecular population genetics, is the calculation of average pairwise distances based on Nei's estimator of nucleotide diversity (π) [128]. Under the Frequentist statistical paradigm, the minimum sample size that is required to estimate a population mean, from a Normal distribution, is given by [3]

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{d} \right)^2 \tag{2.1}$$

where $z_{\alpha/2}$ is the appropriate critical value to estimate μ with a level of significance of $1-\alpha$, σ^2 is the population variance and d is the desired margin of error. From the above equation, the required minimum sample size is controlled by the experimenter through the margin of error. A smaller margin of error results in a larger value of n. Similarly, predicting n with a higher level of accuracy can be achieved through narrowing d. Sample sizes that are computed from the above equation serve as a baseline requirement prior to conducting any quantitative study of interest. Depending on the sampling scheme, for instance stratified sampling, other formulas exist for the appropriate calculation of necessary sample sizes.

In determining the most appropriate sample size required for a particular study, a crude

rule of thumb that is often used in statistics and other scientific disciplines pertains to the use of a sample size of at least $n = 30$ when making comparisons among study groups or when deciding to use probabilities derived from the Standard Normal distribution [33]. Unfortunately, adequate sample sizes, while widely viewed as being central to a given biodiversity research study, are often neglected in practice [104]. In such cases, this may be due to, for example, costs associated with or resources required for adequate specimen collection [22, 84, 121].

Statistical power analysis can be employed to help shed light on sample sizes required in order to detect a given effect prior to carrying out a scientific study. Power, which is defined as the complement of the Type II error rate (β), depends on four factors: effect size (ES), significance level/Type I error rate (α), sample size (n) and population standard deviation (σ) through the proportionality [40]

$$(1 - \beta) \propto \frac{\text{ES} \times \alpha \times \sqrt{n}}{\sigma}. \tag{2.2}$$

Effect size is the difference between an observed quantity and one hypothesized under a null distribution. Larger deviations lead to greater power to detect real effects. It is easily seen from the above proportionality that larger values of effect size, significance level and sample size all generate higher levels of statistical power; whereas, increasing population standard deviation results in loss of power. Together with the sample size equation discussed previously (Equation 1), many factors are at play in determining the most appropriate sample size needed for a given study.

Any sampling scheme that is carried out will be subject to systematic error. Sampling (ascertainment) bias is an important factor to consider in this regard because it can lead to under- or overestimation of population parameters. Ascertainment bias describes the tendency of certain individuals to be less likely sampled than others [137] and is common in molecular biodiversity studies (*e.g.*, [69, 121, 123, 190]). This can occur, for example, when sampling is restricted to certain geographic regions [121] or to particular species (*e.g.*, those known to be of conservation importance) [69]. Sampling bias can be minimized through increasing the geographic breadth of a study, in addition to targeting representative taxa with large specimen sample sizes.

The present review briefly examines current approaches for species genetic variation assessment as it relates to the estimation of intraspecific sample sizes for DNA barcoding. Specifically, the focus will be on COI DNA barcode haplotype sampling completeness. Few studies have focused on DNA barcode sample size prediction for wide-ranging taxa in this regard. Here, methods of haplotype variation assessment are first covered. This is then followed by an examination of existing studies, with particular consideration of important findings to date within the literature. Finally, promising new avenues for further research are explored.

2.2 Current Methods

2.2.1 Methods to Assess Haplotype Variation

Haplotype Diversity

Genetic diversity is manifested within species in several ways. One way is through haplotype variation. While there are many different definitions of what constitutes a haplotype, in the broadest sense, a haplotype is a unique DNA sequence that differs from others at one or more basepair positions within and between species. Nei's [127] haplotype diversity (h), which is a widely-used approach to measuring genetic variation within species populations, is given by the equation

$$h = \frac{n}{n-1}\left(1 - \sum_i p_i^2\right). \tag{2.3}$$

where p_i is the frequency of the i^{th} haplotype in the sample. Two interpretations of *h* are that it expresses the probability of observing a previously unseen haplotype upon sampling a new individual [185] or that it represents the probability that two haplotypes, selected at random from a sample of n DNA sequences, are distinct [60]. Haplotype diversity can also be quantified using the absolute number of haplotypes (H). Both h and H are greatly affected by levels of sampling intensity within species. In particular, undersampling can cause these measures to become under- or overestimated [60]. Several other approaches are in wide use to aid researchers in assessing levels of standing genetic variation existing within species populations. Two of these are haplotype networks and haplotype

accumulation curves.

Haplotype Networks

A widely-used approach to assessing levels of genetic variation within and between species is through the construction of haplotype networks [179]. Haplotype networks accurately represent differences existing among sampled haplotypes through grouping identical DNA sequences within the same vertex. The size of a given vertex is proportional to the number of DNA sequences it contains. Divergent haplotypes are connected via edges that display the number of mutational differences separating adjacent vertices.

Haplotype networks are appealing because they can be used to infer potential cryptic diversity within a taxon or interspecific hybridization between allopatric (*i.e.*, reproductively-isolated) species, but interpretation can sometimes become difficult when multiple species cluster together into one or multiple nodes or subnetworks [70, 73, 193] or when ambiguous/missing nucleotide data are present within DNA sequences (*e.g.*, Ns or gaps (–)) [92]. While haplotype networks, such as the one shown in **Figure 2.1**, cannot give a direct indication of the level of sampling completeness for a given species, the presence of numerous rare haplotypes suggests gross undersampling of intraspecific genetic variation (or alternatively PCR/sequencing error).

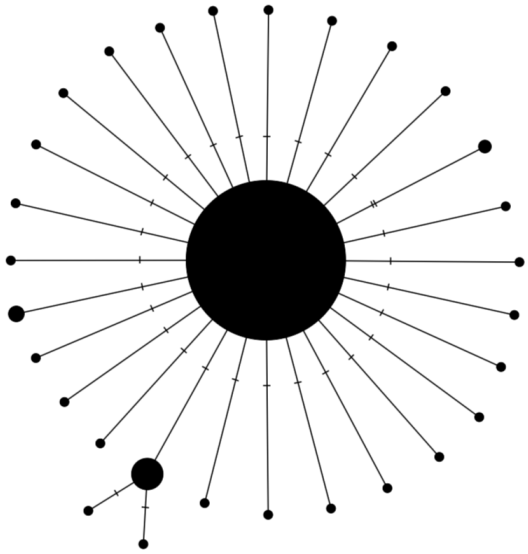

Figure 2.1: Longfin damselfish (*Stegastes diencaeus*) TCS [179] haplotype network depicting an overall skewed distribution of observed haplotypes. Sizes of circles reflect the number of DNA sequences contained within each vertex. Tick marks indicate the number of mutational differences separating sampled haplotypes. DNA barcode sequence data used in the generation of the network were taken from supplemental material accompanying Phillips *et al.* [144]. The software PopArt [103] was used to create the haplotype network.

Haplotype Accumulation Curves

Assessing the completeness of intraspecific haplotype sampling can be carried out through generating haplotype accumulation curves. Such curves are analogous to rarefaction curves used in studies of species richness [61] and depict the degree of asymptotic behaviour as a function of both the number of specimens sampled and the cumulative mean number of haplotypes accumulated. Initially, accumulation curves will increase very rapidly since many new haplotypes will be captured for a given species with minimal sampling effort, but haplotype recovery slows drastically as sampling depth is increased because many haplotypes that are found will have already been observed previously. Thus, species curves showing rapid saturation strongly suggest that the majority

of haplotype diversity has been uncovered; whereas, those curves displaying little to no evidence of reaching an asymptote indicate that further sampling is required [199]. Deciding whether a species should be further sampled can be deduced from the magnitude of the slopes calculated using a fixed number of points occurring on the end of the curve (e.g., ten in the case of [144, 196]). Slopes near or below a predefined threshold, for example, 0.01 (*i.e.*, equivalent to observing one new haplotype for every 100 DNA sequences), suggest that additional sampling is unlikely to reveal any new haplotypes; whereas, those species curves with slopes above 0.1 (*i.e.*, observing one new haplotype for every 10 DNA sequences) strongly indicate that further sampling is necessary [84].

One obvious problem that arises in the use of haplotype accumulation curves to gauge species genetic diversity and levels of sampling effort, however, is the fact that the functional form of such curves is not known and can differ widely across taxa [144]. Furthermore, deciding on appropriate curve slope thresholds necessary for adequate sampling coverage are largely arbitrary [84]. While various parametric model curve-fitting approaches, such as the power, negative exponential and Michaelis-Menten functions, have been heavily employed and debated in the literature to model species-area relationships [38, 181] or species richness, no single approach yet exists that can be readily applied to determine sample sizes that are likely required for intraspecific genetic variation assessment.

A second, lesser-investigated issue, relates to the fact that haplotype accumulation curves are not spatially-explicit. Thus, it becomes difficult to account for correlations that

may exist at the subpopulation or higher taxonomic levels. This has been noted in past studies of species richness employing species accumulation and rarefaction curves (*e.g.*, [16, 30, 180]).

2.2.2 Sampling Models for Genetic Diversity Prediction

In addition to qualitative approaches to assessing standing genetic variation within species, a number of quantitative models to estimate required sample sizes for overall genetic diversity assessment have been proposed. These include Frequentist, Bayesian and coalescent models.

Holt *et al.* [83] reviewed several Frequentist and Bayesian statistical methods of sample size determination for intraspecific haplotype diversity assessment that are most informative over large geographic ranges. The authors note that a lower bound on the probability of sampling a dominant haplotype in a sample of size n with significance level α is given by the inequality

$$p \geq \sqrt[n]{\alpha} \tag{2.4}$$

Grewe *et al.* [62] employed an equivalent approach to Holt *et al.*'s [83] study through utilizing a binomial sampling model to determine the minimum sample size required to assess mtDNA variation in Lake Ontario lake trout (*Salvelinus namaycush*) stocks

according to the equation

$$n = \frac{\ln(1-\beta)}{\ln(1-p)} \tag{2.5}$$

where p is the frequency of a given haplotype and β is the desired confidence level. The authors found that n = 60 individuals are likely needed to be randomly sampled in order to observe a single haplotype having a frequency of at least p = 5% with β = 95% confidence. It is worth noting that this figure increases to $c.$ 460 individuals for a haplotype occurring at frequency of 1% with 99% confidence [62]. This marked increase in sample size is not surprising given that one would need to sample many more individuals in order to be certain that the majority of rare haplotypes have been uncovered. It is important to note, however, that Grewe *et al.* [62] sampled individuals from six different but highly divergent trout strains, each displaying high degrees of population substructure. Population subdivision likely will have an effect on the estimation of required sample sizes needed to gauge levels of standing genetic variation at the species level.

Similar magnitudes of sample sizes were found by Austerlitz *et al.* [9], who employed coalescent theory [96], in order to determine the probability of adequately sampling all genetic variation of a species with sample size n. Coalescent theory attempts to trace the lineage of an ancestral allele (termed the Most Recent Common Ancestor, MRCA) backwards in time within a gene genealogy. Under a geometric distribution, this probability

is given by the equation [9]

$$p = \frac{n-1}{n+1}. \tag{2.6}$$

From the above equation, only n = 39 individuals are required to be sampled at random in order to observe p = 95% of all genetic diversity for a species. It should be noted however that even with increasing sample sizes, one's confidence in having sampled all of a species genetic diversity approaches closely, but never actually reaches, 100% [9]. This is illustrated by the finding that the required sample size increases to n = 1999 individuals necessary to observe p = 99.9% of the total genetic diversity that exists for a given species using Equation (6). This can be explained by the fact that individual haplotypes for a given species become much more difficult to recover as the intensity of specimen sampling is increased because intraspecific genetic variation is expected to increase as a result. The coalescent, as a large-scale sampling model, has found wide application in DNA-based approaches to species identification and delimitation, most notably DNA barcoding [85].

2.2.3 DNA Barcoding

Since its conception in 2003, DNA barcoding [75] has risen to become the largest taxonomically-driven biodiversity initiative to date aimed at identifying and cataloging all assemblages of multicellular life on the planet. DNA barcoding is a genomic technique that relies on DNA sequence variation within short, standardized gene regions in order to rapidly identify specimens to the level of species and to discover new species. The

ideal DNA barcode is one that is found in all organisms, readily distinguishes between taxa, and is easily amplified, sequenced and aligned. In animals, the agreed-upon marker of choice for taxon assignment is a *c.* 650 basepair (bp) fragment from the 5' end of the mitochondrially-encoded cytochrome *c* oxidase subunit I (COI) gene. Mitochondrial loci like COI are particularly suitable as genetic markers for DNA barcoding because they are fast evolving, highly conserved across taxa, present in high copy number, haploid, maternally inherited, lack introns, display few insertion-deletion (indel) mutations, and experience little to no gene recombination [75, 77].

The primary goal of DNA barcoding has been to develop a publicly accessible species reference sequence library to aid in the identification of unknown specimens and accelerate the discovery of potentially undescribed taxa. Obtaining adequate sample sizes for building accurate and reliable specimen reference libraries has culminated in the development of the Barcode of Life Data Systems (BOLD; http://www.boldsystems.org) [150] as the largest collection of user-curated species sequence data specifically for DNA barcoding currently available on the World Wide Web. At present (as of May 1, 2018), BOLD holds over six million DNA barcode records from over 250,000 named species. Certain taxa are well represented in BOLD with upwards of hundreds of barcode sequences for some species. Despite this, barcode reference libraries within BOLD remain largely incomplete, even for the most well-sampled taxa such as fishes and insects. As such, comprehensive coverage of species genetic diversity is still decades away [190]. Wilkinson *et al.* [190] points to strong ascertainment bias as the most likely explanation for this. In the early days of BOLD, DNA

barcode sequence acquisition was high, due to the fact that over 75% of taxon records were mined from already well-established sequence databases such as GenBank [190].

2.2.4 The Importance of Sampling to DNA Barcoding

DNA barcoding works in practice because interspecific (between-species) variation is usually much greater than intraspecific (within-species) divergence [118, 174]. While this observed 'barcoding gap' [118] is a necessary criterion for successful taxonomic resolution using distance-based methods, it may not be a sufficient one for other molecular approaches (*e.g.*, those employing tree- or character-based techniques). Cases are well documented where considerable overlap/separation between (maximum) intraspecific variation and (minimum) interspecific divergence exists [79, 85]. Undersampling can greatly exaggerate the existence of the barcode gap. The inclusion of small sample sizes over large geographic ranges has the effect of obscuring existing mitochondrial sequence diversity at the species level since the finding of divergent haplotypes may be the result of poorly sampled panmictic (*i.e*, randomly-mating) intraspecific variation [31]. Compared to regional scales, with increasing sampling effort across wider spatial scales, intraspecific variation is expected to increase whereas interspecific divergence will decrease in effect since more closely-related species will tend to be found due to allopatric speciation being a dominant mode of diversification [14, 140].

How much variation is actually needed to separate species is not known with certainty because intraspecific sampling has generally been limited to narrow geographic locales. Hebert *et al.* [75] proposed that barcode sequences exhibiting at least 2% nucleotide

divergence should be designated as being from distinct species. Intraspecific distances larger than 2% suggest the presence of cryptic species, whereas those smaller than 2% is evidence for evolutionarily young species with a recent origin (*i.e.*, retention of ancestral polymorphisms due to incomplete lineage sorting), hybridization/introgression or inadequate taxonomy (*e.g.*, cryptic species or species synonymy) [85]. In BOLD, query sequences are matched to reference barcodes based on a genetic distance heuristic of 1% [150]. The use of such threshold estimates for species separation is arbitrary and is often applied to a wide variety of taxa, regardless of species life histories. A later estimate of ten times the mean intraspecific distance (the so-called '$10\times$ rule') was given by Hebert *et al.* [79]. Unlike the previously suggested estimate of 2% sequence divergence, the $10\times$ rule makes use of all available taxon sequences within a dataset in order to calculate an appropriate limit for species separation. Despite this, the $10\times$ rule has been met with criticism: Collins and Cruickshank [35] suggest consideration of the maximum intraspecific distance and the minimum interspecific divergence (*i.e.*, nearest neighbour distance) for each species under investigation. The use of lower thresholds for species discovery may falsely inflate existing genetic diversity, whereas the adoption of higher cutoffs would likely be too conservative for reliable detection of cryptic species [7]. It is well understood however that the most appropriate cutoff necessary to accurately diagnose species on the basis of sequence variation is strongly taxon-dependent [77, 80, 118] and will become more precise with increased sampling effort.

DNA barcoding has its roots in the historic disciplines of Darwinian evolutionary

theory, population genetics and phylogenetics: the coalescent is a modern interpretation that reconciles these domains [153]. While genetic distance-based approaches to species delimitation are commonplace within barcoding studies because they scale well to large taxon datasets, early-proposed arbitrary separation methods like the 2% or 10× rule completely ignore evolutionary relationships that exist among closely-related species. Objective tools for the delimitation of species are well known and generally fall into three overlapping categories: phylogenetic, coalescent, and phylogenetic-coalescent [85]. The well-known neighbour-joining clustering method was advocated for in the early barcoding literature as a means of confirming the presence of reciprocal monophyly across sampled taxa. More recently, novel bioinformatic algorithms, most notably distance-based approaches such as Automatic Barcode Gap Discovery (ABGD; [148]) and tree-based methods including variants of the Generalized Mixed Yule Coalescent (GMYC; [120, 146]) have been put forth in order to facilitate species separation, an otherwise daunting task for even the most highly-skilled and knowledgeable taxonomist. ABGD is a nonparametric technique of partitioning species on the basis of the barcode gap using DNA sequences. On the other hand, GMYC is a likelihood-based method that relies on the premise that bifurcation (*i.e*, fully-resolved branching) within ultrametric species trees is indicative of speciation/diversification events, and therefore suggests the presence of undescribed taxa. A key factor in the success of such methods is sample size, and few groups have been so extensively inventoried [85]. For example, GMYC is especially prone to the under- or overestimation of putative species, which can be magnified due to differences in effective

population sizes as well as historical versus contemporaneous patterns of migration/gene flow among subpopulations [108, 134]. Thus, sufficient sampling is paramount. Often, researchers would like to know whether all unique haplotypes within a lineage or deme have been adequately sampled; unfortunately, this is complicated by the fact that the majority of species are both geographically-widespread and rare. As a result, given that ascertainment and operational biases are inevitable [123], an extensive sampling of all local populations that comprise a given species is unrealistic, even under the best situations (*e.g.*, strong research budget, easy access to sampling locations). Thus, whenever possible, a more comprehensive sampling of study sites is required in order avoid false positives/negatives and to reveal divergent haplotypes that may have been missed with spatially-narrower sampling routines [120]. Incorporation of coalescent and population genetics theory can aid in informing researchers on broad macro-level processes that may be at play in shaping trends seen within haplotype accumulation curves on the basis of extant patterns of intraspecific genetic diversity.

The Barcode Index Number framework for animals, first introduced by Ratnasingham and Hebert [151], represents a potentially novel approach to addressing the issue of sample sizes necessary for barcoding initiatives. The BIN system partitions COI barcodes into distinct Operational Taxonomic Units (OTUs) on the basis of the REfined Single Linkage (RESL) clustering algorithm and Markov clustering [151]. BINs comprise high-quality sequences linked to BARCODE compliant records. The BARCODE standard currently in place stipulates that only barcode sequences with read lengths of at least 500 bp and

containing less than 1% ambiguous nucleotides are designated unique BIN clusters [68]. While BINs generally show high concordance with actual biological species, they can be further employed to gauge instances of suspected cryptic species diversity, especially in the cases where intraspecific distances are not clearcut. Species that fall into two separate BINs (termed a SPLIT) is evidence that they are being overlumped. Further, the occurrence of rare BINs (*i.e.* those represented by a single specimen) may be the result of limited sampling [74, 88]. Stand-alone BINs may also reflect sequencing errors in the form of very low-frequency (VLF) variants or cryptic pseudogenes [173, 174]. Increased sampling coverage can be beneficial in such instances, as true biological variation is less likely to be misidentified as artificial biological variation and unintentionally flagged as potential VLFs.

2.2.5 Consideration of Species' Life Histories

Life history traits, particularly those pertaining to reproductive strategies and sex determination, in well-studied metazoan taxa such as fishes, insects and herpetofauna, are presumed to play a significant role in observed patterns of mtDNA barcode sequence variation at the species level. For instance, the high occurrence of haplodiploidy, a mode of inheritance whereby females develop from fertilized eggs (hence are diploid), while males arise from unfertilized eggs (therefore are haploid), is common across many insect orders such as Hymenoptera, and may explain the large abundances and varying (effective) population sizes seen in representative species that ultimately drives speciation and hybridization [78]. Similar "exceptions to the rule", such as (asexual) modes of

parthenogenesis (*e.g.*, unfertilized eggs producing female-only offspring in Squamata such as species of whiptail lizards), or paternal/biparental organelle inheritance in bivalve molluscs (*e.g.*, mussels of the genus *Mytilus*), will likely help inform researchers on the required level of sampling depth needed to fully characterize broad ranges of COI haplotype diversity in taxa that do not otherwise conform to traditional mtDNA inheritance patterning (*i.e.*, strictly maternal lineage), and thus prevent the naïve implementation of recommendations of any one statistical approach employed in the calculation of intraspecific sample sizes for accurate specimen assignment and rapid species delineation. As an example, because parthenogenetic species display lower standing genetic diversity compared to fully sexually-reproducing species (as a result of being exact clones of their parent due to lack of chromosomal recombination) [13], haplotype frequencies aside, the observation of the faster approach of haplotype accumulation curves to an asymptote is expected. Thus, species exhibiting such mechanisms will require reduced levels of sampling effort. Such a result can be invoked through consideration of Muller's ratchet, as the irreparable accumulation of deleterious mutations that are fixed by genetic drift within asexual genomes directly limits the ability of a species to survive and reproduce [52, 122].

2.3 Key Findings

2.3.1 DNA Barcoding and Sample Size: Past Studies

The ability of DNA barcodes to uncover levels of standing genetic variation within species is strongly influenced by the scale of specimen sampling, which has been

recognized as a major barrier to the success of DNA barcoding since its early days [79, 118, 184]. In spite of this, global barcoding efforts have only been partially successful in capturing the full extent of COI barcode variation in animals due to the majority of studies forgoing deep taxon sampling in favour of maximizing the number of different taxa sampled [115, 199]. Sample sizes of a few individuals per species (typically in the range of 5-10, but one or two specimens is not uncommon since these are often the only representatives available, either due to unclear species boundaries or limited geographic sampling of intraspecific variation) are widespread in barcoding studies [66, 115, 199]. Recommended sample sizes currently in place are by no means sufficient since species abundance is often skewed geographically/ecologically. For example, five specimens per species per FAO (Food and Agriculture Organization) region was initially suggested by the Fish Barcode of Life (FISHBOL; [183]) initiative, but the sampling of up to 25 individuals or more may be necessary for some species exhibiting widespread distribution patterns [12, 170]. Similarly, in assessing haplotype and nucleotide COI variation across wide-ranging animal taxa, Goodall-Copestake *et al.* [60] note that a sample size of five individuals per species population was adequate to differentiate between extremes of h, but as many as 25 specimens would need to be collected in order to achieve maximum accuracy. Jin *et al.* [91], and Matz and Nielsen [115] both point to a sample size of 12 specimens, whereas Ross *et al.* [155] suggest that sampling five or more reference barcodes is sufficient for accurate species identification. Bias toward low sample sizes observed for most species may be the result of many factors (see Bucklin *et al.* [21] for a concise summary in marine

metazoa), including the presence of cryptic diversity, amplification of non-functional gene copies

(*i.e.*, pseudogenes/nuclear-mitochondrial inserts (NUMTs)), contamination by foreign DNA from other species (*e.g.*, bacterial symbionts such as *Wolbachia*), insertion-deletion (indel) mutations, or errors arising from PCR/sequencing runs [60]. Molecular diagnosis of specimens to the species level using DNA barcoding is not definitive; numerous technical sources of error exist that can hamper the ability of reliable taxon assignment, in particular, misidentifications, sequencing errors and lack of taxonomic metadata (*e.g.*, inclusion of GPS coordinates, record linkage to a voucher specimen). While such factors are likely to occur infrequently for interspecific barcodes, this is not the case for intraspecific datasets. Taken together, biases in sample sizes will likely be considerable. In certain cases, the occurrence of biological phenomena can lead to problems encountered later on in the lab, specifically during the sequence amplification stage using PCR. A well-known example of this is the symbiotic association of the bacterium *Wolbachia* with insects. Integration of *Wolbachia* within host genomes of various Hymenoptera, Diptera and Lepidoptera can cause fluctuations in intraspecific distances [163] and thus, observed haplotype diversity between infected and uninfected hosts [28]. Misamplification of host sequences for bacterial symbionts is widely encountered, as is the amplification of pseudogenes/NUMTs. Technical sources of error such as expert taxonomic misidentifications, sequence contamination, as well as errors arising from the amplification/sequencing process can be controlled, and can be minimized to a degree.

Two critical steps in avoiding such issues are: (1) the construction of an NJ tree in order to pinpoint potentially misidentified specimens and/or sequence contaminants (as opposed to solely being used in the establishment of reciprocal monophyly, as argued by Collins and Cruickshank [35]) and (2) the careful inspection of BOLD specimen trace files in order to resolve noisy sequence regions that inflate estimates of standing genetic variation through the introduction of functional (heteroplasmic) sequence variation (as in *e.g.*, Hebert *et al.*, [76]) and/or nonexistent low-frequency species haplotypes occurring in high abundance [173]. The effect of these on generated haplotype accumulation curves is delayed saturation to an asymptote due to larger required sample sizes. Combined with initially large numbers of specimens within intraspecific data sets (*e.g.*, $N > 100$), this effect can be quite substantial. As BOLD is ever-evolving, in part due to the sheer volume of DNA barcode sequences being added on a daily basis, it is crucial that suspected errors within taxon records be dealt with in a timely manner (*e.g,*, through community users flagging problematic records for closer examination by submitters), so that sequence integrity is not compromised. While the issue of determining adequate sample sizes for molecular species diagnosis has largely been aimed at animal taxa, Liu *et al.* [107] explored optimal sample sizes needed for plant DNA barcoding. It was found that relatively small sample sizes were adequate to recover sequence variation in slowly evolving genes (2 or 3 sequences per species population for matK); whereas, higher numbers are necessary for rapidly evolving markers (minimum of 10, 8 and 6 individuals per population for trnH-psbA, trnL-trnF and ITS respectively) [107]. Further, the authors found that a sample size of 8-10 individuals

per species across the entire geographic range appears sufficient for *Taxus* barcoding. Unfortunately, such small sample sizes, likely the result of low information content due to the high presence of sequence artifacts (*e.g.*, indels within mitochondrial/plastid markers), often lack discriminatory power that is needed for accurate identification of specimens on the basis of genetic polymorphism with DNA barcodes.

To date, few studies explicitly exploring simulated sample sizes for DNA barcoding in wide ranging animal taxa have been conducted. One of the first studies to examine the issue of sample sizes for DNA barcoding via haplotype accumulation curves was conducted by Zhang *et al.* [199] using a modified form of the Michaelis-Menten equation. Using this method, the authors found that the random sampling of 250-1188 individuals from the Costa Rican skipper butterfly (*Astraptes fulgerator*) cryptic species complex are likely needed in order to detect 95% of all genetic diversity for this species based on an initial sample size of 407 individuals. Conversely, the same authors found that 156-1985 specimens were needed to retrieve 95% of COI variation using simulated island [194] and stepping-stone [95] coalescent models across three distinct subpopulations and under varying effective population sizes. In addition, a sample size outlier of only 47 individuals was found for one subpopulation of *A. fulgerator* butterflies. The authors note that this may be due to the low level of genetic variation observed in this population: only two haplotypes were observed across 14 sampled individuals. In contrast, a later study on European diving beetles undertaken by Bergsten *et al.* [14] found that based on 419 sampled *Agabus bipustulatus* specimens, a sample size of 250 specimens was required

to be randomly sampled across its range to achieve 95% haplotype recovery. On the other hand, 70 individuals of the same species was necessary to be sampled in order to recover 95% of COI variation when geographic dispersion between a new sample and the closest previous sample was maximized using resampling simulation.

Not all studies find evidence for greatly broadening the scope of comprehensive specimen sampling. Luo *et al.* [111] demonstrate the utility of the Central Limit Theorem (CLT), employing a simple resampling scheme along with the modified Michaelis-Menten saturation model. The CLT states that the distribution of the sample mean tends toward the (Standard) Normal distribution as the sample size increases. It was found that a minimum sample size of only 20 individuals is needed to provide a reliable estimate of genetic polymorphism at the species level on the basis of observed haplotype numbers. The authors note however that sample sizes should be as large as possible, even though new haplotypes will tend to be observed with lower frequency. Compared to present sample size range of 5-10 specimens per species, a slightly larger minimum sample size range of 11-15 individuals per species was recommended by Yao *et al.* [195] for widely-distributed coastal and inland aquatic salt-tolerant plant species of the families Poaceae and Chenopodiaceae across seven different genera, based on results obtained through resampling procedures and nonparametric Mann-Whitney U tests.

Though not devoted to estimating sample sizes for mitochondrial genes such as COI, using resampling simulation, Hale *et al.* [67] found that a sample size of 25-30 individuals was sufficient to accurately estimate microsatellite allele frequencies in hypothetical

populations of hairy wood ants (*Formica lugubris*), kakis (*Himantopus novaezelandia*), black-browed albatrosses (*Thalassarche melanophris*) and red squirrels (*Sciurus vulgaris*). The sampling of 25-30 individuals per species for the assessment of genetic diversity via microsatellite loci was also recommended by Pruett and Winker [147] in an earlier study of song sparrows (*Melospiza melodia*). A more recent simulation study examining minimum sample sizes for accurate estimation of genetic diversity from a large number of single nucleotide polymorphism (SNP) markers in the terrestrial Amazonian plant *Amphirrhox longifolia* found that sample sizes beyond eight are sufficient for genetic diversity assessment and as few as two individuals are needed in order to obtain good estimates of population differentiation [126]. These studies clearly point to the need for large sample sizes in multilocus population genetic studies for the overall assessment of genetic diversity at the species level.

These examples serve to illustrate the fact that, as is the case for species divergence thresholds, there is no one universal sample size that can accurately recover the majority of intraspecific genetic variation across taxa and it appears likely that varying levels of additional sampling will be required within taxa and across geographic ranges [110]. What seems to be clear is the fact that many previous assessments of sample sizes necessary for DNA barcoding studies have underestimated levels of sampling depth that are actually needed in order to recover much of the genetic variation that exists at the species level. Such a trend seems most attributable to restricted geographic sampling and unclear species boundaries, limited funding for adequate specimen retrieval, as well as human-mediated

mechanisms such as errors accrued during the amplification/sequencing process.

2.4 Case Study: Phillips *et al.* (2015)

Phillips *et al.* [144] wished to estimate *sampling sufficiency* (θ) — the sample size at which accuracy is maximized and above which no additional sampling information is likely to be gained. This was applied in the context of haplotype accumulation curves in order to determine the point on the x-axis where curve saturation first becomes evident. If such an estimate exists, it would provide a useful stopping rule for specimen sampling [144]. That is, if a lower bound for specimen sample size exists, then it would provide the best estimate of sampling sufficiency for a given species.

2.4.1 Model Assumptions

In developing their sampling model, Phillips *et al.* [144] made several important assumptions, which together form a baseline "perfect-world" scenario for further exploration of specimen/haplotype sampling. These are:

- that specimen sampling is carried out randomly and without replacement from an infinitely large, panmictic population with constant size
- that species haplotypes are both biologically real and unique; and
- that species haplotypes occur with equal frequency.

In the first assumption, the contribution of genetic drift is presumed to be negligible and it is assumed that population structure is absent. Luo *et al.* [111] presumed a constant

population size, as well as an absence of natural selection, when calculating intraspecific sample sizes for their simulation study. The argument was that a limited number of individuals would be available in species populations undergoing contraction and that coalescence may not be evident. With regard to the second assumption, DNA barcodes are presumed to be of sufficiently high quality such that they are free of both ambiguous and missing nucleotide bases, which can lead to overestimation of observed and total haplotype numbers through creating artificial haplotype variation within species [8, 37, 144, 173, 174].

Assumptions 1 and 3 were employed by Dixon [41] in proposing a method to assess the extent of haplotype sampling completeness utilizing a Bayesian statistical framework based on the use of Stirling numbers. It was noted that the probability of all haplotypes being observed for a species becomes less accurate if the assumptions of random sampling and equal haplotype frequencies are not met and that the presence of rare species haplotypes will lead to overestimation of overall sampling completeness. Similarly, Phillips *et al.* [144] hypothesized that the presence of rare haplotypes within species will lead to inflation of total sample sizes. Further, as noted by Dixon [41], evolutionary mechanisms such as isolation-by-distance, which describes the variation in genetic composition of species populations with increasing geographic distance, will likely cause the true extent of sampling effort to be overestimated. In exploring coalescent simulations, Luo *et al.* [111] treated barcode sequences as panmictic. In this way, all specimens can be regarded as being sampled from a single geographic region. Such an assumption is not uncommon within

DNA barcoding studies, which are often geographically-focused [35]. While Luo *et al*. [111] did not consider spatial heterogeneity within their simulation study, it was proposed that stratified sampling, where individuals are repetitively sampled without replacement from a pre-selected number of strata, can be employed, with the added assumption that gene flow can largely be ignored.

2.4.2 Mathematical Details

Phillips *et al.* [144] derived a simple Method of Moments [139] estimator in order to predict adequate specimen sample sizes necessary to uncover the majority of cytochrome *c* oxidase subunit I (COI) DNA barcode haplotype diversity existing within animal species according to the equation

$$N^* = \left\lceil \frac{NH^*}{H} \right\rceil. \tag{2.7}$$

Above, N^* is considered an estimate of θ, the true sampling sufficiency, which, under the Frequentist statistical paradigm, is a fixed but unknown parameter. The quantity $\left\lceil \frac{N}{H} \right\rceil$ is the number of specimens represented by each haplotype ($\lceil x \rceil$ is the ceiling function applied to a number x, evaluated by rounding up to the nearest integer). Since haplotypes are assumed to be sampled with equal frequency from a species population, in a sample of $N = 100$ sequences comprising $H = 10$ distinct haplotypes, it is expected that each haplotype

is represented by 10 specimens [144]. H^* is found using the equation

$$H^* = \sum_{i=1}^{H} i = \frac{H(H+1)}{2} \tag{2.8}$$

where N is the number of DNA sequences observed for a given species, H is the number of observed haplotypes and H^* is the estimated total number of haplotypes (both observed and unobserved) for a species. The above estimator is similar to estimators of total species richness used widely in ecological settings (*e.g*, the Chao1 estimator of abundance [27]). The central idea around the above estimator is that the majority of haplotypes within a species are rare, being represented by only one (singleton) individual. Thus, once such haplotypes have been accounted for in a species sample, few additional unduplicated haplotypes are likely to be observed, since the majority of remaining haplotypes will be dominant (duplicates) in the population (*i.e.* being represented by two or more specimens); thus, species comprising many singleton haplotypes should be expected to require larger sample sizes in order to capture most of the existing genetic variation for a given species of interest [144, 192].

Phillips *et al.* [144] also proposed both absolute and relative "measures of sampling closeness" in order to quantify the extent of specimen and haplotype sampling effort. These quantities are as follows:

- Mean number of haplotypes sampled: H
- Mean number of haplotypes not sampled: $H^* - H$

- Proportion of haplotypes sampled: $\frac{H}{H^*}$
- Proportion of haplotypes not sampled: $\frac{H^* - H}{H^*}$
- Mean number of individuals not sampled: $N^* - N$

The above equations, which are central to Phillips *et al.*'s [144] sampling model, can be depicted graphically as follows (**Figure 2.2**).

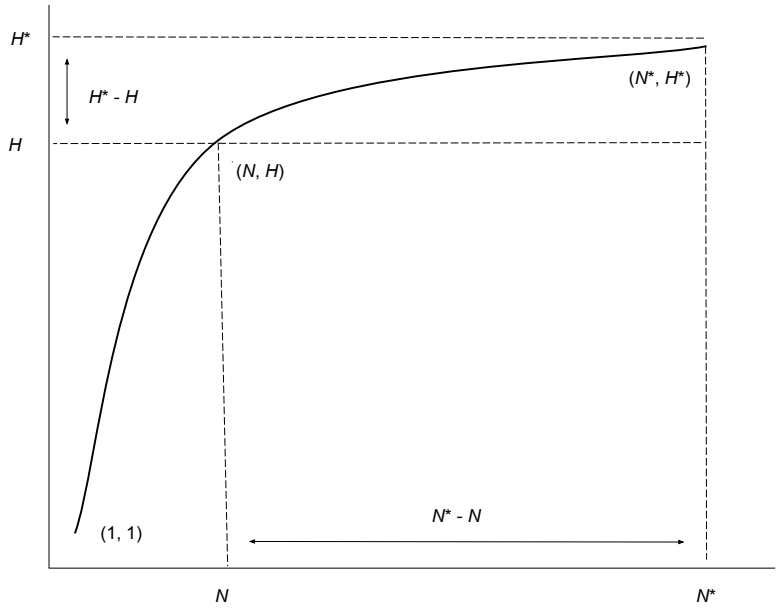

Figure 2.2: Graphical depiction of Phillips *et al.*'s [144] sampling model as described in detail within the main text. The *x*-axis is meant to depict the number of specimens sampled, whereas the *y*-axis is meant to convey the cumulative number of unique haplotypes uncovered for every additional individual that is randomly sampled. N and H refer to specimen and haplotype numbers that are observed for a given species. N^* is the total sample size that is needed to capture all H^* haplotypes that exist for a species.

Figure 2 resembles the general shape of a saturated haplotype accumulation curve for a hypothetically well-sampled species. The point labelled (N, H) on the curve reflects the

current level of sampling effort that has been expended for a given species (*i.e.*, as found in BOLD). The goal is to extrapolate the curve to the point (N^*, H^*) in order to observe the value on the *x*-axis (*i.e., N**) at which levelling off toward an asymptote (on the *y*-axis) first becomes evident (*i.e.*, at the value of *H**). Here, $N^* - N$ is the number of additional specimens that must be randomly sampled in order to observe $H^* - H$ additional haplotypes for a given species. If H is equal to *H**, then *N** will be equal to N, and no further sampling is necessary; otherwise, if H is less than *H**, then *N** will be greater than N, and additional sampling will be required. The curve in **Figure 2** passes through the point (1, 1), which is due to the fact that the sampling of a single individual of a given species corresponds to observing one unique haplotype for that species.

2.4.3 Application to Ray-finned Fishes

Phillips *et al.* [144] investigated levels of existing COI haplotype variation in 18 species of ray-finned fishes (Chordata: Actinopterygii) represented by a minimum of 60 individuals in accordance with Grewe *et al.* [62]. Results showed that 147-5379 specimens likely must be randomly sampled to uncover all predicted haplotype diversity in the selected species (between 3-528 total haplotypes) [144]. Sample size estimates obtained by Phillips *et al.* [144] are comparable in magnitude to those of Zhang *et al.* [199], but not in the case of Luo *et al.* [111], which are closer to practical sample sizes for DNA barcoding. Further, haplotype accumulation curves displayed evidence of reaching an asymptote for only 3/18 examined species: Chinook salmon (*Oncorhynchus tshawytscha*), Rockfish (*Sebastes* sp.) and Siamese fighting fish (*Betta splendens*) based on significance testing of curve slopes

with a one-sided *t*-test using the last 10 points on the end of accumulation curves [144]. Of note is the haplotype accumulation curve for Chinook salmon, which appeared to show premature saturation despite only 12 out of an estimated total of 78 haplotypes being found for the species. At the time of publication of Phillips *et al.*'s [144] study, *Sebastes* sp. was linked to a single BIN. The BIN system is inherently dynamic: as more sequences are added within BOLD, specimens assigned to a single BIN may be allocated to multiple BINs or multiple existing BINs may be coalesced into a single BIN. This is especially the case as species boundaries become clearer or taxonomic revisions are made. As an example, the genus *Sebastes* is a highly speciose group, thought to have undergone an adaptive radiation as recently as 8-9 million years ago [171]. This fact could explain the low haplotype diversity observed for this species (two haplotypes across 98 individuals). Such findings may be due to the underlying assumptions of the model, which are likely to be over-simplistic, particularly that of equality of intraspecific haplotype frequencies. Further, the proposed estimator for the calculation of total haplotype diversity (H^*) (Equation 6) may be a gross overestimate. Despite not being realistic for populations of real species, the reason for adopting a uniform distribution of haplotypes was due to mathematical convenience, in order to make calculations of sample size as simple and as straightforward as possible. This is commonly done in practice, since determining the true distribution of species haplotypes is likely strongly dependent on species under study. Thus, values of N^* are likely overestimates of the true number of specimens that must be randomly sampled in order to observe most haplotype variation that exists for a species [144]. Phillips *et al.*

[144] argue that the use of a limited number of points in the calculation of curve slopes may not be adequate; the authors argue that a fixed proportion of curve points should instead be used. Further, through successively targeting the last 20-15%, 15-10% and the last 10% of species haplotype accumulation curves, in order to observe a statistically-significant change in slope values, the precise point of saturation can be localized [144].

Determining the precise point corresponding to haplotype accumulation curves reaching an asymptote (*i.e.*, having a slope near zero) is difficult. One way this can be accomplished is through employing numerical techniques, specifically iteration. Such methods work by repeatedly recycling computed values into an algorithm; that is, current values are used as starting values to the next iteration until convergence to a solution is achieved. One way this can be realized is through iterating Equation (7) along with the equations for the "measures of sampling closeness" proposed by Phillips *et al.* [144]. This seems to be the most logical way forward in better ascertaining at what level specimen sampling is deemed sufficient and thus, when further collection of specimens should be ceased.

2.5 Future Prospects

The present review explores the issue of sampling in DNA barcoding from the perspective of computational and statistical methodologies. Key sample size studies in the barcoding literature were examined in detail. A lack of consensus exists in the most appropriate number of specimens that must be targeted in order to uncover the majority of

haplotype diversity that exists at the species level for a variety of taxa. This question is similar to the problem of calculating species divergence thresholds for taxon delimitation and is strongly dependent on species abundances, life histories and geographic coverage. To date, few studies exploring sample sizes for DNA barcoding have been conducted. Existing studies ([144, 199]) appear to point to the comprehensive sampling of hundreds to thousands of specimens in order to capture a wide range of standing genetic variation for a given species based on asymptotic behaviour of haplotype accumulation curves.

In order to thoroughly examine the issue of determining specimen sample sizes that are necessary for full assessment of COI DNA barcode haplotype sampling completeness within animal species, relaxation of assumptions inherent in Phillips *et al.*'s [144] sampling model is necessary. Specifically, subsequent approaches should investigate the following:

1. relaxing the assumption of uniformity of species haplotype frequencies;

2. loosening the assumption of panmixia within species; and,

3. testing both above assumptions in tandem.

The incorporation of population structure into models of haplotype sampling is not straightforward, as sampling strategies for DNA barcoding are quite variable and highly dependent on the taxa under study. Thus, this necessitates the introduction of a more spatially-explicit systematic sampling (*e.g.*, phylogeographic) of species genetic variation across distinct taxon boundaries and along phenotypic gradients (*i.e.*, clines). The view of DNA barcoding metaphorically as a "molecular transect", along which a wide range

of intraspecific haplotype diversity can be uncovered, is fitting. Within-species genetic variation has been limited to over-representation of deep sampling of a single or a few populations. If the ultimate goal is to account for levels of standing genetic variation with species, then constraining taxon sampling to narrow geographic regions is not ideal, as this can be considered a form of pseudo-replication. This seems to be an issue of nestedness in sampling and while some depth of sampling within a population is certainly warranted, it cannot be conflated with depth of sampling across populations within a species. In addition, future research should aim to answer the question: is there an optimal threshold for specimen sampling above which no new DNA barcode haplotype variation is likely to be observed for a species? While it should be possible to find this limit for already well-sampled taxa based on trends seen in haplotype accumulation curves, the use of haplotype accumulation curves to estimate sample sizes that are required for full

assessment of COI DNA barcode haplotype sampling completeness has only been tested in one previous study (Zhang *et al.* [199]). Phillips *et al.* [144] expanded on previous studies through proposing a simple and easily implemented method to estimate specimen sample sizes for a number of ray-finned fish species, which are among the most densely sampled to date within BOLD. Sample size optimization for the identification of animal species across wide-ranging geographic scales is key since intraspecific variation within DNA barcodes is not easy to measure, and obtaining large numbers of barcodes that reflect a wide range of intraspecific genetic divergence is sometimes challenging [15]. In addition to being able to report likely required specimen sample sizes necessary to achieve saturation in species

haplotype curves, it would be ideal if DNA barcoding studies could also provide a global measure of geographic dispersion in order to reliably test for cases of isolation-by-distance within species. Unfortunately, no such measure yet exists in this regard, making these kinds of analyses problematic. While model estimates may not be practical, having such a framework at hand that easily allows for the calculation of lower bounds for sample size offers researchers a glimpse into the most appropriate taxon sample sizes to target, and potentially where those taxa should be sampled. More crucially, the present simulation proposed herein can be employed in order to best determine the proper allocation of sampling effort, time and resources [84]. Such work finds application in studies of metabarcoding [185] as well as more broadly to global climate change [141].

The development of a computational simulation of haplotype accumulation curves, a tool that can greatly aid biodiversity scientists in targeting species that will benefit from increased sampling effort, can be employed in order to build and grow BOLD with statistically defensible taxon records, which ultimately will allow more reliable specimen identification. This work is crucial because many taxon records currently in BOLD are known from only single specimens. Further, such a simulation algorithm could aid in species discovery through providing more reliable estimates of intraspecific sample sizes used in the calculation of the barcode gap. Through developing statistically-relevant sample size estimation tools that capture geographic and genetic variation within and between species, researchers will be able to improve sampling design strategies, which will lead to a better understanding (and improved database) of intra and interspecies genetic variation.

As such, new methodologies will fill this void and contribute to the growing literature on sample size estimation for DNA barcoding as well as be implemented as another tool to add to the biodiversity toolbox.

Acknowledgments

We wish to greatly acknowledge the efforts of Rodger Gwiazdowski in providing valuable edits to this manuscript. In addition, comments by Sarah (Sally) Adamowicz improved overall readablity and flow of the manuscript considerably. Finally, two anonymous revievers lent constructive feedback on this work, for which we are greatly appreciative.

This work was supported by a 2016/17 University of Guelph College of Physical and Engineering Science (CPES) Graduate Excellence Entrance Scholarship awarded to JDP.

The Dish With One Spoon Covenant speaks to our collective responsibility to steward and sustain the land and environment in which we live and work, so that all peoples, present and future, may benefit from the sustenance it provides. As we continue to strive to strengthen our relationships with and continue to learn from our Indigenous neighbours, we recognize the partnerships and knowledge that have guided the research conducted in our labs. We acknowledge that the University of Guelph resides in the ancestral and treaty lands of several Indigenous peoples, including the Attawandaron people and the Mississaugas of the Credit, and we recognize and honour our Anishinaabe, Haudenosaunee, and Métis neighbours. We acknowledge that the work presented here has occurred on their traditional lands so that we might work to build lasting partnerships that respect, honour, and value the culture, traditions, and wisdom of those who have lived here since time immemorial.

Author Contributions

JDP conducted the literature review and wrote the manuscript. DJG acted as an advisor in statistics. RHH acted as an advisor in DNA barcoding. All authors contributed to the revision of this manuscript and approved the final version.

Conflict of Interest

None declared.

Chapter 3

`HACSim`: An R package to estimate intraspecific sample sizes for genetic diversity assessment using haplotype accumulation curves

Jarrett D. Phillips[1], Steven H. French[1], Daniel J. Gillis[1] and Robert H. Hanner[2,3]

[1]School of Computer Science

[2]Biodiversity Institute of Ontario

[3]Department of Integrative Biology

Published in *PeerJ Computer Science*

ABSTRACT

Assessing levels of standing genetic variation within species requires a robust sampling for the purpose of accurate specimen identification using molecular techniques such as DNA barcoding; however, statistical estimators for what constitutes a robust sample are currently lacking. Moreover, such estimates are needed because most species are currently represented by only one or a few sequences in existing databases, which we can safely assume are undersampled. Unfortunately, sample sizes of 5-10 specimens per species typically seen in DNA barcoding studies are often insufficient to adequately capture within-species genetic diversity.

Here, we introduce a novel iterative extrapolation simulation algorithm of haplotype accumulation curves, called `HACSim` (**H**aplotype **A**ccumulation **C**urve **Sim**ulator) that can be employed to calculate likely sample sizes needed to observe the full range of DNA barcode haplotype variation that exists for a species. Using uniform haplotype and non-uniform haplotype frequency distributions, the notion of sampling sufficiency θ (the sample size at which sampling accuracy is maximized and above which no new sampling information is likely to be gained) can be gleaned.

`HACSim` can be employed in two primary ways to estimate specimen sample sizes: (1) to simulate haplotype sampling in hypothetical species, and (2) to simulate haplotype sampling in real species mined from public reference sequence databases like the Barcode of Life Data Systems (BOLD) or GenBank for any genomic marker of interest. While our algorithm is globally convergent, runtime is heavily dependent on initial sample

sizes and skewness of the corresponding haplotype frequency distribution.

3.1 Introduction

3.1.1 Background

Earth is in the midst of its sixth mass extinction event and global biodiversity is declining at an unprecedented rate [26]. It is therefore important that species genetic diversity be catalogued and preserved. One solution to address this mounting crisis in a systematic, yet rapid way is DNA barcoding [75]. DNA barcoding relies on variability within a small gene fragment from standardized regions of the genome to identify species, based on the fact that most species exhibit a unique array of barcode haplotypes that are more similar to each other than those of other species (*e.g.*, a barcode "gap"). In animals, the DNA barcode region corresponds to a 648 bp fragment of the 5' terminus of the cytochrome *c* oxidase subunit I (COI) mitochondrial marker [75, 77]. A critical problem since the inception of DNA barcoding involves determining appropriate sample sizes necessary to capture the majority of existing intraspecific haplotype variation for major animal taxa [79, 118, 184]. Taxon sample sizes currently employed in practice for rapid assignment of a species name to a specimen, have ranged anywhere from 1-15 specimens per species [60, 91, 115, 155, 195]; however, oftentimes only 1-2 individuals are actually collected. This trend is clearly reflected within the Barcode of Life Data Systems (BOLD) [150], where an overwhelming number of taxa have only a single record and sequence.

A fitting comparison to the issue of adequacy of specimen sample sizes can be made to the challenge of determining suitable taxon distance thresholds for species separation

on the basis of the DNA barcode gap [118]. It has been widely demonstrated that certain taxonomic groups, such as Lepidoptera (butterflies/moths), are able to be readily separated into distinct clusters largely reflective of species boundaries derived using morphology [23]. However, adoption of a fixed limit of 2% difference between maximum intraspecific distance and minimum interspecific (*i.e*, nearest-neighbour) divergence is infeasible across all taxa [35, 77]. Species divergence thresholds should be calculated from available sequence data obtained through deep sampling of taxa across their entire geographic ranges whenever possible [197]. There is a clear relationship between specimen sample sizes and observed barcoding gaps: sampling too few individuals can give the impression of taxon separation, when in fact none exists [23, 37, 80, 118, 189], inevitably leading to erroneous conclusions [35]. It is thus imperative that barcode gap analyses be based on adequate sample sizes to minimize the presence of false positives. Introducing greater statistical rigour into DNA barcoding appears to be the clear way forward in this respect [23, 111, 129, 142]. The introduction of computational approaches for automated species delimitation such as Generalized Mixed Yule Coalescent (GMYC) [57, 120, 146], Automatic Barcode Gap Discovery (ABGD) [148] and Poisson Tree Processes (PTP; [202]) has greatly contributed to this endeavour in the form of web servers (GMYC, ABGD, PTP) and R packages (GMYC: Species' LImits by Threshold Statistics, `splits` [51]).

Various statistical resampling and population genetic methods, in particular coalescent simulations, for the estimation of sample sizes, have been applied to Lepidoptera (Costa Rican skipper butterflies (*Astraptes fulgerator*)) [199] and European diving beetles (*Agabus*

bipustulatus) [14]. Using Wright's equilibrium island model [194] and Kimura's stepping stone model [95] under varying effective population sizes and migration rates, Zhang *et al.* [199] found that between 156-1985 specimens per species were necessary to observe 95% of all estimated COI variation for simulated specimens of *A. fulgerator*. Conversely, real species data showed that a sample size of 250-1188 individuals is probably needed to capture the majority of COI haplotype variation existing for this species [199]. A subsequent investigation carried out by Bergsten *et al.*[14] found that a random sample of 250 individuals was required to uncover 95% COI diversity in *A. bipustulatus*; whereas, a much smaller sample size of 70 specimens was necessary when geographic separation between two randomly selected individuals was maximized.

Others have employed more general statistical approaches. Based on extensive simulation experiments, through employing the Central Limit Theorem (CLT), Luo *et al.* [111] suggested that no fewer than 20 individuals per species be sampled. Conversely, using an estimator of sample size based on the Method of Moments, an approach to parameter estimation relying on the Weak Law of Large Numbers [139], sample sizes ranging from 150-5400 individuals across 18 species of ray-finned fishes (Chordata: Actinopterygii) were found by Phillips *et al.* [144].

Haplotype accumulation curves paint a picture of observed standing genetic variation that exists at the species level as a function of expended sampling effort [142, 144]. Haplotype sampling completeness can then be gauged through measuring the slope of the curve, which gives an indication of the number of new haplotypes likely to be uncovered

with additional specimens collected. For instance, a haplotype accumulation curve for a hypothetical species having a slope of 0.01 suggests that only one previously unseen haplotype will be captured for every 100 individuals found. This is strong evidence that the haplotype diversity for this species has been adequately sampled. Thus, further recovery of specimens of such species provide limited returns on the time and money invested to sequence them. Trends observed from generated haplotype accumulation curves for the 18 actinopterygian species assessed by Phillips *et al.* [144], which were far from reaching an asymptote, corroborated the finding that the majority of intraspecific haplotypes remain largely unsampled in Actinopterygii for even the best-represented species in BOLD. Estimates obtained from each of these studies stand in sharp contrast to sample sizes typically reported within DNA barcoding studies.

Numerical optimization methods are required to obtain reasonable approximations to otherwise complex questions. Many such problems proceed via the iterative method, whereby an initial guess is used to produce a sequence of successively more precise (and hopefully more accurate) approximations. Such an approach is attractive, as resulting solutions can be made as precise as desired through specifying a given tolerance cutoff. However, in such cases, a closed-form expression for the function being optimized is known *a priori*. In many instances, the general path (behaviour) of the search space being explored is the only information known, and not its underlying functional form. In this paper, we take a middle-ground approach that is an alternative to probing sampling completeness on the basis of haplotype accumulation curve slope measurement. To this

end, iteration is applied to address the issue of relative sample size determination for DNA barcode haplotype sampling completeness, a technique suggested by Phillips *et al.* [142]. Given that specimen collection and processing is quite a laborious and costly endeavour [22, 168], the next most direct solution to an otherwise blind search strategy is to employ computational simulation that approximates specimen collection in the field. The main contribution of this work is the introduction of a new, easy-to-use R package implementing a novel statistical optimization algorithm to estimate sample sizes for assessment of genetic diversity within species based on saturation observed in haplotype accumulation curves. Here, we present a novel nonparametric stochastic (Monte Carlo) iterative extrapolation algorithm for the generation of haplotype accumulation curves based on the approach of [144]. Using the statistical environment R [149], we examine the effect of altering species haplotype frequencies on the shape of resulting curves to inform on likely required sample sizes needed for adequate capture of within-species haplotype variation. Proof-of-concept of our method is illustrated through both hypothetical examples and real DNA sequence data.

3.1.2 Motivation

Consider N DNA sequences that are randomly sampled for a given species of interest across its known geographic range, each of which correspond to a single specimen. Suppose further that H^* of such sampled DNA sequences are unique (*i.e.*, are distinct haplotypes). This scenario leads naturally to the following question: What is N^*, the estimated total number of DNA sequence haplotypes that exist for a species? Put another

way, what sample size (number of specimens) is needed to capture the existing haplotype variation for a species?

The naïve approach (adopted by Phillips *et al.* [144]) would be to ignore relative frequencies of observed haplotypes; that is, assume that species haplotypes are equally probable in a species population. Thus, in the absence of any information, the best one can do is adopt a uniform distribution for the number of sampled haplotypes. Such a path leads to obtaining gross overestimates for sufficient sampling [144]. A much better approach uses all available haplotype data to arrive at plausible estimates of required taxon sample sizes. This latter method is explored here in detail.

3.2 Methods

3.2.1 Haplotype Accumulation Curve Simulation Algorithm

Algorithm Functions

Our algorithm, `HACSim` (short for **H**aplotype **A**ccumulation **C**urve **Sim**ulator), consisting of two user-defined R functions, `HAC.sim()` and `HAC.simrep()`, was created to run simulations of haplotype accumulation curves based on user-supplied parameters. The simulation treats species haplotypes as distinct character labels relative to the number of individuals possessing a given haplotype. The usual convention in this regard is that Haplotype 1 is the most frequent, Haplotype 2 is the next most frequent, *etc.* [65]. A haplotype network represents this scheme succinctly (**Fig. 3.1**).

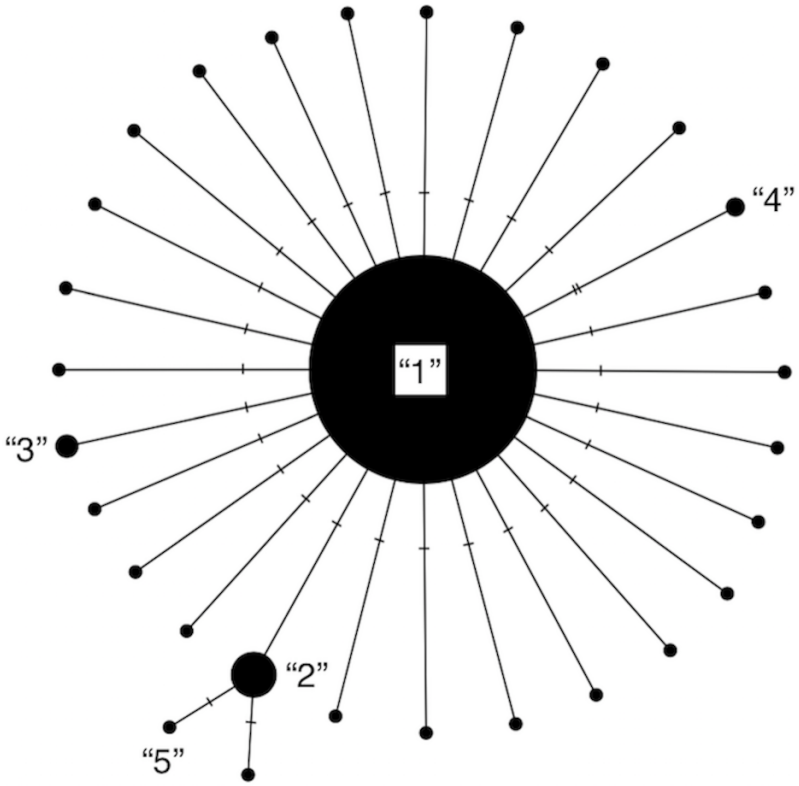

Figure 3.1: Modified haplotype network from Phillips *et al.* [142]. Haplotypes are labelled according to their absolute frequencies such that the most frequent haplotype is labelled "1", the second-most frequent haplotype is labelled "2", *etc*. and is meant to illustrate that much species locus variation consists of rare haplotypes at very low frequency (typically only represented by 1 or 2 specimens). Thus, species showing such patterns in their haplotype distributions are probably grossly under-respresented in public sequence databases like BOLD and GenBank.

Such an implementation closely mimics that seen in natural species populations, as each character label functions as a unique haplotype linked to a unique DNA barcode sequence. The algorithm then randomly samples species haplotype labels in an iterative fashion with replacement until all unique haplotypes have been observed. This process continues until all species haplotypes have been sampled. The idea is that levels of species haplotypic

variation that are currently catalogued in BOLD can serve as proxies for total haplotype diversity that may exist for a given species. This is a reasonable assumption given that, while estimators of expected diversity are known (*e.g.*, Chao1 abundance) [27], the frequencies of unseen haplotypes are not known *a priori*. Further, assuming a species is sampled across its entire geographic range, haplotypes not yet encountered are presumed to occur at low frequencies (otherwise they would likely have already been sampled).

Because R is an interpreted programming language (*i.e.*, code is run line-by-line), it is slow compared to faster alternatives which use compilation to convert programs into machine-readable format; as such, to optimize performance of the present algorithm in terms of runtime, computationally-intensive parts of the simulation code were written in the C++ programming language and integrated with R via the packages `Rcpp` [44] and `RcppArmadillo` [45]. This includes function code to carry out haplotype accumulation (via the function `accumulate()`, which is not directly called by the user). A further reason for turning to C++ is because some R code (e.g. nested 'for' loops) is not easily vectorized, nor can parallelization be employed for speed improvement due to loop dependence. The rationale for employing R for the present work is clear: R is free, open-source software that it is gaining widespread use within the DNA barcoding community due to its ease-of-use and well-established user-contributed package repository (Comprehensive R Archive Network (CRAN)). As such, the creation and disemination of `HACSim` as a R framework to assess levels of standing genetic variation within species is greatly facilitated.

A similar approach to the novel one proposed here to automatically generate haplotype accumulation curves from DNA sequence data is implemented in the R package `spider` (SPecies IDentity and Evolution in R; [20]) using the `haploAccum()` function. However, the approach, which formed the basis of earlier work carried out by Phillips *et al.* [144], is quite restrictive in its functionality and, to our knowledge, is currently the only method available to generate haplotype accumulation curves in R because `spider` generates haplotype accumulation curves from DNA sequence alignments only and is not amenable to inclusion of numeric inputs for specimen and haplotype numbers. Thus, the method could not be easily extended to address our question. This was the primary reason for the proposal of a statistical model of sampling sufficiency by Phillips *et al.* [144] and its extension described herein.

Algorithm Parameters

At present, the algorithm (consisting of `HAC.sim() and HAC.simrep()`) takes 13 arguments as input (**Table 3.1**).

Table 3.1: Parameters inputted (first 7) and outputted (last six) by `HAC.sim()` and `HAC.simrep()`, along with their definitions. **Range** refers to plausible values that each parameter can assume within the haplotype accumulation curve simulation algorithm. [and] indicate that a given value is included in the range interval; whereas, (and) indicate that a given value is excluded from the range interval. Simulation progress can be tracked through setting `progress = TRUE` within `HACHypothetical()` or `HACReal()`. Users can optionally specify that a file be created containing all information outputted to the R console (via the argument `filename`, which can be named as the user wishes).

Parameter	Definition	Range
N	total number of specimens/DNA sequences	(1, ∞)
H^*	total number of unique haplotypes	(1, N]
`probs`	haplotype probability distribution vector	(0, 1)
p	proportion of haplotypes to recover	(0, 1]
`perms`	total number of permutations	(1, ∞)
`input.seqs`	analyze FASTA file of species DNA sequences	`TRUE`, `FALSE`
`conf.level`	desired confidence level for confidence interval calculation	(0, 1)
H	cumulative mean number of haplotypes sampled	[1, H^*]
$H^* - H$	cumulative mean number of haplotypes not sampled	[0, H^*)
$R = \frac{H}{H^*}$	cumulative mean fraction of haplotypes sampled	(0, 1]
$\frac{H^* - H}{H^*}$	cumulative mean fraction of haplotypes not sampled	[0, 1)
N^*	mean specimen sample size corresponding to H^*	[N, ∞)
$N^* - N$	mean number of individuals not sampled	[0, N]

A user must first specify the number of observed specimens/DNA sequences (N) and the number of observed haplotypes (*i.e.*, unique DNA sequences) (H^*) for a given species. Both N and H^* must be greater than one. Clearly, N must be greater than or equal to H^*.

Next, the haplotype frequency distribution vector must be specified. The `probs` argument allows for the inclusion of both common and rare species haplotypes according to user interest (*e.g.*, equally frequent haplotypes, or a single dominant haplotype). The resulting `probs` vector must have a length equal to H^*. For example, if $H^* = 4$, `probs` must contain four elements. The total probability of all unique haplotypes must sum to one.

The user can optionally input the fraction of observed haplotypes to capture p. By

default, p = 0.95, mirroring the approach taken by both Zhang *et al.* [199] and Bergsten *et al.* [14] who computed intraspecific sample sizes needed to recover 95% of all haplotype variation for a species. At this level, the generated haplotype accumulation curve reaches a slope close to zero and further sampling effort is unlikely to uncover any new haplotypes. However, a user may wish to obtain sample sizes corresponding to different haplotype recovery levels, *e.g.*, p = 0.99 (99% of all estimated haplotypes found). In the latter scenario, it can be argued that 100% of species haplotype variation is never actually achieved, since with greater sampling effort, additional haplotypes are almost surely to be found; thus, a true asymptote is never reached. In any case, simulation completion times will vary depending on inputted parameter values, such as `probs`, which controls the skewness of the observed haplotype frequency distribution.

The `perms` argument is in place to ensure that haplotype accumulation curves "smooth out" and tend to H^* asymptotically as the number of permutations (replications) is increased. The effect of increasing the number of permutations is an increase in statistical accuracy and consequently, a reduction in variance. The proposed simulation algorithm outputs a mean haplotype accumulation curve that is the average of `perms` generated haplotype accumulation curves, where the order of individuals that are sampled is randomized. Each of these `perms` curves is a randomized step function (a sort of random walk), generated according to the number of haplotypes found. A permutation size of 1000 was used by Phillips *et al.* [144] because smaller permutation sizes yielded non-smooth (noisy) curves. Permutation sizes larger than 1000 typically resulted in greater computation

time, with no noticeable change in accumulation curve behaviour [144]. By default, `perms` = 10000 (in contrast to Phillips *et al.* [144]), which is comparable to the large number of replicates typically employed in statistical bootstrapping procedures needed to ensure accuracy of computed estimates [47]. Sometimes it will be necessary for users to sacrifice accuracy for speed in the presence of time constraints. This can be accomplished through decreasing `perms`. Doing so however will result in only near-optimal solutions for specimen sample sizes. In some cases, it may be necessary to increase `perms` to further smooth out the curves (to ensure monotonicity), but this will increase algorithm runtime substantially.

Should a user wish to analyze their own intraspecific COI DNA barcode sequence data (or sequence data from any single locus for that matter), setting `input.seqs = TRUE` allows this (via the `read.dna()` function in `ape`). In such a case, a pop-up file window will prompt the user to select the formatted FASTA file of aligned/trimmed sequences to be read into R. When this occurs, arguments for N, H^* and `probs` are set automatically by the algorithm via functions available in the R packages `ape` (Analysis of Phylogenetics and Evolution) [136] and `pegas` (Population and Evolutionary Genetics Analysis System) [135]. Users must be aware however that the number of observed haplotypes treated by `pegas` (via the `haplotype()` function) may be overestimated if missing/ambiguous nucleotide data are present within the final alignment input file. Missing data are explicitly handled by the `base.freq()` function in the `ape` package. When this occurs, R will output a warning that such data are present within the alignment. Users should therefore

consider removing sequences or sites comprising missing/ambiguous nucleotides. This step can be accomplished using external software such as MEGA (Molecular Evolutionary Genetics Analysis; [99]). The BARCODE standard [68] was developed to help identify high quality sequences and can be used as a quality filter if desired. Exclusion of low-quality sequences also has the advantage of speeding up compution time of the algorithm significantly.

Options for confidence interval (CI) estimation and graphical display of haplotype accumulation is also available via the argument `conf.level`, which allows the user to specify the desired level of statistical confidence. CIs are computed from the sample $\frac{\alpha}{2}100\%$ and $(1-\frac{\alpha}{2})100\%$ quantiles of the haplotype accumulation curve distribution. The default is `conf.level` = 0.95, corresponding to a confidence level of 95%. High levels of statistical confidence (*e.g.*, 99%) will result in wider confidence intervals; whereas low confidence leads to narrower interval estimates.

How Does HACSim Work?

Haplotype labels are first randomly placed on a two-dimensional spatial grid of size `perms` $\times$ N (read `perms` rows by N columns) according to their overall frequency of occurrence (**Fig. 3.2**).

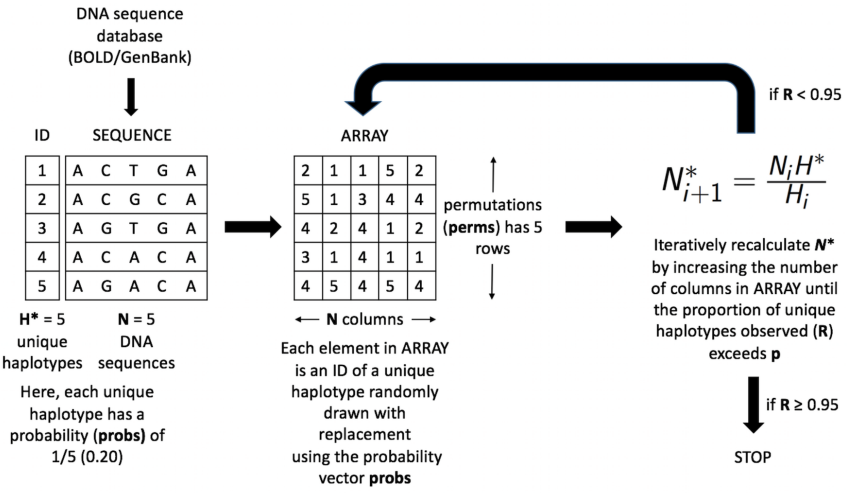

Figure 3.2: Schematic of the `HACSim` optimization algorithm (setup, initialization and iteration). Shown is a hypothetical example for a species mined from a biological sequence database like BOLD or GenBank with N = 5 sampled specimens (DNA sequences) possessing H^* = 5 unique haplotypes. Each haplotype has an associated numeric ID from 1-H^* (here, 1-5). Haplotype labels are randomly assigned to cells on a two-dimensional spatial array (ARRAY) with `perms` rows and N columns. All haplotypes occur with a frequency of 20%, (*i.e.*, `probs` = (1/5, 1/5, 1/5, 1/5, 1/5)). Specimen and haplotype information is then fed into a black box to iteratively optimize the likely required sample size (N^*) needed to capture a proportion of at least `p` haplotypes observed in the species sample.

The cumulative mean number of haplotypes is then computed along each column (*i.e.*, for every specimen). If all H^* haplotypes are not observed, then the grid is expanded to a size of `perms` $\times$ N^* and the observed haplotypes enumerated. Estimation of specimen sample sizes proceeds iteratively, in which the current value of N^* is used as a starting value to the next iteration (**Fig. 3.2**). An analogy here can be made to a game of golf: as one aims towards the hole and hits the ball, it gets closer and closer to the hole; however, one does not know the number of times to hit the ball before it lands in the hole. It is important to note that since sample sizes must be whole values, estimates of N^* found at each iteration

are rounded up to the next whole number. Even though this approach is quite conservative, it ensures that estimates are adequately reflective of the population from which they were drawn. `HAC.sim()`, which is called internally from `HAC.simrep()`, performs a single iteration of haplotype accumulation for a given species. In the case of real species, resulting output reflects current levels of sampling effort found within BOLD (or another similar sequence repository such as GenBank) for a given species. If the desired level of haplotype recovery is not reached, then `HAC.simrep()` is called to perform successive iterations until the observed fraction of haplotypes captured (R) is at least p. This stopping criterion is the termination condition necessary to halt the algorithm as soon as a "good enough" solution has been found. Such criteria are widely employed within numerical analysis. At each step of the algorithm, a dataset, in the form of a dataframe (called "d") consisting of the mean number of haplotypes recovered (called `means`), along with the estimated standard deviation (`sd`) and the number of specimens sampled (`specs`) is generated. The estimated required sample size (*N**) to recover a given proportion of observed species haplotypes corresponds to the endpoint of the accumulation curve. An indicator message is additionally outputted informing a user as to whether or not the desired level of haplotype recovery has been reached. The algorithm is depicted in **Fig. 3.3**.

Iterative Extrapolation Algorithm to Calculate N^*

INPUT: N, H^*, `probs`, `perms`, p, H, R $(= \frac{H}{H^*})$
OUTPUT: N^*

(1) **SET** $i = 1$ (*initialize iterations*);
(2) **SET** $N^* = N$ (*specify initial guess*)

WHILE R $< p$
(3) **SET** $i = i + 1$ (*update iterations*);
(4) **SET** $N^*_{i+1} = \frac{N_i H^*}{H_i}$ (*compute* N^*);
(5) **IF** $N^*_{i+1} = N_i$, **STOP**, **ELSE** return to (3)

END.

Figure 3.3: Iterative extrapolation algorithm pseudocode for the computation of taxon sampling sufficiency employed within `HACSim`. A user must input N, H^* and `probs` to run simulations. Other function arguments required by the algorithm have default values and are not necessary to be inputted unless the user wishes to alter set parameters.

In **Fig. 3.3**, all input parameters are known *a priori* except H_i, which is the number of haplotypes found at each iteration of the algorithm, and $R_i = \frac{H_i}{H^*}$, which is the observed fraction of haplotype recovery at iteration i. The equation to compute N^*

$$N^*_{i+1} = N_i + \frac{N_i}{H_i}(H^* - H_i) = \frac{N_i H^*}{H_i} = \frac{N_i}{R_i} \tag{3.1}$$

is quite intuitive since as H_i approaches H^*, $H^* - H_i$ approaches zero, $R_i = \frac{H_i}{H^*}$ approaches one, and consequently, N_i approaches N^*. In the first part of the above equation, the quantity $\frac{N_i}{H_i}(H^* - H_i)$ is the amount by which the haplotype accumulation curve is extrapolated, which incorporates random error and uncertainty regarding the true value of θ in the search space being explored. Nonparametric estimates formed from the above iterative method produce a convergent monotonically-increasing sequence, which becomes

closer and closer to *N** as the number of iterations increase; that is,

$$N_1^* \leq N_2^* \leq ... \leq N_i^* \leq N_{i+1}^* \rightarrow N^* \tag{3.2}$$

which is clearly a desirable property. Since haplotype accumulation curves are bounded below by one and bounded above by *H**, then the above sequence has a lower bound equal to the initial guess for specimen sampling sufficiency (N) and an upper bound of *N**.

Along with the iterated haplotype accumulation curves and haplotype frequency barplots, simulation output consists of the five initially proposed "measures of sampling closeness", the estimate of θ (*N**) based on Phillips *et al.*'s [144] sampling model, in addition to the number of additional samples needed to recover all estimated total haplotype variation for a given species ($N^* - N$; **Fig. 3.4**) (**Table 3.1**).

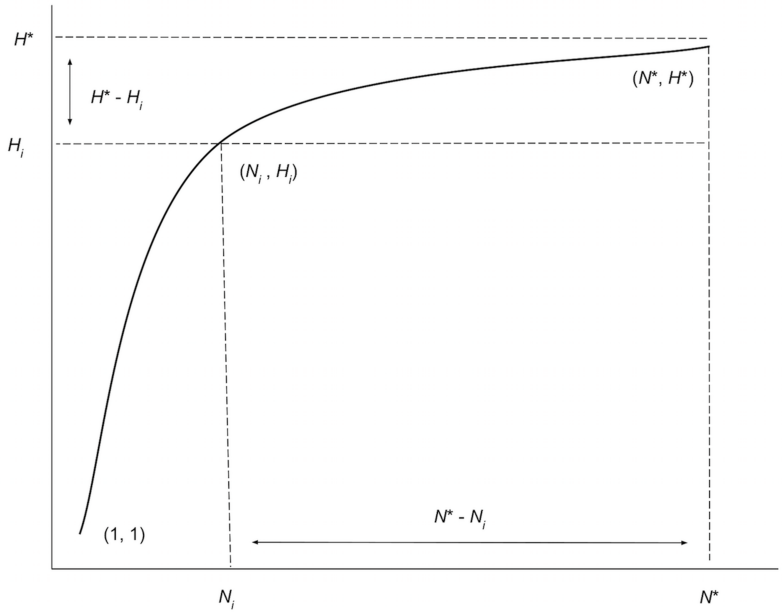

Figure 3.4: Graphical depiction of the iterative extrapolation sampling model as described in detail herein. The figure is modified from Phillips *et al.* [142]. The *x*-axis is meant to depict the number of specimens sampled, whereas the *y*-axis is meant to convey the cumulative number of unique haplotypes uncovered for every additional individual that is randomly sampled. N_i and H_i refer respectively to specimen and haplotype numbers that are observed at each iteration (i) of `HACSim` for a given species. N^* is the total sample size that is needed to capture all H^* haplotypes that exist for a species.

These five quantities are given as follows: (1) Mean number of haplotypes sampled: H_i, (2) Mean number of haplotypes not sampled: $H^* - H_i$, (3) Proportion of haplotypes sampled: $\frac{H_i}{H^*}$, (4) Proportion of haplotypes not sampled: $\frac{H^* - H_i}{H^*}$, (5) Mean number of individuals not sampled: $N^* - N_i = \frac{N_i}{H_i}(H^* - H_i)$ and are analogous to absolute and relative approximation error metrics seen in numerical analysis. It should be noted that the mean number of haplotypes captured at each iteration, H_i, will not necessary be increasing, even though estimates of the cumulative mean value of N^* are. It is easily seen above that H_i approaches H^* with increasing number of iterations. Similarly, as the simulation

progresses, $H^* - H_i$, $\dfrac{H^* - H_i}{H^*}$ and $N^* - N_i = \frac{N_i}{H_i}(H^* - H_i)$ all approach zero, while $\dfrac{H_i}{H^*}$ approaches one. The rate at which curves approach H^* depends on inputs to both `HAC.sim()` and `HAC.simrep()`. Once the algorithm has converged to the desired level of haplotype recovery, a summary of findings is outputted consisting of (1) the initial guess (N) for sampling sufficiency; (2) the total number of iterations until convergence and simulation runtime (in seconds); (3) the final estimate (N^*) of sampling sufficiency, along with an approximate $(1 - \alpha)100\%$ confidence interval (see next paragraph); and, (4) the number of additional specimens required to be sampled ($N^* - N$) from the initial starting value. Iterations are automatically separated by a progress meter for easy visualization.

An approximate symmetric $(1 - \alpha)100\%$ CI for θ is derived using the (first order) Delta Method [25]. This approach relies on the asymptotic normality result of the CLT and employs a first-order Taylor series expansion around θ to arrive at an approximation of the variance (and corresponding standard error) of N^*. Such an approach is convenient since the sampling distribution of N^* would likely be difficult to compute exactly due to specimen sample sizes being highly taxon-dependent. An approximate (large sample) $(1 - \alpha)100\%$ CI for θ is given by

$$N^* \pm z_{1-\frac{\alpha}{2}} \left(\frac{\hat{\sigma}_H}{H} \sqrt{N^*} \right) \tag{3.3}$$

where $z_{1-\frac{\alpha}{2}}$ denotes the appropriate critical value from the standard Normal distribution and $\hat{\sigma}_H$ is the estimated standard deviation of the mean number of haplotypes recovered at

*N**. The interval produced by this approach is quite tight, shrinking as H_i tends to *H**. By default, `HACSim` computes 95% confidence intervals for the abovementioned quantities.

It is important to consider how a confidence interval for θ should be interpreted. For instance, a 95% CI for θ of (L, U), where L and U are the lower and upper endpoints of the confidence interval respectively, does *not* mean that the true sampling sufficiency lies between (L, U) with 95% probability. Instead, resulting confidence intervals for θ are themselves random and should be interpreted in the following way: with repeated sampling, one can be $(1 - \alpha)100\%$ confident that the true sampling sufficency for $p\%$ haplotype recovery for a given species lies in the range (L, U) $(1 - \alpha)100\%$ of the time. That is, on average, $(1 - \alpha)100\%$ of constructed confidence intervals will contain θ $(1 - \alpha)100\%$ of the time. It should be noted however that as given computed confidence intervals are only approximate in the limit, desired nominal probability coverage may not be achieved. In other words, the proportion of times calculated $(1 - \alpha)100\%$ intervals actually contain θ may not be met.

`HACSim` has been implemented as an object-oriented framework to improve modularity and overall user-friendliness. Scenarios of hypothetical and real species are contained within helper functions which comprise all information necessary to run simulations successfully without having to specify certain function arguments beforehand. To carry out simulations of sampling haplotypes from hypothetical species, the function `HACHypothetical()` must first be called. Similarly, haplotype sampling for real species is handled by the function `HACReal()`. In addition to all input parameters rquired

by `HAC.sim()` and `HAC.simrep()` outlined in **Table 3.1**, both `HACHypothetical()` and `HACReal()` take further arguments. Both functions take the optional argument `filename` which is used to save results outputted to the R console to a CSV file. When either `HACHypothetical()` or `HACReal()` is invoked (*i.e.*, assigned to a variable), an object herein called `HACSObj` is created containing the 13 arguments employed by `HACSim` in running simulations. Note the generated object can have any name the user desires. Further, all simulation variables are contained in an environment called 'envr' that is hidden from the user.

3.3 Results

Here, we outline some simple examples that highlight the overall functionality of `HACSim`. When the code below is run, outputted results will likely differ from those depicted here since our method is inherently stochastic. Hence, it should be stressed that there is not one single solution for the problem at hand, but rather multiple solutions [165]. This is in contrast to a completely deterministic model, where a given input always leads to the same unique output. To ensure reproducibility, the user can set a random seed value using the base R function `set.seed()` prior to running `HAC.simrep()`. It is important that a user set a working directory in R prior to running `HACSim`, which will ensure all created files ('seqs.fas' and 'output.csv') are stored in a single location for easy access and reference at a later time. In all scenarios, default parameters were unchanged (`perms` = 10000, $p = 0.95$).

3.3.1 Application of `HACSim` to Hypothetical Species

Equal Haplotype Frequencies

Fig. 3.5 shows sample graphical output of the proposed haplotype accumulation curve simulation algorithm for a hypothetical species with $N = 100$ and $H^* = 10$. All haplotypes are assumed to occur with equal frequency (*i.e.*, `probs` = 0.10). Algorithm output is shown below.

```
## Set parameters for hypothetical species ##
> N <- 100 # total number of sampled individuals
> Hstar <- 10 # total number of haplotypes
> probs <- rep(1/Hstar, Hstar) # equal haplotype frequency

### Run simulations ###
> HACSObj <- HACHypothetical(N = N, Hstar = Hstar, probs = probs) # call helper function
# set seed here if desired, e.g., set.seed(12345)
> HAC.simrep(HACSObj)
Simulating haplotype accumulation...
|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  10
Mean number of haplotypes not sampled:  0
Proportion of haplotypes sampled:  1
Proportion of haplotypes not sampled:  0

Mean value of N*:  100
Mean number of specimens not sampled:  0

Desired level of haplotype recovery has been reached

---------- Finished.  ----------
The initial guess for sampling sufficiency was N = 100 individuals
The algorithm converged after 1 iterations and took 3.637 s
The estimate of sampling sufficiency for p = 95% haplotype recovery is N* = 100
individuals ( 95% CI: 100-100 )
The number of additional specimens required to be sampled for p = 95% haplotype recovery
is N* - N = 0 individuals
```

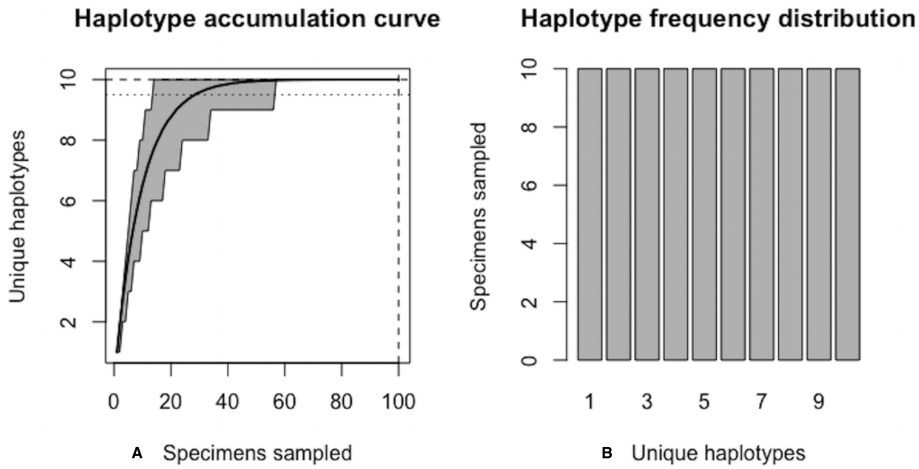

Figure 3.5: Graphical output of `HAC.sim()` for a hypothetical species with equal haplotype frequencies. **A**: Iterated haplotype accumulation curve. **B**: Corresponding haplotype frequency barplot. For the generated haplotype accumulation curve, the 95% confidence interval for the number of unique haplotypes accumulated is depicted by gray error bars. Dashed lines depict the observed number of haplotypes (*i.e.*, *RH**) and corresponding number of individuals sampled found at each iteration of the algorithm. The dotted line depicts the expected number of haplotypes for a given haplotype recovery level (here, $p = 95\%$) (*i.e.*, *pH**). In this example, $R = 100\%$ of the $H^* = 10$ estimated haplotypes have been recovered for this species based on a sample size of only $N = 100$ specimens.

Algorithm output shows that $R = 100\%$ of the $H^* = 10$ haplotypes are recovered from the random sampling of $N = 100$ individuals, with lower and upper 95% confidence limits of 100-100. No additional specimens need to be collected ($N^* - N = 0$). Simulation results, consisting of the six "measures of sampling closeness" computed at each iteration, can be optionally saved in a comma-separated value (CSV) file called 'output.csv' (or another filename of the user's choosing). **Fig. 3.5** shows that when haplotypes are equally frequent in species populations, corresponding haplotype accumulation curves reach an asymptote very quickly. As sampling effort is increased, the confidence interval becomes narrower, thereby reflecting one's increased confidence in having likely sampled the majority of

haplotype variation existing for a given species. Expected counts of the number of specimens possessing a given haplotype can be found from running `max(envr$d$specs) * envr$probs` in the R console once a simulation has converged. However, real data suggest that haplotype frequencies are not equal.

Unequal Haplotype Frequencies

Fig. 3.6 and **Fig. 3.7** show sample graphical output of the proposed haplotype accumulation curve simulation algorithm for a hypothetical species with N = 100 and H^* = 10. All haplotypes occur with unequal frequency. Haplotypes 1-3 each have a frequency of 30%, while the remaining seven haplotypes each occur with a frequency of *c.* 1.4%.

```
## Set parameters for hypothetical species ##
> N <- 100
> Hstar <- 10
> probs <- c(rep(0.30, 3), rep(0.10/7, 7)) # three dominant haplotypes each with 30%
frequency

### Run simulations ###
> HACSObj <- HACHypothetical(N = N, Hstar = Hstar, probs = probs)
> HAC.simrep(HACSObj)
Simulating haplotype accumulation...
|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  8.3291
Mean number of haplotypes not sampled:  1.6709
Proportion of haplotypes sampled:  0.83291
Proportion of haplotypes not sampled:  0.16709

Mean value of N*:  120.061
Mean number of specimens not sampled:  20.06099

Desired level of haplotype recovery has not yet been reached
|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  9.2999
Mean number of haplotypes not sampled:  0.7001
Proportion of haplotypes sampled:  0.92999
Proportion of haplotypes not sampled:  0.07001
```

```
Mean value of N*:  179.5718
Mean number of specimens not sampled:  12.57182

Desired level of haplotype recovery has not yet been reached

|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  9.5358
Mean number of haplotypes not sampled:  0.4642
Proportion of haplotypes sampled:  0.95358
Proportion of haplotypes not sampled:  0.04642

Mean value of N*:  188.7623
Mean number of specimens not sampled:  8.762348

Desired level of haplotype recovery has been reached

---------- Finished.  ----------
The initial guess for sampling sufficiency was N = 100 individuals
The algorithm converged after 6 iterations and took 33.215 s
The estimate of sampling sufficiency for p = 95% haplotype recovery is N* = 180
individuals ( 95% CI: 178.278-181.722)


The number of additional specimens required to be sampled for p = 95% haplotype recovery


is N* - N = 80 individuals
```

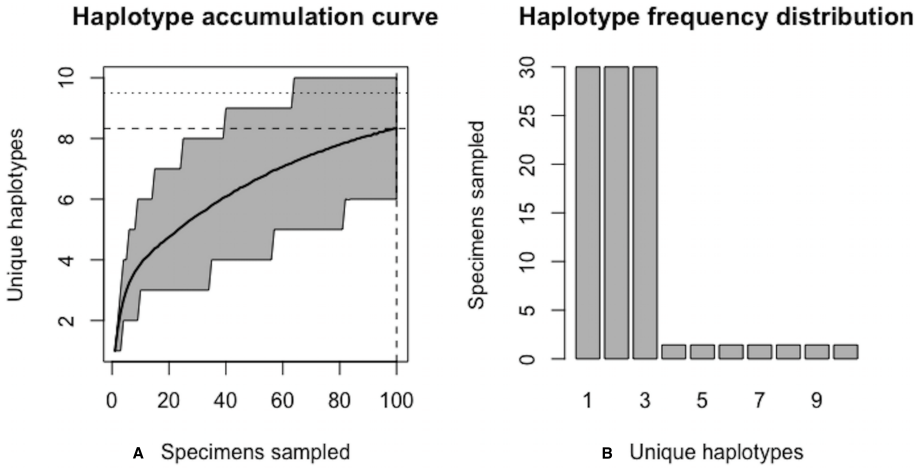

Figure 3.6: Initial graphical output of `HAC.sim()` for a hypothetical species having three dominant haplotypes. In this example, initially, only R = 83.3% of the H^* = 10 estimated haplotypes have been recovered for this species based on a sample size of N = 100 specimens.

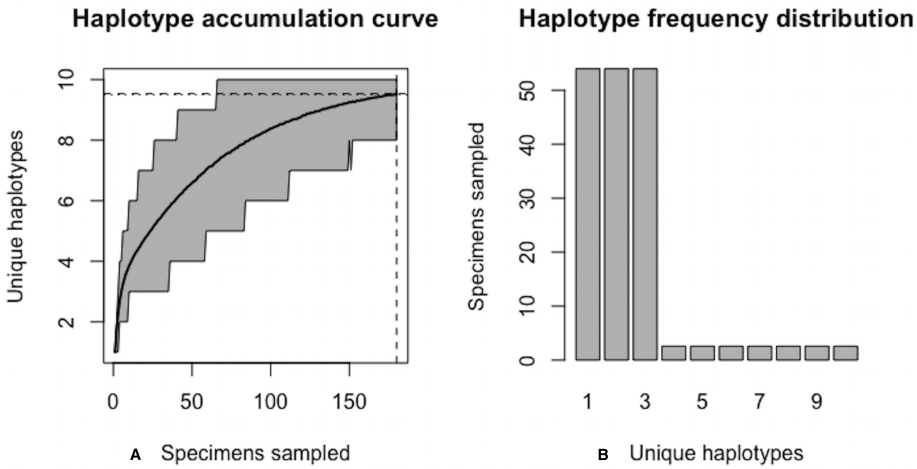

Figure 3.7: Final graphical output of `HAC.sim()` for a hypothetical species having three dominant haplotypes. In this example, upon convergence, R = 95.4% of the H^* = 10 estimated haplotypes have been recovered for this species based on a sample size of N = 180 specimens.

Note that not all iterations are displayed above for the sake of brevity; only the first and last two iterations are given. With an initial guess of N = 100, only R = 83.3% of all H^* = 10 observed haplotypes are recovered. The value of N^* = 121 in the first iteration above serves as an improved initial guess of the true sampling sufficiency, which is an unknown quantity that is being estimated. This value is then fed back into the algorithm and the process is repeated until convergence is reached.

Using Equation (1), the improved sample size was calculated as $N^* = 100 + \frac{100}{8.3291}(10 - 8.3291) = 120.061$. After one iteration, the curve has been extrapolated by an additional $N^* - N_i$ = 20.06099 individuals. Upon convergence, R = 95.4% of all observed haplotypes are captured with a sample size of N^* = 180 specimens, with a 95% CI of 178.278-181.722. Given that $N = 100$ individuals have already been

sampled, the number of additional specimens required is $N^* - N$ = 80 individuals. The user can verify that sample sizes close to that found by `HACSim` are needed to capture 95% of existing haplotype variation. Simply set N = N* = 180 and rerun the algorithm. The last iteration serves as a check to verify that the desired level of haplotype recovery has been achieved. The value of N* = 188.7623 that is outputted at this step can be used as a good starting guess to extrapolate the curve to higher levels of haplotype recovery to save on the number of iterations required to reach convergence. To do this, one simply runs `HACHypothetical()` with N = 189.

3.3.2 Application of `HACSim` to Real Species

Because the proposed iterative haplotype accumulation curve simulation algorithm simply treats haplotypes as numeric labels, it is easily generalized to any biological taxa and genetic loci for which a large number of high-quality DNA sequence data records is available in public databases such as BOLD. In the following examples, `HACSim` is employed to examine levels of standing genetic variation within animal species using 5'-COI.

Lake Whitefish (*Coregonus clupeaformis*)

An interesting case study on which to focus is that of Lake whitefish (*Coregonus clupeaformis*). Lake whitefish are a commercially, culturally, ecologically and economically important group of salmonid fishes found throughout the Laurentian Great Lakes in Canada and the United States, particularly to the Saugeen Ojibway First Nation

(SON) of Bruce Peninsula in Ontario, Canada, as well as non-indigenous fisheries [156].

The colonization of refugia during Pleistocene glaciation is thought to have resulted in high levels of cryptic species diversity in North American freshwater fishes [5, 6, 7, 86]. [132] wished to investigate this hypothesis in larval Lake Huron lake whitefish. Despite limited levels of gene flow and likely formation of novel divergent haplotypes in this species, surprisingly, no evidence of deep evolutionary lineages was observed across the three major basins of Lake Huron despite marked differences in larval phenotype and adult fish spawning behaviour [132]. This may be the result of limited sampling of intraspecific genetic variation, in addition to presumed panmixia [132]. While lake whitefish represent one of the most well-studied fishes within BOLD, sampling effort for this species has nevertheless remained relatively static over the past few years. Thus, lake whitefish represent an ideal species for further exploration using `HACSim`.

In applying the developed algorithm to real species, sequence data preparation methodology followed that which is outlined in Phillips *et al.* [144]. Curation included the exclusion of specimens linked to GenBank entries, since those records without the BARCODE keyword designation lack appropriate metadata central to reference sequence library construction and management [68]. Our approach here was solely to assess comprehensiveness of single genomic sequence databases rather than incorporating sequence data from multiple repositories; thus, all DNA barcode sequences either originating from, or submitted to GenBank were not considered further. As well, the presence of base ambiguities and gaps/indels within sequence alignments can lead to bias

in estimate haplotype diversity for a given species.

Currently (as of November 28, 2018), BOLD contains public (both barcode and non-barcode) records for 262 *C. clupeaformis* specimens collected from Lake Huron in northern parts of Ontario, Canada and Michigan, USA. Of the barcode sequences, N = 235 are of high quality (full-length (652 bp) and comprise no missing and/or ambiguous nucleotide bases). Haplotype analysis reveals that this species currently comprises H^* = 15 unique COI haplotypes. Further, this species shows a highly-skewed haplotype frequency distribution, with a single dominant haplotype accounting for *c.* 91.5% (215/235) of all individuals (**Fig. 3.8**).

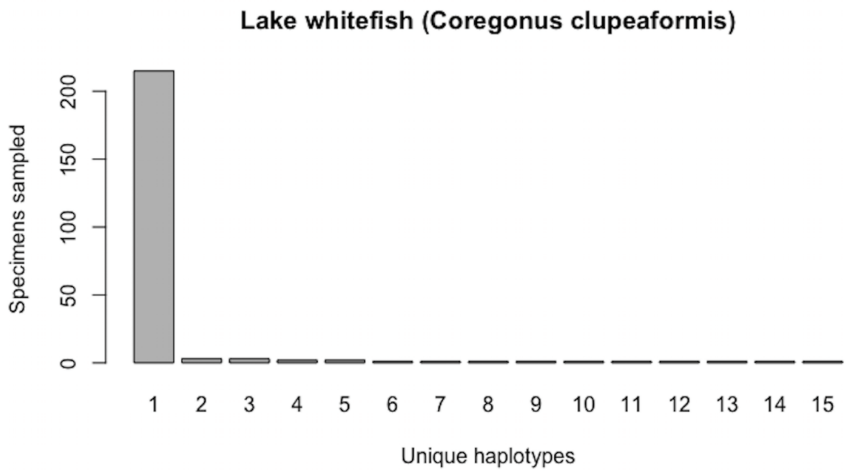

Figure 3.8: Initial haplotype frequency distribution for N = 235 high-quality lake whitefish (*Coregonus clupeaformis*) COI barcode sequences obtained from BOLD. This species displays a highly-skewed pattern of observed haplotype variation, with Haplotype 1 accounting for *c.* 91.5% (215/235) of all sampled records.

The output of `HACSim` is displayed below.

```
### Run simulations ###
> HACSObj <- HACReal()
> HAC.simrep(HACSObj)
```

```
Simulating haplotype accumulation...

|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  11.0705
Mean number of haplotypes not sampled:  3.9295
Proportion of haplotypes sampled:  0.7380333
Proportion of haplotypes not sampled:  0.2619667

Mean value of N*:  318.4138
Mean number of specimens not sampled:  83.4138

Desired level of haplotype recovery has not yet been reached

|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  13.8705
Mean number of haplotypes not sampled:  1.1295
Proportion of haplotypes sampled:  0.9247
Proportion of haplotypes not sampled:  0.0753

Mean value of N*:  603.439
Mean number of specimens not sampled:  45.43895

Desired level of haplotype recovery has not yet been reached

|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  14.3708
Mean number of haplotypes not sampled:  0.6292
Proportion of haplotypes sampled:  0.9580533
Proportion of haplotypes not sampled:  0.04194667

Mean value of N*:  630.4451
Mean number of specimens not sampled:  26.44507

Desired level of haplotype recovery has been reached

---------- Finished.  ----------
The initial guess for sampling sufficiency was N = 235 individuals
The algorithm converged after 8 iterations and took 241.235 s
The estimate of sampling sufficiency for p = 95% haplotype recovery is N* = 604
individuals ( 95% CI: 601.504-606.496 )


The number of additional specimens required to be sampled for p = 95% haplotype recovery


is N* - N = 369 individuals
```

From the above output, it is clear that current specimen sample sizes found within BOLD for *C. clupeaformis* are probably not sufficient to capture the majority of within-species COI haplotype variation. An initial sample size of N = 235 specimens corresponds to recovering only 73.8% of all H^* = 15 unique haplotypes for this species

(**Fig. 3.9**).

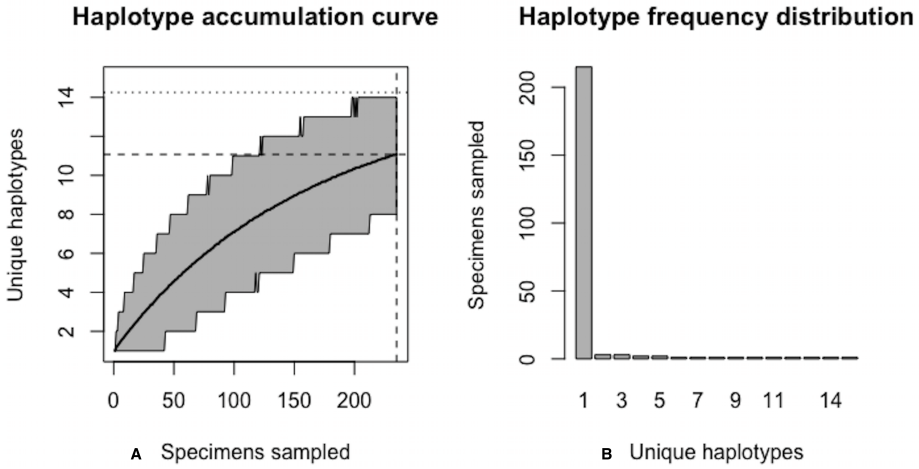

Figure 3.9: Initial graphical output of `HAC.sim()` for a real species (Lake whitefish, *C. clupeaformis*) having a single dominant haplotype. In this example, initially, only R = 73.8% of the H^* = 15 estimated haplotypes for this species have been recovered based on a sample size of N = 235 specimens. The haplotype frequency barplot is identical to that of **Fig. 8**.

A sample size of N^* = 604 individuals (95% CI: 601.504-606.496) would likely be needed to observe 95% of all existing genetic diversity for lake whitefish (**Fig. 3.10**).

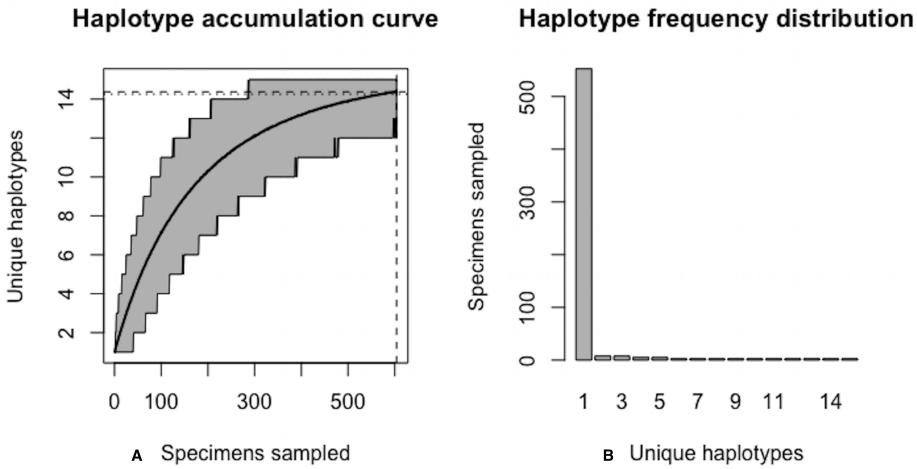

Figure 3.10: Final graphical output of `HAC.sim()` for Lake whitefish (*C. clupeaformis*) having a single dominant haplotype. Upon convergence, R = 95.8% of the H^* = 15 estimated haplotypes for this species have been uncovered with a sample size of N = 604 specimens.

Since N = 235 individuals have been sampled previously, only $N^* - N$ = 369 specimens remain to be collected.

Deer tick (*Ixodes scapularis*)

Ticks, particularly the hard-bodied ticks (Arachnida: Acari: Ixodida: Ixodidae), are well-known as vectors of various zoonotic diseases including Lyme disease [130]. Apart from this defining characteristic, the morpohological identification of ticks at any lifestage, by even expert taxonomists, is notoriously difficult or sometimes even impossible [131]. Further, the presence of likely high cryptic species diversity in this group means that turning to molecular techniques such as DNA barcoding is often the only feasible option for reliable species diagnosis. Lyme-competent specimens can be accurately detected through employing a sensitive quantitative PCR (qPCR) procedure [131]. However, for such a

workflow to be successful, wide coverage of within-species haplotype variation from across broad geographic ranges is paramount to better aid design of primer and probe sets for rapid species discrimination. Furthermore, the availability of large specimen sample sizes for tick species of medical and epidemiological relevance is necessary for accurately assessing the presence of the barcode gap.

Notably, the deer tick (*Ixodes scapularis*), native to Canada and the United States, is the primary carrier of *Borrelia burgdorferi*, the bacterium responsible for causing Lyme disease in humans in these regions. Because of this, *I. scapularis* has been the subject of intensive taxonomic study in recent years. For instance, in a recent DNA barcoding study of medically-relevant Canadian ticks, [131] found that out of eight specimens assessed for the presence of *B. burgdorferi*, 50% tested positive. However, as only exoskeletons and a single leg were examined for systemic infection, the reported infection rate may be a lower bound due to the fact that examined specimens may still harbour *B. burgdorferi* in their gut. As such, this species is well-represented within BOLD and thus warrants further examination within the present study.

As of August 27, 2019, 531 5'-COI DNA barcode sequences are accessble from BOLD's Public Data Portal for this species. Of these, N = 349 met criteria for high quality outlined in Phillips *et al.* [144]. A 658 bp MUSCLE alignment comprised H^* = 83 unique haplotypes. Haplotype analysis revealed that Haplotypes 1-8 were represented by more than 10 specimens (range: 12-46; **Fig. 3.11**).

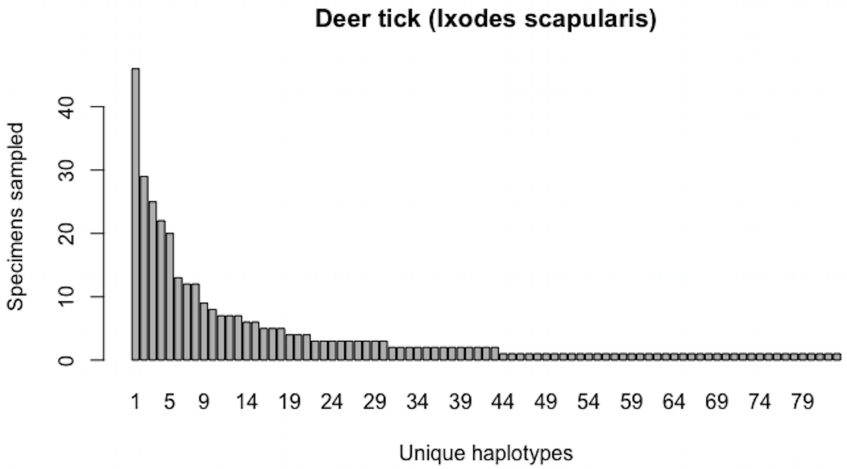

Figure 3.11: Initial haplotype frequency distribution for N = 349 high-quality deer tick (*Ixodes scapularis*) COI barcode sequences obtained from BOLD. In this species, Haplotypes 1-8 account for *c.* 51.3% (179/349) of all sampled records.

Simulation output of `HACSim` is depicted below.

```
### Run simulations ###
> HACSObj <- HACReal()
> HAC.simrep(HACSObj)
Simulating haplotype accumulation...
|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  65.3514
Mean number of haplotypes not sampled:  17.6486
Proportion of haplotypes sampled:  0.7873663
Proportion of haplotypes not sampled:  0.2126337

Mean value of N*:  443.2499
Mean number of specimens not sampled:  94.24988

Desired level of haplotype recovery has not yet been reached
|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  78.3713
Mean number of haplotypes not sampled:  4.6287
Proportion of haplotypes sampled:  0.9442325
Proportion of haplotypes not sampled:  0.05576747

Mean value of N*:  802.7684
Mean number of specimens not sampled:  44.76836

Desired level of haplotype recovery has not yet been reached
|================================================================================| 100%
```

```
--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  79.2147
Mean number of haplotypes not sampled:  3.7853
Proportion of haplotypes sampled:  0.954394
Proportion of haplotypes not sampled:  0.04560602

Mean value of N*:  841.3716
Mean number of specimens not sampled:  38.37161

Desired level of haplotype recovery has been reached

---------- Finished.  ----------
The initial guess for sampling sufficiency was N = 349 individuals
The algorithm converged after 8 iterations and took 1116.468 s
The estimate of sampling sufficiency for p = 95% haplotype recovery is N* = 803
individuals ( 95% CI: 801.551-804.449 )


The number of additional specimens required to be sampled for p = 95% haplotype recovery


is N* - N = 454 individuals
```

The above results clearly demonstrate the need for increased specimen sample sizes in deer ticks. With an initial sample size of N = 349 individuals, only 78.7% of all observed haplotypes are recovered for this species (**Fig. 3.12**).

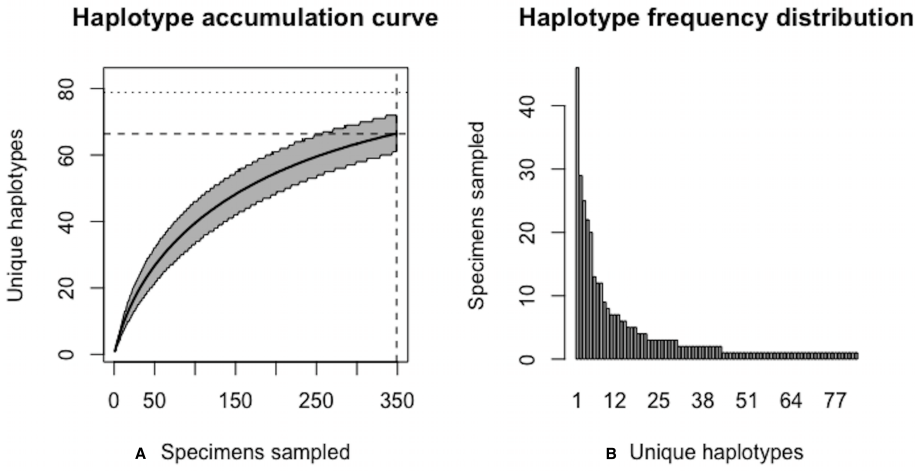

Figure 3.12: Initial graphical output of `HAC.sim()` for a real species (Deer tick, *I. scapularis*) having eight dominant haplotypes. In this example, initially, only R = 78.7% of the H^* = 83 estimated haplotypes for this species have been recovered based on a sample size of N = 349 specimens. The haplotype frequency barplot is identical to that of **Fig. 3.11**.

N^* = 803 specimens (95% CI: 801.551-804.449) is necessary to capture at least 95% of standing haplotype variation for *I. scapularis* (**Fig. 3.13**) .

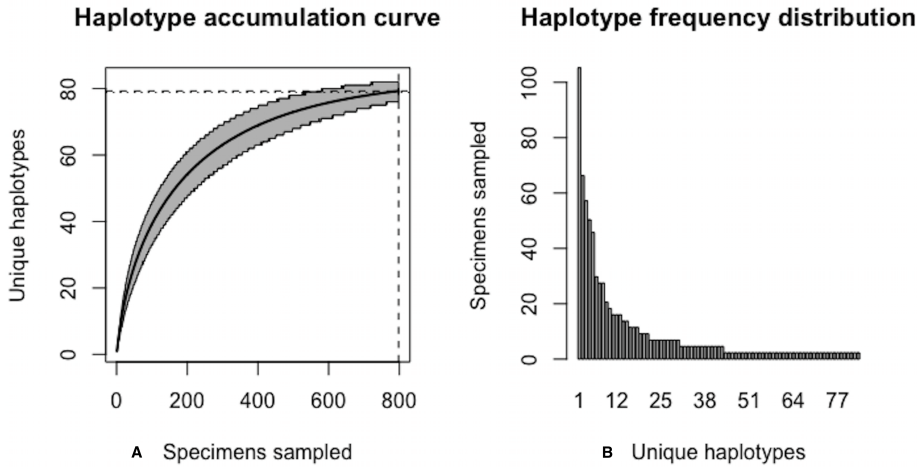

Figure 3.13: Final graphical output of `HAC.sim()` for deer tick (*I scapularis*) having eight dominant haplotypes. Upon convergence, R = 95.4% of the H^* = 83 estimated haplotypes for this species have been uncovered with a sample size of N = 803 specimens.

Thus, a further $N^* - N$ = 454 specimens are required to be collected.

Scalloped hammerhead (*Sphyrna lewini*)

Sharks (Chondrichthyes: Elasmobranchii: Selachimorpha) represent one of the most ancient extant lineages of fishes. Despite this, many shark species face immediate extinction as a result of overexploitation, together with a unique life history (*e.g.*, K-selected, predominant viviparity, long gestation period, lengthy time to maturation) and migration behaviour [71]. A large part of the problem stems from the increasing consumer demand for, and illegal trade of, shark fins, meat and bycatch on the Asian market. The widespread, albeit lucrative practice of "finning", whereby live sharks are definned and immediately released, has led to the rapid decline of once stable populations

[169]. As a result, numerous shark species are currently listed by the International Union for the Conservation of Nature (IUCN) and the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). Interest in the molecular identification of sharks through DNA barcoding is multifold. The COI reference sequence library for this group remains largely incomplete. Further, many shark species exhibit high intraspecific distances within their barcodes, suggesting the possibility of cryptic species diversity. Instances of hybridization between sympatric species has also been documented. As establishing species-level matches to partial specimens through morphology alone is difficult, and such a task becomes impossible once fins are processed and sold for consumption or use in traditional medicine, DNA barcoding has paved a clear path forward for unequivocal diagnosis in most cases.

The endangered hammerheads (Family: Sphyrnidae) represent one of the most well-sampled groups of sharks within BOLD to date. Fins of the scalloped hammerhead (*Sphyrna lewini*) are especially highly prized within IUU (Illegal, Unregulated, Unreported) fisheries due to their inclusion as the main ingredient in shark fin soup.

As of August 27, 2019, 327 *S. lewini* specimens (sequenced at both barcode and non-barcode markers), collected from several Food and Agriculture Organization (FAO) regions, including the United States, are available through BOLD's Public Data Portal. Of these, all high-quality records (N = 171) were selected for alignment in MEGA7 and assessment via `HACSim`. The final alignment was found to comprise H^* = 12 unique haplotypes, of which three were represented by 20 or more specimens (range: 28-70; **Fig.**

3.14).

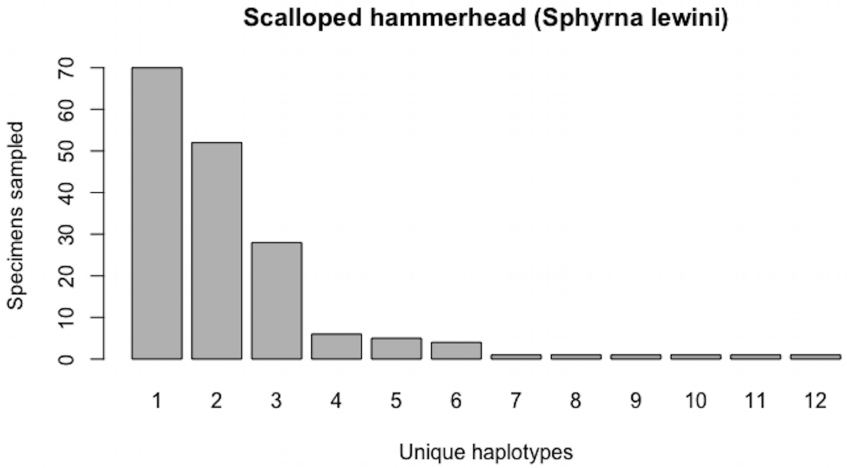

Figure 3.14: Initial haplotype frequency distribution for N = 171 high-quality scalloped hammerhead (*Sphyrna lewini*) COI barcode sequences obtained from BOLD. In this species, Haplotypes 1-3 account for *c.* 87.7% (150/171) of all sampled records.

`HACSim` results are displayed below.

```
### Run simulations ###
> HACSObj <- HACReal()
> HAC.simrep(HACSObj)
Simulating haplotype accumulation...
|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  9.9099
Mean number of haplotypes not sampled:  2.0901
Proportion of haplotypes sampled:  0.825825
Proportion of haplotypes not sampled:  0.174175

Mean value of N*:  207.0657
Mean number of specimens not sampled:  36.06566

Desired level of haplotype recovery has not yet been reached
|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  11.3231
Mean number of haplotypes not sampled:  0.6769
Proportion of haplotypes sampled:  0.9435917
Proportion of haplotypes not sampled:  0.05640833

Mean value of N*:  413.3144
Mean number of specimens not sampled:  23.31438

Desired level of haplotype recovery has not yet been reached
```

```
|================================================================================| 100%

--- Measures of Sampling Closeness ---

Mean number of haplotypes sampled:  11.4769
Mean number of haplotypes not sampled:  0.5231
Proportion of haplotypes sampled:  0.9564083
Proportion of haplotypes not sampled:  0.04359167

Mean value of N*:  432.8695
Mean number of specimens not sampled:  18.8695

Desired level of haplotype recovery has been reached

---------- Finished.  ----------
The initial guess for sampling sufficiency was N = 171 individuals
The algorithm converged after 9 iterations and took 174.215 s
The estimate of sampling sufficiency for p = 95% haplotype recovery is N* = 414
individuals ( 95% CI: 411.937-416.063 )


The number of additional specimens required to be sampled for p = 95% haplotype recovery


is N* - N = 243 individuals
```

Simulation output suggests that only 82.6% of all unique haplotypes for the scalloped hammerhead have likely been recovered (**Fig. 15**) with a sample size of N = 171.

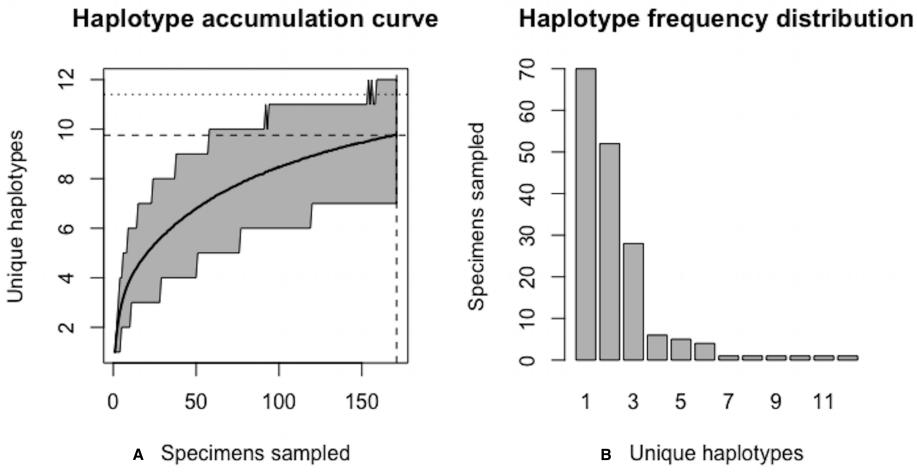

Figure 3.15: Initial graphical output of `HAC.sim()` for a real species (Scalloped hammerhead, *S. lewini*) having three dominant haplotypes. In this example, initially, only R = 82.6% of the H^* = 12 estimated haplotypes for this species have been recovered based on a sample size of N = 171 specimens. The haplotype frequency barplot is identical to that of **Fig. 3.14**.

Further, `HACSim` predicts that N^* = 414 individuals (95% CI: 411.937-416.063) probably need to be randomly sampled to capture the majority of intraspecific genetic diversity within 5'-COI (**Fig. 3.16**).

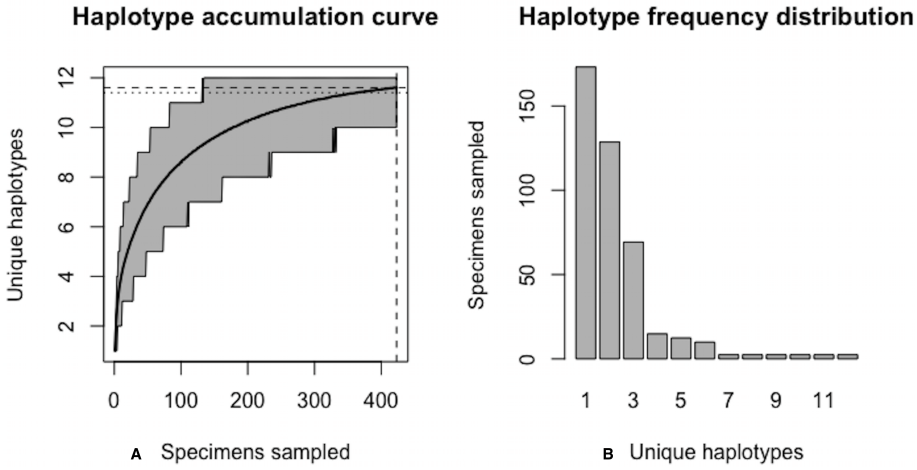

Figure 3.16: Final graphical output of `HAC.sim()` for scalloped hammerhead (*S. lewini*) having three dominant haplotypes. Upon convergence, $R = 95.6\%$ of the $H^* = 12$ estimated haplotypes for this species have been uncovered with a sample size of $N = 414$ specimens.

Since 171 specimens have already been collected, this leaves an additional $N^* - N = 243$ individuals which await sampling.

3.4 Discussion

3.4.1 Initializing `HACSim` and Overall Algorithm Behaviour

The overall stochastic behaviour of `HACSim` is highly dependent on the number of permutations used upon algorithm initialization. Provided that the value assigned to the `perms` argument is set high enough, numerical results ouputted by `HACSim` will be found to be quite consistent between consecutive runs whenever all remaining parameter values

remain unchanged. It is crucial that `perms` not be set to too low a value to prevent the algorithm from getting stuck at local maxima and returning suboptimal solutions. This is a common situation with popular optimization algorithms such as hill-climbing. Attention therfore must be paid to avoid making generalizations based on algorithm performance and obtained simulation results [165].

In applying the present method to simulated species data, it is important that selected simulation parameters are adequately reflective of those observed for real species. Thus, initial sample sizes should be chosen to cover a wide range of values based on those currently observed within BOLD. Such information can be gauged through examining species lists associated with BOLD records, which are readily accessible through Linnean search queries within the Taxonomy browser.

As with any iterative numerical algorithm, selecting good starting guesses for initialization is key. While `HACSim` is globally convergent (*i.e.*, convergence is guaranteed for any value of $N \geq H^*$), a good strategy when simulating hypothetical species is to start the algorithm by setting $N = H^*$. In this way, the observed fraction of haplotypes found, R, will not exceed the desired level of haplotype recovery p, and therefore lead to overestimation of likely required specimen sample sizes. Setting N high enough will almost surely result in R exceeding p. Thus, arbitrarily large values of N may not be biologically meaningful or practical. However, in the case of hypothetical species simulation, should initial sample sizes be set too high, such that $R > p$, a straightforward workaround is to observe where the dashed horizontal line intersects the final haplotype

accumulation curve (*i.e.*, not the line the touches the curve endpoint). The resulting value of N at this point will correspond with p quite closely. This can be seen in **Fig. 5**, where an eyeball guess just over $N^* = 20$ individuals is necessary to recover $p = 95\%$ haplotype variation. A more reliable estimate can be obtained through examining the dataframe "d" outputted once the algorithm has halted (via `envr$d`). In this situation, simply look in the row corresponding to $pH^* \geq 0.95(10) \geq 9.5$. The required sample size is the value given in the first column (`specs`). This is accomplished via the R code

```
envr$d[which(envr$d$means >= envr$p * envr$Hstar), ][1, 1].
```

The novelty of `HACSim` is that it offers a systematic means of estimating likely specimen sample sizes required to assess intraspecific haplotype diversity for taxa within large-scale genomic databases like GenBank and BOLD. Estimates of sufficient sampling suggested by our algorithm can be employed to assess barcode coverage within existing reference sequence libraries and campaign projects found in BOLD. While comparison of our method to already-established ones is not yet possible, we anticipate that `HACSim` will nevertheless provide regulatory applications with an unprecedented view and greater understanding of the state of standing genetic diversity (or lack thereof) within species.

3.4.2 Additional Capabilities and Extending Functionality of `HACSim`

In this paper, we illustrate the application of haplotype accumulation curves to the broad assessment of species-level genetic variation. However, `HACSim` is quite flexible in that one can easily explore likely required sample sizes at higher taxonomic levels (*e.g.* order, family, genus) or specific geographic regions (*e.g.*, salmonids of the Great Lakes)

with ease. Such applicability will undoubtedly be of interest at larger scales (*i.e.*. entire genomic sequence libraries). For instance, due to evidence of sampling bias in otherwise densely-sampled taxa housed in BOLD (*e.g.*, Lepidoptera), D'Ercole *et al.* (J. D'Ercole, 2019, unpublished data) wished to assess whether or not intraspecific haplotype variation within butterfly species remains unsampled. To test this, the authors employed `HACSim` to examine sampling comprehensiveness for species comprising a large barcode reference library for North American butterflies spanning 814 species and 14623 specimens.

We foresee use of `HACSim` being widespread within the DNA barcoding community. As such, improvements to existing code in terms of further optimization and algorithm runtime, as well as implementation of new methods by experienced R programmers in the space of DNA-based taxonomic identification, seems bright.

Potential extensions of our algorithm include support for the exploration of genetic variation at the Barcode Index Number (BIN) level [151], as well as high-throughput sequencing (HTS) data for metabarcoding and environmental DNA (eDNA) applications. Such capabilities are likely to be challenging to implement at this stage until robust operational taxonomic unit (OTU) clustering algorithms are developed (preferably in R). One promising tool in this regard for barcoding of bulk samples of real species and mock communities of known species composition is JAMP (**J**ust **A**nother **M**etabarcoding **P**ipeline) devised for use in R by Elbrecht and colleagues [49]. JAMP includes a sequence read denoising tool that can be used to obtain haplotype numbers and frequency information (*H** and `probs`). However, because JAMP relies on third-party software (particularly

USEARCH [46] and VSEARCH [152]), it cannot be integrated within `HACSim` itself and will thus have to be used externally. In extending `HACSim` to next-generation space, two issues arise. First, it is not immediately clear how the argument N, is to be handled since multiple reads could be associated with single individuals. That is, unlike in traditional Sanger-based sequencing, there is not a one-to-one correspondence between specimen and sequence [2, 185]. Second, obtaining reliable haplotype information from noisy HTS datasets is challenging without first having strict quality filtering criteria in place to minimize the occurrence of rare, low-copy sequence variants which may reflect artifacts stemming from the Polymerase Chain Reaction (PCR) amplification step or sequencing process [19, 49, 182]. Turning to molecular population genetics theory might be the answer [2]. Wares and Pappalardo [185] suggest three different approaches to estimating the number of specimens of a species that may have contributed to a metabarcoding sample: (1) use of prior estimates of haplotype diversity, together with observed number of haplotypes; (2) usage of Ewens' sampling formula [50] along with estimates of Watterson's θ (not to be confused with the θ denoting true sampling sufficency herein) [187], as well as total number of sampled haplotypes; and (3) employment of an estimate of θ and the number of observed variable sites (S) within a multiple sequence alignment. A direct solution we propose might be to use sequencing coverage/depth (*i.e.*, the number of sequence reads) as a proxy for number of individuals. The outcome of this would be an estimate of the mean/total number of sequence reads required for maximal haplotype recovery. However, the use of read count as a stand-in for number of specimens sampled would require the

unrealistic assumption that all individuals (*i.e.*, both alive and dead) shed DNA into their environment at equal rates. The obvious issue with extending `HACSim` to handle HTS data is computing power, as such data typically consists of millions of reads spanning multiple gigabytes of computer memory.

3.4.3 Summary

Here, we introduced a new statistical approach to assess specimen sampling depth within species based on existing gene marker variation found in public sequence databanks such as BOLD and GenBank. `HACSim` is both computationally efficient and easy to use. We show utility of our proposed algorithm through both hypothetical and real species genomic sequence data. For real species (here, lake whitefish, deer tick and scalloped hammerhead), results from `HACSim` suggest that comprehensive sampling for species comprising large barcode libraries within BOLD, such as Actinopterygii, Arachnida and Elasmobranchii is far from complete. With the availability of `HACSim`, appropriate sampling guidelines based on the amount of potential error one is willing to tolerate can now be established. For the purpose of addressing basic questions in biodiversity science, the employment of small taxon sample sizes may be adequate; however, this is not the case for regulatory applications, where greater than 95% coverage of intraspecific haplotype variation is needed to provide high confidence in sequence matches defensible in a court of law.

Of immediate interest is the application of our method to other ray-finned fishes, as well as other species from deeply inventoried taxonomic groups such as Elasmobranchii (*e.g.*

sharks), Insecta (*e.g.* Lepidoptera, Culicidae (mosquitoes)), Arachnida (*e.g.*, ticks) and Chiroptera (bats) that are of high conservation, medical and/or socioeconomic importance. Although we explicitly demonstrate the use of `HACSim` through employing COI, it would be interesting to extend usage to other barcode markers such as the ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) and maturase K (matK) chloroplast genes for land plants, as well as the nuclear internal transcribed spacer (ITS) marker regions for fungi. The application of our method to non-barcode genes routinely employed in specimen identification like mitochondrial cytochrome *b* (cyt*b*) in birds for instance [11, 101], nuclear rhodopsin (rho) for marine fishes [70] or the phosphoenolpyruvate carboxykinase (PEPCK) nuclear gene for bumblebees [191] is also likely to yield interesting results since sequencing numerous individuals at several different genomic markers can often reveal evolutionary patterns not otherwise seen from employing a single-gene approach (*e.g.*, resolution of cryptic species or confirmation/revision of established taxonomic placements) [191].

While it is reasonable that `HACSim` can be applied to genomic regions besides 5'-COI, careful consideration of varying rates of molecular evolution within rapidly-evolving gene markers and the effect on downsteam inferences is paramount, as is sequence quality. Previous work in plants (Genus: *Taxus*) by Liu *et al.* [107] has found evidence of a correlation between mutation rate and required specimen sampling depth: genes evolving at faster rates will likely require larger sample sizes to estimate haplotype diversity compared to slowly-evolving genomic loci. We simply focused on 5'-COI because it is by far the most

widely sequenced mitochondrial locus for specimen identification, owing to its desirable biological properties as a DNA barcode for animal taxa and because it has an associated data standard to help filter out poor-quality data. [142]. However, it should be noted that species diagnosis using COI and other barcode markers is not without its challenges. While COI accumulates variation at an appreciable rate, certain taxonomic groups are not readily distinguished on the basis of their DNA barcodes (*e.g.*, the so-called "problem children", such as Cnidaria, which tend to lack adequate sequence divergence [21]). Other taxa, like Mollusca, are known to harbour indel mutations [102]. Introns within Fungi greatly complicate sequence alignment [119]. Thus, users of `HACSim` must exercise caution in interpreting end results with other markers, particularly those which are not protein-coding.

It is necessary to consider the importance of sampling sufficiency as it pertains to the myriad regulatory applications of specimen identification established using DNA barcoding (*e.g.*, combatting food fraud) in recent years. It since has become apparent that the success of such endeavours is complicated by the ever-evolving state of public reference sequence libraries such as those found within BOLD, in addition to the the inclusion of questionable sequences and lack of sufficent metadata for validation purposes in other genomic databases like GenBank (*e.g.*, [72]). Dynamic DNA-based identification systems may produce multiple conflicting hits to otherwise corresponding submissions over time. This unwanted behaviour has led to a number of regulatory agencies creating their own *static* repositories populated with expertly-identified sequence records tied to known voucher specimens deemed fit-for-purpose for molecular species diagnosis and forensic

compliance (*e.g.* the United States Food and Drug Administration (USFDA)'s Reference Standard Sequence Library (RSSL) employed to identify unknown seafood samples from species of high socioeconomic value). While such a move has partially solved the problem of dynamism inherent in global sequence databases, there still remains the issue of low sample sizes that can greatly inflate the perception of barcode gaps between species. Obtaining adequate representation of standing genetic variation, both within and between species, is therefore essential to mitigating false assignments using DNA barcodes. To this end, we propose the use of `HACSim` to assess the degree of saturation of haplotype accumulation curves to aid regulatory scientists in rapidly and reliably projecting likely sufficient specimen sample sizes required for accurate matching of unknown queries to known Linnean names.

A defining characteristic of `HACSim` is its convergence behaviour: the method converges to the desired level of haplotype recovery p for any initial guess N specified by the user. Based on examples explored herein, it appears likely that already-sampled species within repositories like BOLD are far from being fully characterized on the basis of existing haplotype variation. In addition to this, it is important to consider the current limitations of our algorithm. We can think of only one: it must be stressed that appropriate sample size trajectories are not possible for species with only single representatives within public DNA sequence databases because haplotype accumulation is unachievable with only one DNA sequence and/or a single sampled haplotype. Hence, `HACSim` can only be applied to species with at least two sampled specimens. Thus, application of our method to assess

necessary sample sizes for full capture of extant haplotype variation in exceedingly rare or highly elusive taxa is not feasible. Despite this, we feel that `HACSim` can greatly aid in accurate and rapid barcode library construction necessary to thoroughly appreciate the diversity of life on Earth.

3.5 Conclusions

Herein, a new, easy-to-use R package was presented that can be employed to estimate intraspecific sample sizes for studies of genetic diversity assessment, with a particular focus on animal DNA barcoding using the COI gene. `HACSim` employs a novel nonparametric stochastic iterative extrapolation algorithm with good convergence properties to generate haplotype accumulation curves. Because our approach treats species' haplotypes as numeric labels, any genomic locus can be targeted to probe levels of standing genetic variation within multicellular taxa. However, we stress that users must exercise care when dealing with sequence data from non-coding regions of the genome, since these are likely to comprise sequence artifacts such as indels and introns, which can both hinder successful sequence alignment and lead to overestimation of existing haplotype variation within species. The application of our method to assess likely required sample sizes for both hypothetical and real species produced promising results. We argue the use of `HACSim` will be of broad interest in both academic and industry settings, most notably, regulatory agencies such as the Canadian Food Inspection Agency (CFIA), Agriculture and Agri-Food Canada (AAFC), United States Department of Agriculture (USDA), Public Health Agency

of Canada (PHAC) and the USFDA. While `HACSim` is an ideal tool for the analysis of Sanger sequencing reads, an obvious next step is to extend usability to Next-Generation Sequencing (NGS), especially HTS applications. With these elements in place, even the full integration of `HACSim` to assess comprehensiveness of taxon sampling within large sequence databases such as BOLD seems like a reality in the near future.

Acknowledgments

We wish to greatly acknowledge the efforts of Rodger Gwiazdowski in providing valuable edits to this manuscript. In addition, comments by Sarah (Sally) Adamowicz improved overall readablity and flow of the manuscript considerably.

This work was supported by a University of Guelph College of Physical and Engineering Science (CPES) Graduate Excellence Entrance Scholarship awarded to JDP.

The Dish With One Spoon Covenant speaks to our collective responsibility to steward and sustain the land and environment in which we live and work, so that all peoples, present and future, may benefit from the sustenance it provides. As we continue to strive to strengthen our relationships with and continue to learn from our Indigenous neighbours, we recognize the partnerships and knowledge that have guided the research conducted in our labs. We acknowledge that the University of Guelph resides in the ancestral and treaty lands of several Indigenous peoples, including the Attawandaron people and the Mississaugas of the Credit, and we recognize and honour our Anishinaabe, Haudenosaunee, and Métis neighbours. We acknowledge that the work presented here has occurred on their traditional lands so that we might work to build lasting partnerships that respect, honour, and value the culture, traditions, and wisdom of those who have lived here since time immemorial.

Author Contributions

JDP conducted the literature review and wrote the manuscript. DJG acted as an advisor in statistics. RHH acted as an advisor in DNA barcoding. All authors contributed to the

revision of this manuscript and approved the final version.

Conflict of Interest

None declared.

Chapter 4

Lack of statistical rigor in DNA barcoding likely invalidates the presence of a true species' barcode gap

Jarrett D. Phillips[1*], Daniel J. Gillis[1], Robert H. Hanner[2,3]

[1]*School of Computer Science, University of Guelph, Guelph, ON., Canada, N1G2W1*

[2]*Biodiversity Institute of Ontario, University of Guelph, Guelph, ON., Canada, N1G2W1*

[3]*Department of Integrative Biology, University of Guelph, Guelph, ON., Canada, N1G2W1*

Submitted to *Systematic Biology*

ABSTRACT

When DNA barcoding was first conceived at the start of the 21st century as a means to rapidly characterize Earth's dwindling biodiversity faster than traditional taxonomy ever could over the last 250 years, the proposal was met with both high praise and stark criticism. Within the first few years of barcoding's introduction, what followed was the polarization of the biodiversity community-at-large, the sparking of heated discussion among researchers, and the incitement of widespread debate amidst established schools of biological thought. Flashing forward almost two decades later, DNA barcoding has proved itself to be both fully capable of rising to the challenge and highly resilient to change. However, the story does not end here. As reference sequence libraries continue to grow exponentially in size, there is now the need to identify novel ways of meaningfully analyzing vast amounts of available DNA barcode data.

Here, it is demonstrated that the interpretation of DNA barcoding data is lacking in statistical rigor. To highlight this, focus is set specifically on one key concept that has become a household name in the field: the DNA barcode gap. Arguments outlined herein stem from three angles: (1) the improper allocation of specimen sampling effort necessary to capture adequate levels of within-species genetic variation, (2) failing to properly visualize intraspecific and interspecific genetic distances, and (3) the inconsistent, inappropriate use, or absence of statistical inferential procedures in DNA barcoding gap analyses. Furthermore, simple statistical solutions are outlined which can greatly propel the use of DNA barcoding as a tool to irrefutably match unknowns to knowns on the basis

of the barcoding gap with a high degree of confidence. Proposed methods examined herein are illustrated through application to DNA barcode sequence data from Canadian Pacific fish species as a case study.

4.1 Introduction

4.1.1 DNA Barcoding: A Brief Tour

In its infancy, DNA barcoding [75] was envisaged as a means to resolve a longstanding problem facing biodiversity science: the taxonomic impediment. Accelerating the description of novel taxa, as well as revising the status of existing ones, through the assembly of genetic "signatures" within a centralized repository more rapidly than customary Linnean classification was even capable, seemed, at first, like wishful thinking within some academic circles, in a time marked by global species extinction and ongoing environmental crisis. DNA barcoding employs short molecular sequence tags from standardized genomic regions, such as the *c.* 650 bp fragment from the 5' end of the mitochondrial cytochrome *c* oxidase subunit I (COI) gene in animals, to establish taxon-level matches to unknown specimen queries at any life stage across the Eukaryotic Tree of Life [77]. The isolation of a barcode sequence from mitochondrial DNA (mtDNA), as opposed to its nuclear counterpart, is appealing due to mtDNA's high copy number given its haploid structure, its low rate of homologous recombination, and its uniparental (maternal) mode of inheritance. The specific choice of COI as the currently unattested DNA barcode for animals is justified in several respects: (1) that it is protein-coding and plays a central role in oxidative phosphorylation, (2) that it possesses a reasonably high rate of nucleotide substitution, and (3) that it lacks introns and comprises few insertions/deletions (indels) and no stop codons. In addition to these desirable characteristics, its ease of amplification, sequencing and alignment across most taxa, due to its highly conserved

nature, makes COI the preferred gene marker over other such loci that meet only some of the abovementioned requirements [142]. Despite this, the attractiveness of DNA barcoding's use as both a specimen identification and species discovery tool is fraught with much controversy as its success rests crucially on extant species-level haplotype diversity and the distinction between intraspecific and interspecific genetic variation across taxa [142] that readily explains observable biogeographic, demographic, geneologic, phylogenetic and phylogeographic patterns.

While DNA barcoding has found myriad applications in diverse subdisciplines of evolutionary biology, ecology, and more broadly in biodiversity science, one surprising area, namely applied regulatory forensics, has reaped the benefits barcoding has to offer in unparalleled ways throughout the years. The identification of regulated species of socioeconomic importance through the accumulation of DNA barcodes has been instrumental in combatting instances of seafood market fraud as well as monitoring the introduction and spread of invasive pests, particularly in Canada and the USA (*e.g.*, seafood: [**?**, 160]; meat products: [124, 161]; invasive arthropods: [113]). Despite this, several obstacles still remain. The inherent dynamism characteristic of public genomic databases, such as the Barcode of Life Data Systems (BOLD; [150]; http://v4.boldsystems.org) and GenBank, precludes their routine use for such a task. The fact that the addition of new specimen records to community databanks may produce contradictory findings over time is problematic [143]. Instead, regulatory sequence databases should be populated with static taxon records traced to voucher specimens

whenever possible so that such issues can be mitigated. While the inclusion of fit-for-purpose DNA sequences in governmental repositories like the European and Mediterranean Plant Protection Organization (EPPO)'s Q-bank (https://qbank.eppo.int) for agricultural/quarantine pests and the United States Food and Drug Administration (USFDA)'s Reference Standard Sequence Library for Seafood Identification (RSSL; https://www.fda.gov/food/dna-based-seafood-identification/reference-standard-sequence-library-seafood-identification-rssl) for seafood species represents a step in the right direction, sample size issues continue to plague the arena [143]. Further, the deep sampling of an adequate number of specimens necessary to capture sufficient levels of standing haplotype variation within species is critical if high confidence in specimen assignments is desired [42, 142, 143, 144]. DNA-based identification accomplished through DNA barcoding places heavy reliance on the accuracy and completeness of reference sequence libraries to enable the rapid assignment of unknown specimens to valid or putative species, depending on whether the ultimate goal is specimen identification or species discovery respectively. As distance-based methods strongly outweigh other identification approaches (*e.g.,* tree-based algorithms) within most DNA barcoding studies, a means of directly testing the overall performance of DNA barcoding is needed. Such a path forward is provided by the DNA barcode gap.

4.1.2 DNA Barcoding and the Barcode Gap: A Perfect Harmony?

A well-established tenet in the field is that the majority of DNA barcode variation found across species exceeds genetic variability seen within species. This apparent "barcode gap"

[118] was recognized early on as a critical factor to the success of DNA barcoding as a discipline transcending modern biodiversity science at a fundamental level. The existence of a species' barcoding gap is often invoked as evidence that DNA barcoding "works" in practice [174]. Under current sampling efforts and morphological identifications associated with DNA barcodes, a large number of species show greater than 2% genetic distance to their nearest heterospecific and typically exhibit less than 1% intraspecific distance [85].

Whereas Meyer *et al.* [118] advocated the use of the mean genetic distance to the nearest neighbour, employment of the minimum interspecific distance is now commonplace. Reliance on the former metric tends to exaggerate the presence of a real species barcoding gap through inflating false positives, leading to misidentification of specimens [117]. Thus, on the basis of this paradigm shift, it is wholly conceivable that published DNA barcoding studies have likely reported biased estimates of the barcoding gap at the species level and therefore warrant revisitation and cautious interpretation. Perhaps this is why the barcoding gap is depicted using both the mean and maximum interspecific genetic distance within BOLD's Barcode Gap Analysis tool available in the user Workbench. Further still, Meyer and Paulay [118] differentiate between two variants of species barcoding gaps, depending on whether specimen identification or species discovery is the end goal: 'local' and 'global' respectively. A 'local' DNA barcode gap can be applied whenever an individual specimen of a particular species is closer in distance to another member of the same species; whereas, a 'global' barcoding gap is applicable whenever a threshold can be identified that separates all species [35]. While each of

these hold great importance for the identification of unknown specimens, the absence of a sufficiently wide 'global' barcode gap that readily distinguishes higher-level taxonomic diversity (*e.g.*, phyla) does not immediately rule out the existence and usefulness of 'local' gaps at lower levels of taxonomic organization (*i.e.*, genus, species) [98, 100].

Many studies use a barcoding gap approach as a reliable species, genus, or higher-level (*i.e.*, order, family) separation criterion with little discussion as to its overall utility. However, both the existence and application of the DNA barcode gap are equally important: taken together, they reconcile both morphological identifications established through Linnean taxonomy with molecular identifications based on DNA sequence variation found across all multicellular taxa. Recognizing this, Collins and Cruickshank [35]'s outline of "the seven sins of DNA barcoding" touched briefly on "inappropriate use of fixed distance thresholds" and "incorrectly interpreting the barcoding gap" as the sixth and seventh deadly sins, respectively. Despite strong support early-on in the DNA barcoding enterprise for the existence of the DNA barcode gap, subsequent studies have since gone on to suggest that the presence of a barcode gap at any taxonomic level is simply an artifact of insufficient specimen sampling across narrow geographic and morphologic space [23, 37]. In light of this observation, the interpretation of the barcode gap across taxa is not a straightforward task (*e.g.*, marine gastropods molluscs: [118], butterflies: [189], spiders: [23], annelids: [100], leaf-footed bugs: [201], dragonflies/damselflies: [98]). This may be due to the fact that the very definition of the DNA barcode gap has undergone refinement over the years. As a result, no one "true" quantitative approach exists for measuring the

DNA barcoding gap that can be unanimously agreed upon within the barcoding community. This means that use of arbitrary ("fixed") distance cutoffs to separate out all taxa, such as the well-known 2% heuristic [75, 77] no longer holds [201]. Rather, because barcoding data is in a state of continual flux, as more specimens are collected and as taxonomic revisions are made, taxon distance thresholds should instead be directly computed from specimen DNA sequences when possible [35, 197]. This assertion comes as no surprise since species themselves, while obscure in nature and broad in concept, are, in reality, simple testable cladistic hypotheses refuted solely on the basis of existing expert knowledge and newly acquired information [133].

Numerous computational and statistical methods have been proposed over the years to better quantify the magnitude of the barcode gap. For instance, the $10\times$ rule [79] goes some way into accomplishing this, but it cannot be readily applied to distinguish among all taxa due to differences in both taxon evolutionary and life histories. The absence of a DNA barcoding gap to reliably discriminate species can be attributed to three primary factors: (1) the recent/rapid splitting of species from the Most Recent Common Ancestor (MRCA), leading to the retention of ancestral polymorphisms as a result of incomplete lineage sorting, introgressive hybridization or species synonymy; (2) the likely presence of cryptic species diversity due to lack of fixed morphological differences among closely-related taxa; and, (3) human-mediated errors (*e.g.*, overlumping/oversplitting of taxa) in the identification of specimens by experts [85, 98]. Thus, the employment of taxon-specific distance thresholds, as opposed to generic cutoffs, seems more reasonable.

As a consequence, the adoption of a number of query-based criteria designed to aid in the reliable separation of intraspecific and interspecific distances has propagated throughout the DNA barcoding community and literature over the years. These include for instance the Best Close Match criterion employed within the TaxonDNA software [116], and methods available in the `spider` [20], the `adhoc` [164] and the `BarcodingR` [198] R packages. In spite of the introduction of various methods to aid the solving of the species genetic separation problem, a significant knowledge gap persists: the apparent dearth of statistical thoroughness that accompanies the majority of published DNA barcoding studies.

4.2 A Need to Improve and Maintain Statistical Rigor in DNA Barcoding Studies

Here, evidence is presented that has greatly hampered the acknowledged and untapped potential of DNA-based specimen identification and species delineation: the lack of statistical rigor in DNA barcoding. Despite having been pointed out as a clear limitation multiple times in various capacities by authors of previous studies [111, 115, 129, 142], this issue has not yet garnered the scrutiny it desperately deserves. In fact, not having been explicitly addressed as a major problem at all until now is highly disconcerting. This may be due to the fact that there is no one set definition for statistical rigor in the literature, partly because, like science, statistics is rooted deeply in epistemology, and more generally in philosophy [106]. The problem faced here however is that the majority of researchers, particularly those in life science fields, lack an appropriate level of statistical knowledge

necessary for the proper application of statistical methods [54]. As a result, misuse, abuse and misinterpretation of quantitative results is rampant in academic settings. By no means are DNA barcoding studies immune to this. Such naïveté has led to the overuse of 'basic' parametric statistical procedures such as *t*-tests and the drawing of incorrect conclusions from *p*-values [186]. These and other statistical "sins" are so widespread in academic publications that some statisticians have devoted much of their time, and even their entire careers, to writing about the most common errors made by non-statisicians and steps to take to avoid making them (*e.g.*, [59]). Thus, here, statistical rigor is informally and simply defined as the use of appropriate quantitative methods to test and justify hypotheses in light of empirical evidence (*i.e.*, data) and uncertainty. This definition is adopted herein. Notably, it is stressed that the ubiquitous barcoding gap should be better defined on a statistical level, contingent on its use for the task of identifying unknown specimens or in describing novel species. However, as most DNA "barcoders" are not also statisticians, the lack of a statistically-precise static definition for the DNA barcode gap is understandable, albeit one that is absolutely necessary. Although there is much that could be elaborated on here, in this brief investigation, focus is specifically placed within the context of the need for the sound interpretation of the DNA barcoding gap as a necessary and sufficient criterion to assess the overall perfomance of DNA barcoding.

In the following subsections, problems with barcode gap interpretation from the standpoints of (1) requiring higher intraspecific specimen sample sizes to adequately capture standing genetic variation, (2) needing better descriptive statistics, along with

visualization methods, to concisely and accurately summarize taxon genetic sequence distance data, and (3) necessitating more appropriate statistical inference procedures to draw meaningful conclusions from limited DNA sequence data are outlined. Throughout the present work, Meier *et al.*'s [117] version of the DNA barcoding gap is employed; however, one can easily replace 'maximum' with 'mean' everywhere in the context of interspecific distances with the understanding that discrepancies as to which species show a barcode gap may (and often do) result. Viable solutions are then proposed which better harmonize the seemingly disparate disciplines of DNA barcoding and statistics. Moreover, the methods proposed herein also extend to the notion of (statistical) *reproducibility*. Many scientific studies lack sufficient information (including detailed explanations, quantitative data and metadata) necessary to replicate original experiments. A prime example where a sufficient level of detail is crucial to convey to researchers is in the description of agent-based models (ABMs), which are used widely in ecology. Typically, ABMs incorporate numerous assumptions needed to establish baseline individual- and group-level behaviour in "perfect-world" scenarios. To ensure that such rigor is not compromised, Grimm *et al.* [63] introduced and outlined a standard protocol that sought to bridge challenges of ABM **O**verview, **D**esign concepts and **D**etails (ODD). This work has since been expanded upon to more fully encapsulate the elements needed to adequately describe ABMs in a complete but succinct manner [64]. The approaches outlined and examined below go some way into better enabling reproducibility, much like the ODD Protocol, as they are not only planted firmly in solid statistical theory, but are also

straightforward to implement and easy to understand by the statistical nonexpert.

4.3 Case Study: DNA Barcoding of Pacific Canada's Fishes

From this point onward, statistical approaches to better characterizing the DNA barcode gap will be framed in the context of the barcoding of Canadian Pacific fishes as a focal case study. Many fish species native to the Pacific (*e.g.* Sockeye salmon (*Oncorhynchus nerka*)) hold strong socioeconomic and conservation importance globally, particularly as central food commodities within the supply chain. As such, in recent decades, much work has gone towards better understanding patterns of standing genetic diversity in this group to aid recovery of declining fish stocks.

DNA barcodes were downloaded from BOLD on December 1, 2020. Specifically, sequence data were taken from Steinke *et al.* [171] (BOLD Project: TZFPC Fishes of Pacific Canada Part I) and consist of 1219 specimens representing 197 species (as of the date of download). At the time of project release and publication of Steinke *et al.* [171], data comprised 1225 specimens records from 201 species. Within the current dataset two specimen records (Process IDs: TZFPA062-06 and TZFPB406-05) were flagged as problematic (*i.e.* misidentified) in BOLD and an additional sequence (Process ID: TZFP069-04) was outside the barcode region length necessary for BARCODE compliance (*i.e*, said sequence was shorter than 500 bp; [68]). Only the latter record was excluded from further analysis, leaving a final sequence count of 1218. Through the BOLD Public Data

Portal, the excluded record was found to be submitted as *Ophiodon elongatus* (Lingcod) based on a 490 bp barcode sequence, whereas the two misidentified specimens were identified at the time of record submission to the species level as *Arctozenus risso* (Spotted barracudina) and *Lipolagus ochotensis* (Eared blacksmelt), respectively. Using BOLD's Animal Identification [COI] Engine with the default current Species Level Barcode Records (for both public and private records, including interim species, and a mimimum 500 bp sequence length) database however, the specimen previously queried as *Arctozenus risso* is actually *Lestidiops ringens* (Slender barracudina), based on a match to a 647 bp fragment with 100% similarity as of December 1, 2020. This match was additionally verified through examination of a neighbour-joining tree (not shown), where the unknown query barcode clustered tightly with a single *Lestidiops ringens* individual. Said sequence was matched to the top-most *Arctozenus risso* published record with only 97.67% similarity. *Lipolagus ochotensis* was still diagnosed by the BOLD ID Engine as such with 100% probability based on a 652 bp segment. Sequence alignment necessary for calculation of the DNA barcoding gap was carried out directly using the built-in amino acid-based Hidden Markov Model (HMM) aligner due to dataset size. The default Kimura-2-Parameter DNA substitution model was maintained, along with the default Pairwise Deletion option for ambiguous base and gap handling.

Using the Barcode Gap Analysis tool available through the BOLD Workbench, results revealed a total of 38 species (19.3%) had nearest neighbour distances less than the 2% threshold. The species *Lipolagus ochotensis* showed a maximum intraspecific distance of

1.24% and a minimum interspecific distance of 13.43% (nearest neighbour: Northern smoothtongue (*Leuroglossus schmidti*)), while distances for *Arctozenus risso* (a singleton) were 0% and 17.72% (nearest neighbour: Northern pearleye (*Benthalbella dentata*)), respectively. Although observed magnitudes of genetic distances for both of these species suggest that DNA barcoding "works" and is an effective tool when it comes to specimen identification, it is nevertheless unsettling that all mentioned species' nearest neighbours fall into separate genera. This finding suggests a lack of overall specimen sampling depth for these species and perhaps Pacific fishes in general [171]. One species, the Deepwater bristlemouth (*Cyclothone atraria*), displayed a maximum intraspecific distance of 9.22% (nearest neighbour distance: 22.78%; nearest neighbour: Stout blacksmelt (*Pseudobathylagus milleri*)). This is strong indication of *potential* cryptic species diversity. All other within-species distances were below 2%. It should be noted here that specimens assigned as correctly-identifed or misidentified to a given species, as well as those individuals displaying cryptic genetic variation or evidence of barcode sharing may in fact not bear these characteristics. Because specimen sample size information was not provided by Steinke *et al.* [171], it is impossible to directly discern how reliable reported genetic distance measures are, and therefore the trustworthiness of estimated DNA barcode gaps.

4.4 Evidence for the Lack of Statistical Rigor in DNA Barcoding

Prior to delving any further into the primary elements that constitute the lack of statistical rigor in DNA barcoding, along with the discussion of simple solutions to help aid its mitigation, astute readers may have noted thus far the use of the term "distance" to describe both genetic variability within species as well as among species. This is no mistake. To the untrained eye, these terms are synonymous from a lexical point of view, and can be (and often are) used interchangeably within general writing. However in scientific writing, this constitutes a major *faux-pas*. Recently, DeSalle and Goldstein [39] reiterated the importance of carefully balancing word meaning and word choice in barcoding papers so that author(s)' overall intent is not blurred. Numerous highly-cited past DNA barcoding studies employing barcode gap analyses have unknowingly used the term "divergence" to denote gene variation seen across species. Even some authors of the current work are guilty of this. Such word usage bears similarity to the confusion between the terms "species identification" and "specimen identification", as raised by Collins and Cruickshank [35] as the first of seven deadly sins of DNA barcoding. There is an important mathematical/statistical distinction between distance and divergence which must be stressed: distances are *symmetric*, whereas divergences are not. Considering two different specimens (or species), calling them A and B, then the distance between A and B is equal to the distance between B and A. That is,

$$d(A, B) = d(B, A).$$

Such a pattern is easily observed from examining a pairwise distance matrix of intraspecific and interspecific genetic distances. Values are identical (zero) with respect to the main diagonal (moving top left to bottom right), as a specimen or species will display zero distance from itself to itself. This can also be seen through exchanging matrix rows for columns and *vice versa*. On the other hand, the notion of divergence speaks to how different two probability distributions are from one another. Thus the term "distance" is employed everywhere throughout the current work when referring to intraspecific and interspecific differences. While this confusion does not directly contribute a lack of statistical rigor *per se*, to this end, all future DNA barcoding studies employing barcode gap analyses should use the term "distance" to avoid any potential ambiguity and confusion.

4.4.1 Improper Allocation of Specimen Sampling Effort

Current specimen sampling efforts for DNA barcoding have been improperly delegated to further the growth of public reference sequence databases such as BOLD and GenBank. Both geographic and taxonomic barcoding projects and campaigns have been far too focused on exhaustively sampling as many species as possible [142, 144]. This assertion is immediately evident from examining BOLD species lists, where an overwhelming majority of species are singletons or doubletons. From this observation, it is clear that the sampling of intraspecific rather than interspecific genetic variation has

been severely limited. While both extremes of genetic variation are necessary to fully comprehend and assess the scope and magnitude of species limits, taxon rarity combined with narrow sampling typical in DNA barcoding studies precludes one's ability to paint a full picture [4]. Barcoding initiatives should therefore instead be focused on the dense sampling of an *optimal* number of *specimens* per species, which should be strongly calibrated by factors such as research budget, cost and funding [22, 168].

In the early days of the DNA barcoding endeavour, it was decided by the Consortium for the Barcode of Life (CBOL) that at least 5-10 specimens per species be collected from wide geographic regions for assembly of reference sequence libraries; indeed, this heuristic has been globally adopted by barcoding campaigns such as the Fish Barcode of Life (FISHBOL) [183] in an attempt to limit project costs and maximize returns. However, while collection of only a few individuals of every species is a good starting point, recent studies have highlighted that such small sample sizes are likely far from adequate to capture the majority of standing haplotype variation found within species; instead, hundreds to thousands of individuals may be needed based on both empirical findings and simulation studies [142, 144, 199]. Further, it is imperative that, in addition to target species, sister species also be adequately sampled. This is necessary for both the strong detection and the correct interpretation of the DNA barcode gap. In the case of monotypic genera, representatives from the closest allied genus should also be targeted.

Estimating Intraspecific Specimen Sample Sizes with the R Package `HACSim`

The lack of a comprehensive and robust sampling of within-taxon genetic variation is a very real problem for molecular species diagnosis because it impedes the ability of DNA barcode researchers to acquire a full understanding of standing levels of intra-taxon haplotype diversity that enables rapid and reliable species differentiation.

To this end, the R package `HACSim` [143] can aid biodiversity researchers and regulatory scientists in assessing current levels of specimen sampling effort reflected in genomic sequence libraries like those housed in BOLD (*i.e.*, through computing the observed fraction of haplotype diversity that has likely been sampled within species). The method can further assist researchers in obtaining optimal specimen sample sizes likely required to adequately capture the majority of haplotype diversity found within presumably panmictic species randomly sampled across their entire geographic/ecologic ranges (*i.e.*, through extrapolating haplotype accumulation curves and observing the point on the x-axis where curves begin to saturate toward an asymptote). Thus, the likelihood of observing a true species barcode gap is increased when specimen sampling effort is high. Furthermore, the employment of `HACSim` to better gauge required sampling depths within species means that less reliance will ultimately be placed on arbitrary distance thresholds such as the 1% cutoff employed within BOLD [150, 151] to assign Linnean names to user-submitted query sequences based on expertly-verified references. Since it has long been recognized that a given taxonomic level is not equivalent across different evolutionary lineages (*e.g.*, a family of insects is not equal to a family of fishes), it is reasonable to expect that species falling on

separate branches of the Eukaryotic Tree of Life will warrant the use of different distance thresholds when it comes to specimen identification. In fact, it seems resonable that the output of `HACSim` can be employed to calculate optimal distance thresholds for reliable species separation. This is because, with larger specimen sample sizes and increasing spatial scale, intraspecific genetic distances will tend to increase, while distances observed among species will shrink [118].

`HACSim` has been designed with user-friendliness in mind. The tool is specifically relevant to assessing genetic variation derived from Sanger-based amplicon reads obtained from any taxon under study and any molecular marker of interest. It is our belief that such a method will be of invaluable use to the DNA barcoding community-at-large. However, as we progress deeper into the realm of big data, the overarching potential of `HACSim` to aid in the characterization of next-generation sequencing (NGS) and High-Throughput Sequencing (HTS) data for environmental DNA (eDNA) and metabarcoding applications becomes clear. This said, it is critical that the capabilities of `HACSim` be expanded upon, especially the ability to handle multiple specimen reads. Thus, all computational DNA barcoders should consider contributing to this endeavour.

4.4.2 Failing to Properly Visualize Intraspecific and Interspecific Genetic Distances

A large majority of published DNA barcoding studies infer the detection (presence or absence) of a species' barcode gap through visualization of specimen pairwise sequence distances as either histograms or dotplots [35]. Collins and Cruickshank [35] were correct

to suggest the employment of dotplots as opposed to frequency histograms to better depict the estimated distribution of species' interspecific and intraspecific distances, but they failed to offer a more thorough quantitative treatment as to why this is the case.

Circumventing the Problem with Histograms and Dotplots for Barcode Gap Display

Histograms partition numerical data into *discrete* class intervals called bins to more easily visualize how sample data is distributed. However, the use of histograms, while both ubiquitous as a statistical summarization method and widely-understood by many, can often muddy the true shape of probability distributions if both the bin width and number of bins in which to group data are not chosen wisely. Histograms with narrow bins tend to be more precise when density of the sample data is low; whereas, when density of observations is high, wider bin widths should be preferred because of the tendency to better expose true data signal relative to noise [157]. Despite the added benefit of experimenting with bin widths to reveal hidden structure within data, most software now routinely employed to construct histograms, such as R's `graphics` [149] and `ggplot2` [188] packages, utilize equal bin widths in generating histograms by default. Too small a choice of the number of bins and the histogram will be very rugged (*i.e.*, have high bias); too large the number of bins and the histogram will be oversmoothed (*i.e.*, possess high variance) [157]. If DNA barcode researchers choose to continue to use equal histogram bin widths to display the barcode gap, then consideration of the optimal number of bins to employ needs to be carefully investigated. Several measures of appropriate bin numbers to use have been proposed in the statistical literature such as the robust Freedman-Diaconis rule [55],

which makes use of the sample interquartile range (IQR), or Scott's Normal reference rule [157], which employs the estimated (sample) standard deviation calculated from Normal distributions. Unfortunately, most heuristics (including the ones mentioned here) place a strong dependence on sample size. For instance, Microsoft® Excel sets the number of histogram bins to be equal to the square root of the number of data observations, whereas `graphics` employs Sturges rule [176], basing the number of bins to scale proportionally with the base-two logarithm of the number of samples, while `ggplot2` defaults to using 30 bins regardless of dataset size. For Sturges rule, bin width is computed from dividing the sample range of the data by the optimal bin number. The validity of Sturges rule in particular has been called into question as it tends to oversmooth data in the case of large samples [90] and there have been calls for the usage of more reliable methods. Because studies potentially employ different software for histogram generation, results are no longer directly comparable; thus, care must be exercised when making generalizations.

A much better alternative to displaying the DNA barcoding gap is to rely on the *continuous* variant of the histogram, kernel density estimation (KDE) plots [138, 154], to more accurately inform on the actual population distribution of the barcoding gap through depiction of intraspecific and interspecific pairwise genetic distances as smooth curves. KDE works by weighting data observations relative to their distance to other similar-magnitude data points. Much like histograms however, KDE often requires careful parameter selection, in particular regarding the kernel type and the kernel bandwidth. The kernel type strongly defines the overall shape that the density curve takes on, whereas

the kernel density bandwidth controls the amount of smoothness of the generated curve. Optimal choice of these parameters is crucial so as to not distort real patterns present within the data. Most modern software (such as R) employ defaults which tend to work well under a wide variety of situations, letting the data do all of the talking, but also give the user fine control over parameter initialization. However, automatic settings can sometimes lead to undesirable results. R for instance employs a Gaussian kernel and chooses the kernel bandwidth to be equal to the standard deviation of the kernel itself; this should be sufficient as far as estimation of the DNA barcode gap is concerned. Often with kernel density estimation, data may extend beyond those observed from histograms. In particular, data that are constrained to only positive support values can end up having negative density values, which for genetic distances, is not biologically meaningful. In practice however, this is not an immediate concern since truncation methods exist to ensure that data located at the boundaries of KDE plots have positive support.

The dotplot approach to inferring the barcode gap (**Fig. 1**) is simple: on a plot of maximum intraspecific genetic distances (displayed on the x-axis) *versus* minimum interspecific distances (shown on the y-axis), represented by points for every barcoded species, a line corresponding to the function $y = x$ is drawn. Points occurring above this line suggest that a barcode gap is present for a given species and that DNA barcoding "works". In contrast, points falling below the 1:1 line for any species suggest that the DNA barcode gap is absent, and thus barcoding fails to tell specimens apart. Often, points plotted in this fashion overlap tightly, making species-by-species visual inspections

difficult. **Figure 4.1** clearly shows that many Canadian Pacific fish species exhibit maximum intraspecific distances very close to, or equal to, zero. This strongly indicates that adequate specimen sampling needed to characterize standing haplotype diversity at the species level is severely lacking.

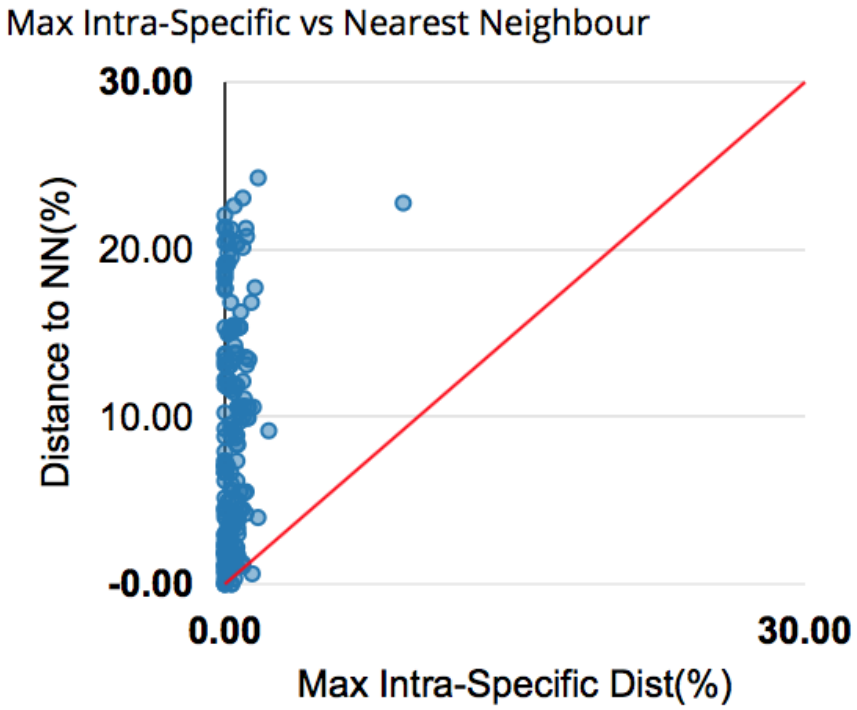

Figure 4.1: Traditional dotplot for visualizing the DNA barcode gap for a range of Canadian Pacific fishes assessed by [171] and generated using the BOLD Workbench's Barcode Gap Analysis tool. Data comprise those specimens currently found in the TZFPC BOLD project (as of December 1, 2020) and represent 1219 specimens from 197 species (*c.* 6.18 specimens per species on average). Points lying above the 45° line indicate that species show a barcode gap and are readily identified via molecular barcodes. On the other hand, points falling below the 1:1 line suggest that species lack a barcoding gap and thus are not easily diagnosed through their DNA barcodes. Most species assessed here (38 of 197 species (19.3%)) display a barcode gap since the minimum interspecific genetic distance exceeds the maximum intraspecific genetic distance. Despite this, evidence of species showing a barcode gap may in fact be an artifact of limited sampling of within-species haplotype variation. The species *Cyclothone atraria* at the point (9.22, 22.78) is clearly visible as an extreme outlier and signals possible cryptic species variation.

The use of traditional dotplots to display the barcode gap would be better represented as half-logarithm dotplots (**Fig. 4.2**) which plot sorted log-transformed genetic distances for every species included in a taxon dataset against the number of species sampled [171, 172]. A horizontal line is then drawn at the 1% mark (or similar threshold), allowing for good separation of intraspecific distances from nearest neighbour distances. Plotting sorted genetic distances in this manner allows for relative differences to be easily seen among species [171]. Further, through employing a log transformation of species' genetic distance data, interesting patterns are more easily spotted without worry for any loss of information. This is the case for two reasons. First, since $y = log_a(x)$ is *monotone*, the order of plotted points is preserved. Second, the log transform is *variance-stabilizing* because it has the effect of making positively-skewed data less skewed through removing any dependence existing between the mean and variance of a set of data observations. Without such a transformation in place, sample observations would likely display varying levels of heteroscedasticity (*i.e.*, non-constant variance). As with **Figure 4.1**, numerous data points (representing over 60 fish species) lie directly on the x-axis, indicating a complete lack of sufficient specimen sampling (**Fig. 4.2**). Despite its promise, it appears that the modified dotplot has not caught on within the DNA barcoding community outside a select few fish DNA barcoding studies [171, 172].

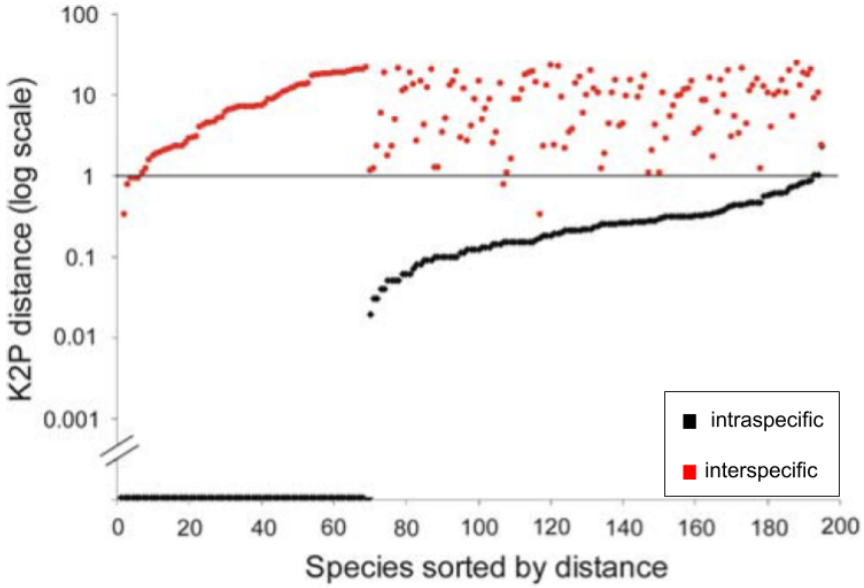

Figure 4.2: Half logarithm dotplot for the display of species' genetic distances modified from [171] for fishes from Pacific Canada. Plotted data comprise those specimens orginally analyzed by [171] (*i.e.*, 1225 specimens from 201 species). Most sampled species are resolved at the 1% log level of genetic distance.

A much more intuitive means of displaying intraspecific distances and interspecific distances is through using "quadrant plots" (**Fig. 4.3**) because they can be employed to directly detect "outlier" and problematic species in need of closer examination. In this approach, as in the generation of traditional and half-logarithm dotplots, barcoded species are depicted as points on a plot of maximum intraspecific distances on the x-axis *versus* minimum interspecific distances on the y-axis. Points fall into one of four categories in positive Cartesian space, depending on a predefined species distance cutoff (2% typically). Moving in a clockwise fashion from the top left corner, each category can be viewed as a case of either "barcoding success" or "barcoding failure". Quadrant I corresponds to the case where species are easily discriminated using DNA barcoding and reflects concordance

with currently accepted Linnean taxonomy (*i.e.*, interspecific distances are greater than the prespecified level cutoff, while intraspecific distances are less than the chosen threshold — a "success"). Species falling into Quadrant II likely represent cryptic complexes (*i.e.*, both intraspecific distances and interspecific distances are greater than the prespecified level cutoff — a "failure"). Species in this partition are indistinguishable through morphology alone and as a result are lumped under a single species name by taxonomists. Quadrant III encompasses evolutionarily young species that have recently diverged from the MRCA (i.e., not enough time has elapsed to allow nucleotide differences in the barcode region to accumulate — a "failure"). This category can also include species that are known by various synonyms. Finally, Quadrant IV includes likely misidentified specimens or instances of hybridization between closely-related species — a "failure"). *Cyclothone atraria* is the only species that would fall into Quadrant II. Based on computed genetic distances, both *Arctozenus risso* and *Lipolagus ochotensis* would be classified as belonging to Quadrant I; yet BOLD categorizes each of them as misidentified. This result is telling: it strongly suggests that *Lipolagus ochotensis* was represented in Steinke *et al.* [171]'s dataset by only a handful of collected specimens. Thus, the plausibility of both *Arctozenus risso* and *Lipolagus ochotensis* as barcoding "successes" (and therefore presenting a real barcode gap) should be immediately called into question. Like the half-logarithm dotplot, the quadrant plot approach has seen very limited use in barcoding studies, despite its inherent simplicity. Such plots appear to have only been employed in two previous publications [79, 85].

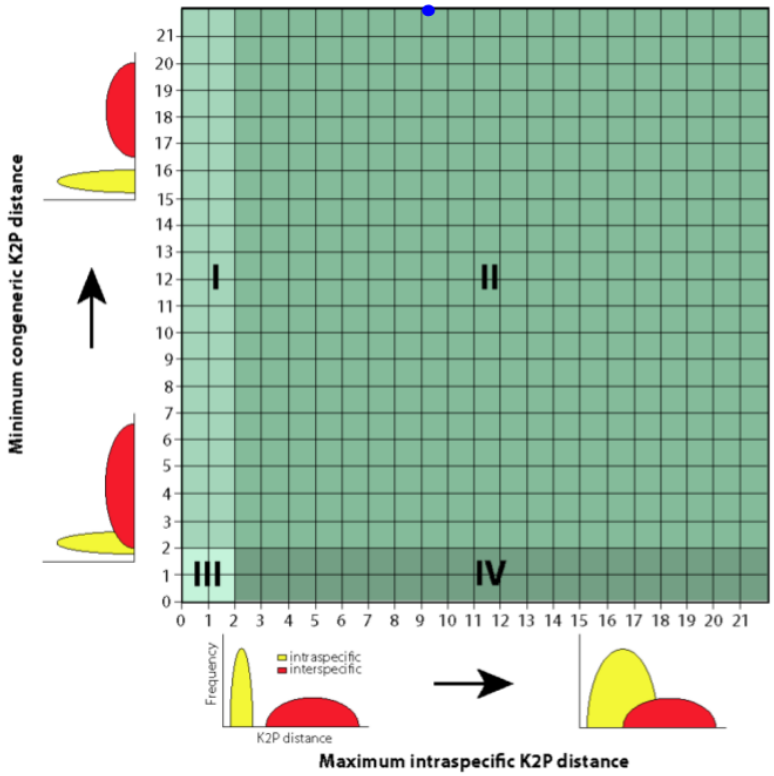

Figure 4.3: plot for the depiction of species' genetic distances reproduced with modification from [85]. Such plots are informative since problematic and/or "outlier" species can be easily detected and the success/performance of DNA barcoding assessed. Species are partitioned into four mutually exclusive groups based on observed magnitudes of intraspecific and interspecific genetic distances. Here, a 2% distance threshold is assumed to separate most taxa. The blue dot at the approximate point (9.22, 22.78) within Quadrant II represents *Cyclothone atraria*, a likely cryptic species complex. Plot axes show the relationship to the density curves shown in [118]. While usage of the Kimura-2-Parameter (K2P) DNA evolution model is both widespread and criticized in DNA barcoding studies and the community-at-large [34, 167], other more parsimonious nucleotide substitution models (such as the uncorrected p-distance) can be adopted without loss of information (and may even be better suited in the long run). Theoretically, between-species genetic variation should greatly exceed barcode sequence variation observed within species (Quadrant I; minimum interspecific distance $-$ maximum intraspecific distance $> 2\%$). Practically, this will only be the case when specimens have been adequately sampled.

One element that visual tools fail to reveal however is whether a barcoding gap likely *exists*. To properly address this question, more rigorous statistical methods are required.

4.4.3 Inconsistent, Inappropriate Use, or Absence of Inferential Statistical Procedures in DNA Barcoding

Attempts to place DNA barcoding on more statistically-solid ground have been undertaken several times before, particularly with regard to specimen classification (*e.g.*, [1, 9, 109, **?**, 115, 129, 200]) and species delineation. Many of these proposed methods have seen widespread usage, while others seem to be rarely employed in certain instances due to their inherent mathematical complexity and/or black-box nature.

Perhaps the first instance of the use of statistical algorithmic approaches in DNA barcoding was for the purpose of specimen classification. Such methods relied mostly on ideas from classical inferential paradigms such as likelihood theory and subjectivist (Bayesian) thinking, whereas others took inspiration from more modern models, particularly machine learning. One promising, yet grossly underrated technique worth mentioning here is the probabilty of correct identification (PCI) [114, 166]. While the PCI has many variants, its primary function is to serve as a simple metric of DNA barcoding efficacy given a richly-populated and fine-tuned reference database. The PCI statistic has mostly seen use around appropriate marker selection for DNA barcoding, particularly in regard to challenging taxa such as fungi, plants and protists. At its heart, the PCI is nothing more than a binomial proportion whose sampling distribution is easily estimated using resampling procedures. From here, it is trivial to calculate quantities of interest such as standard errors and confidence intervals. In essence, the strong mathematical and statistical theory that underlies the PCI is what is missing and should be emulated in future DNA

barcoding studies.

The use of statistical approaches for species delimitation has also generated much interest. However, it has been strongly cautioned to tread down this road carefully, considering multiple strategies to approaching species demarcation tasks, as well as explanations for species existence, origin and formation over space and time [24]. The majority of proposed approaches in this regard have been centered on coalescent theory [96]. In 2011, Puillandre *et al.* [148] introduced the widely-popular Automatic Barcode Gap Discovery (ABGD), a nonparametric statistical method to discriminate species based on the existence of the barcode gap, using avaliable DNA sequence data, as opposed to generating taxon
phylograms beforehand. Prior to this, heavy reliance fell upon the Generalized Mixed Yule Coalescent [57, 120, 146], an extremely time- and memory-consuming model-based approach to species delimitation based on branching patterns observed within ultrametric phylogenies generated using third-party software such as Bayesian Evolutionary Analysis Sampling Trees (BEAST; [43]) or MrBayes [87], and analyzed using the `splits` (SPecies LImits by Threshold Statistics) R package [51]. Since then, other methods to delimit species have been introduced to analyze barcode data (*e.g.*, Poisson Tree Processes (PTP); [202]). Earlier approaches such as haplotype parsimony networks [179], constructed using software like TCS [32], have found their way into DNA barcoding, despite known interpretational issues such as the tendency to form disconnected subnetworks, or the inclination to group species together within the same node [73, 142]. In addition, the

default 95% detection limit (*i.e.*, the probability of parsimony) employed within TCS is largely arbitary; users can set this value to range anywhere from 90-99% [32]. Thus, the choice of distance cutoff can have a large effect on the ouputted network topology. The above methods can differ greatly in the number of species delimited. Luo *et al.* [112] notes that GMYC tends to overestimate (oversplit) species, whereas underestimation of species (*i.e.*, undersplitting) is evident for ABGD. The Barcode Index Number (BIN) approach [151] seems to be a good compromise as it is fast to run, straightforward to implement and resulting output is easily interpreted [94]. A BIN comprises a unique alphanumeric code corresponding to a tight cluster of closely-related haplotypes. The BIN framework employs hierarchical clustering (via the REfined Single Linkage (RESL) algorithm), along with Markov clustering, and often suggests species numbers between the extremes of ABGD and GMYC through partitioning DNA sequences into four mutually exclusive groups largely reflective of actual species: MATCH, MERGE, SPLIT and MIXTURE, on the basis of genetic distances [151]. These presumptive groups or operational taxonomic units (OTUs) are biologically interpretable: MATCHES conform to established Linnean taxonomy; MERGES indicate that distinct species are indistinguishable through DNA barcoding and should be combined under a single species name; SPLITS reflect the presence of multiple species under a common Linnean name (*i.e.*, cryptic species diversity); finally, MIXTURES reveal possible specimen misidentifications or instances of introgression/hybridization [151, 159]. A direct relationship exists between BIN categories and quadrant plot categories mentioned previously: MATCHES correspond to Quadrant I; SPLITS make

up Quadrant II; MERGES fall into Quadrant III and MIXTURES lie in Quadrant IV. However, despite its promise, a major drawback to the use of the BIN system as a suitable species proxy is that it is a black box whose underlying algorithm is not well-understood by researchers outside the DNA barcoding community, such as regulatory scientists.

Framing the DNA Barcode Gap as a Statistical Hypothesis

There is a need to define the barcode gap more formally as a composite (one-sided) statistical hypothesis test. An analogy here can be made to testing the hypothesis that a gene evolves neutrally in a species population. Such hypotheses can be assessed using a wide variety of tests such as Tajima's D [177]. In the present case, the null hypothesis of no barcode gap for a species would be tested against the alternative hypothesis that a barcode gap exists. Mathematically,

$$H_0 : \text{minimum interspecific distance} - \text{maximum intraspecific distance} \leq d_0$$

versus

$$H_1 : \text{minimum interspecific distance} - \text{maximum intraspecific distance} < d_0.$$

where d_0 is a predefined cutoff for species separation (say, d_0 = 2%). Here, the null hypothesis (H_0) is assumed to be true and the ultimate goal is to reject it in favor of the alternative (H_1), in light of data already observed. That is, it is assumed that present DNA sequence data support the existence of a barcode gap at the species level. Under this

scheme, it is easy to distinguish between Type I errors (false positives) and Type II errors (false negatives). A false positive is analogous to taxonomic oversplitting (*i.e.*, nearest neighbour distance $<$ maximum intraspecific distance); whereas, excessive lumping of species (*i.e.*, nearest neighbour distance $>$ maximum intraspecific distance) strongly indicates that a false negative error has been made. A one-tailed test is chosen here, as opposed to the more widely employed two-sided test since between-species genetic variation usually exceeds that seen within species, with few exceptions. Such an approach leads to a more powerful test with greater flexibility than would be allowed using a two-sided test.

An immediate challenge exists in formulating an appropriate hypothesis test statistic for the barcode gap. Test statistics are usually of the form

$$T = \frac{d - d_0}{\text{SE}[d]}$$

where d is the observed difference between minimum interspecific distance and maximum interspecific distance and SE denotes estimator standard error. For already well-sampled species (*i.e.*, those with a large number of collected specimens), the above test statistic would approximately follow the standard Normal distribution whenever H_0 is true. Unfortunately, in the case of small specimen sample sizes, deriving an expression for the standard error of the estimated barcode gap would be difficult and the distribution of said test statistic would be also not be obvious.

Framing DNA barcoding in a statistical way is clearly needed, since for

densely-sampled clades, a barcode gap is almost surely to exist. Through employing deep taxon sampling schemes, DNA barcode researchers will be able to more easily detect a true species' barcode gap when one is actually present.

The Use of Nonparametric Bootstrapping to Estimate the DNA Barcode Gap

In addition to simple point estimates (and associated standard errors) of the barcoding gap for varying taxa which are widely reported (*e.g.*, [189]) future studies should also report confidence interval (CI) estimates around the estimated population (or "true") maximum intraspecific distance, minimum interspecific distance and the barcode gap using sample data of intra- and interspecific distances. A simple but naïve solution in this regard is to form CIs using the data at hand; however, this requires the strong assumption that genetic distances are drawn from a large normally-distributed population; in reality, the sampling distribution of pairwise distances is unknown since it is likely to be highly taxon-dependent. This should come as no surprise since genomic markers employed to assign taxon-level matches to unknown specimens using DNA barcoding show varying rates of molecular evolution both within and across taxonomic groups. These observed differences in taxon molecular evolutionary rates strongly affect fundamental processes at both the microevolutionary (*e.g.*, random genetic drift, mutation, natural selection) and macroevolutionary (*e.g.*, speciation) scales.

Thus, a better approach to reporting parameter estimates, which does not require the sampling distribution to be known *a priori*, and relaxes distributional assumptions through allowing for reasonably small sample sizes, is to employ nonparametric bootstrapping to

continually resample from observed distances a large number of times (say, 10000 times) uniformly (*i.e.*, with equal probability) with replacement [47]. Sampling with replacement ensures that drawn observations are both independent and identically distributed; that is, sampling a given observation has no bearing on the occurrence of a future observation and all observations are generated from the same underlying statistical population. The idea here is that, for a large number of bootstrap replicates, the distribution of resampled distances (*i.e.*, the bootstrap sampling distribution) mimics the actual distance distribution for the taxon under study quite closely. Such a scheme is analogous to bootstrapping in phylogenetic inference to assess how well nodes within neighbour-joining trees support the observed data [48, 53]. Because a test statistic need not be known in advance, bootstrap results can be immediately used to form appropriate level (*e.g.*, 95%) bootstrap confidence intervals for the population barcoding gap. Statistical interpretation of such constructed intervals is relatively straightforward: if the intervals contain the value d_0, then the hypothesis that the maximum intraspecific distance does not differ significantly from the minimum interspecific distance at the hypothesized value d_0 cannot be rejected at the stated significance level (*e.g.* α = 5% for 95% confidence). Put another way, if d_0 falls within the obtained CI, then the hypothesis that no barcode gap is present cannot be rejected at the chosen level of statistical significance.

Nonparametric bootstrapping is known to perform poorly in certain situations. One such failure of the traditional bootstrap is in the estimation of extreme order statistics such as the population minimum or the population maximum. Standard bootstrapping,

sometimes termed the n-out-of-n bootstrap, works by drawing resamples of the same size as the original sample. As the "revised" DNA barcode gap is defined in terms of the maximum intraspecific distance and minimum interspecific distance, the usual bootstrapping procedure detailed above is not applicable. It is worth mentioning that the n-out-of-n bootstrap would indeed work as expected under the "old" definition of the barcode gap, used prior to [117], since that definition involved only statistical means. Fortunately, there is an immediate remedy available. The trick is to take resamples of a *smaller* size than the original dataset [17]. This technique is known as the m-out-of-n bootstrap, where $m < n$. In employing such a method, the variability of corresponding estimates will be higher (larger) than in the regular bootstrapping procedure [29]. While this result may appear counterintuitive at first, assuming the variance of an estimator of interest is both constant and finite, said estimator's standard error will be smaller for a larger number of observations and larger for smaller sample sizes. Since $m < n$, another approach worth examining is random subsampling, which involves sampling *without* replacement [145]. The optimal choice of m however is not obvious and can have a signficant impact on obtained results. Therefore, algorithms for selecting appropriate values of m (such as that presented in [18]) must be investigated. Regardless, the above bootstrapping approach should be used to report point estimates and desired level CIs for "true" maximum intraspecific distances and interspecific distances, as well as population barcoding gaps in any and all future taxon-specific DNA barcoding studies (especially reference sequence library publications).

4.5 New Avenues for Estimating the DNA Barcode Gap

Finally, it is important to draw upon future promising avenues for continued work on accurately estimating the DNA barcoding gap. One potential application in this regard includes statistical mixture models which can account for genetic differences observed within and among species for the purpose of molecular specimen assignment. Mixture models offer great flexibility when it comes to accomplishing this task because correlations in haplotype diversity existing at the species level can be easily incorporated into such modelling frameworks. Much effort has gone into the development of easy to use computational tools to fit mixture models to a wide variety of data. A notable example in this regard that may prove valuable for barcode gap estimation is the R package `mclust` [158], software that has seen widespread use for the task of parametric model-based clustering in recent years.

Statistical methods for delineating species can inherently be viewed as "mixture models". All proposed species delimitation approaches to date find the optimal partition of DNA sequences into mutually-distinct groups that are highly reflective of actual species. Thus, the problem of species separation boils down to that of a simple clustering/classification task. The majority of methods generate these clusters on the basis of estimated phylogenetic relationships (*e.g.*, GMYC, PTP), along with an assumed parametric model of species generation (*e.g.*, birth-death model, Yule model), whereas others simply use the DNA sequences themselves (*i.e.*, ABGD) to arrive at a plausible solution in a nonparametric fashion. In recent years, novel "hybrid" approaches to tease

out species have been published. Notably, algorithmic methods such as [56] and [93] stray away from objective single-locus likelihood inference to also include subjective multilocus Bayesian inferential frameworks.

Another approach that should be investigated is the employment of nearest-neighbour and other machine learning methods used in clustering and classification tasks. However, the widespread success of machine learning methods is due greatly to the availability of large amounts of training data that feed and nurture artificial intelligence (AI) algorithms, a factor that poses problems for undescribed species, rare taxa and those with narrow geographic distributions (*e.g.* endemic species, monotypic taxa). With an arsenal of statistical tools like mixture models and nearest-neighbour methods in hand, practitioners will be better equipped to estimate important quantities central to DNA barcoding, including species separation thresholds.

Although not a statistical issue *per se*, the increased need for the sequencing of multiple genetic loci, particularly nuclear genes, to solidify confidence in specimen assignment and aid resolution of taxon boundaries, cannot be stressed enough. Much like the adoption of the rbcL and matK chloroplast genes for DNA barcoding of land plants, a similar case can be made for a dual, or better yet, multiple, mitochondrial-nuclear gene system for barcoding of metazoan taxa. COI has been demonstrated to lack sufficient discriminatory power for identification in groups such as sharks and the aptly named "problem children" (Cnidaria and Porifera) to name a few — all which show remarkably low rates of molecular evolution. In the case of animal DNA barcoding, several molecular regions (preferrably

both mitochondrial and nuclear) should be sequenced across the same sampled specimen whenever possible; in reality however, this is rarely done. While other International Barcode of Life (iBOL) member nations (*e.g.*, those in Europe) have accepted the move toward multilocus DNA barcoding with open arms, it seems that Canada, ironically, is not one of them. The largest global hub for DNA barcoding, the Centre for Biodiversity Genomics (CBG), appears to remain completely fixated on the promise of single-marker barcoding for the construction of reference sequence libraries and for the progression of biodiversity science as a whole. One can even argue that Canadian DNA barcoding's staunch position on maintaining the *status quo* and its blatant refusal to embrace necessary change, due to the overwhelming fear of becoming irrelevant, has greatly hindered the timely transition into the vast and exciting realm of "next-generation DNA barcoding" [178]. This is clearly evident from the fact that the majority of specimen sequence records found in BOLD are derived from just a single marker (COI). Within BOLD, substantially fewer specimen records orginate from other mitochondrial markers like cytochrome *b* (cyt*b*) and the mitochondrial D-loop; even fewer come from nuclear gene regions such as ribosomal DNA (rDNA) and rhodopsin (rho). Thus, sequence reference databases should strive to incorporate genetic information from multiple genomic sources to better aid specimen identification to the species level, especially since the DNA barcode gap is nonexistent in most taxonomic groups [97].

4.6 Concluding Remarks

In this brief piece, it was demonstrated that DNA barcoding currently lacks the statistical rigor needed to properly interpret results of species barcode gap analyses through focusing on three key areas: (1) the need for larger specimen sample sizes reflective of standing genetic variation within species; (2) the misleading display of intraspecific and interspecific distances, and (3) the absence of formal statistical inference procedures in DNA barcoding. A past study of Pacific Canada's fish fauna by [171] was employed to illustrate flaws in the presntation of the DNA barcode gap, as well as the need for larger specimen sample sizes to avoid biases in the reporting of within- and between-species genetic distances critical for reliably estimating the gap. First, the routine use of the novel R package `HACSim` will allow researchers to better assess the efficacy of current taxon sampling schemes and develop more robust collection protocols that will permit greater statistical power in detecting a true species barcode gap. Next, a more careful consideration of the depiction of the DNA barcode gap as a frequency histogram is warranted, as are alternative representations, including density estimation curves and the half-logarithm dotplot, due to interpretation issues surrounding default graphical parameters employed by many popular statistical analysis programs such as R and Excel. In addition, better ways to reconcile DNA barcoding with statistical inference include proposing the framing of the barcode gap as a one-tailed statistical hypothesis test, and backing the use of the nonparametric bootstrap to compute standard errors and confidence intervals for maximum intraspecific distances, nearest-neighbour distances, as well as the barcode gap. Finally,

new directions are offered for thinking critically about the robust estimation of the DNA barcode gap. Taken together, the methods outlined herein have the potential to open closed doors, giving biodiversity researchers and regulatory scientists an unprecedented view of key evolutionary mechanisms and processes responsible for shaping Earth's biodiversity over millions of years.

With these considerations in mind, both biodiversity and regulatory scientists alike will be well-equipped to constructively analyze vast amounts of DNA barcode data with greater confidence and as a result feel more secure in making critical assessments as to the performance of DNA barcoding on the basis of the barcode gap. The widespread adoption of the methods discussed herein will be of great importance in moving forward with the building of large-scale DNA barcode reference libraries within BOLD through global iBOL initiatives such as BIOSCAN [81, 82].

In closing, a word on ethics surrounding the need for comprehensive specimen collection for DNA barcoding is essential, while at the same time, first doing no harm whenever reasonably possible. To this end, specimen sampling efforts oftentimes result in the partial or complete destruction of individual organisms (*e.g.*, pinning whole insects for inclusion in museum collections, clipping part of a fish fin or bird wing for curation to aid downstream identification, electrofishing for the purpose of taking biometric measurements). Such practices seem to be in direct conflict with the increased urgency with which specimen sampling must be undertaken. For instance, in the case of rare or at-risk species, it may not be feasible to implement desired or traditional sampling strategies that

would result in sacrifice of entire specimens. Thus, it is important that DNA barcoders rely on alternative sources of specimen biomaterial such as museum collections or herbaria that seek to limit direct harm.

Conflict of Interest

None declared.

Chapter 5

Solving the genetic specimen sample size problem for DNA barcoding with a local search optimization algorithm

Jarrett D. Phillips[1*], Scarlett E. Bootsma[1], Daniel J. Gillis[1], Robert H. Hanner[2,3]

[1]*School of Computer Science, University of Guelph, Guelph, ON., Canada, N1G2W1*

[2]*Biodiversity Institute of Ontario, University of Guelph, Guelph, ON., Canada, N1G2W1*

[3]*Department of Integrative Biology, University of Guelph, Guelph, ON., Canada, N1G2W1*

In preparation for submission to

5.1 Abstract

5.2 Introduction

5.3 Methods

5.4 Results

5.5 Discussion

References

[1] ABDO, Z., AND GOLDING, G. A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. *Systematic Biology 56*, 1 (2007), 44–56.

[2] ADAMS, C., KNAPP, M., GEMMELL, N., JEUNEN, G.-J., BUNCE, M., LAMARE, M., AND TAYLOR, H. Beyond biodiversity: Can environmental DNA (eDNA) cut it as a population genetics tool? *Genes 10*, 192 (2019), 1.

[3] ADCOCK, C. Sample size determination: A review. *Journal of the Royal Statistical Society: Series D (The Statistician) 46*, 2 (1997), 261–283.

[4] AHRENS, D., FUJISAWA, T., KRAMMER, H.-J., EBERLE, J., FABRIZI, S., AND VOGLER, A. Rarity and incomplete sampling in DNA-based species delimitation. *Systematic Biology 65*, 3 (2016), 478–494.

[5] APRIL, J., HANNER, R. H., DION-CÔTÉ, A.-M., AND BERNATCHEZ, L. Glacial cycles as an allopatric speciation pump in north-eastern American freshwater fishes. *Molecular Ecology 22*, 2 (2013), 409–422.

[6] APRIL, J., HANNER, R. H., MAYDEN, R. L., AND BERNATCHEZ, L. Metabolic rate and climatic fluctuations shape continental wide pattern of genetic divergence and biodiversity in fishes. *PLOS ONE 8*, 7 (2013), e70296.

[7] APRIL, J., MAYDEN, R. L., HANNER, R. H., AND BERNATCHEZ, L. Genetic calibration of species diversity among North America's freshwater fishes. *Proceedings of the National Academy of Sciences 108*, 26 (2011), 10602–10607.

[8] ATHEY, T. Assessing Errors in DNA Barcode Sequence Records. Master's thesis, University of Guelph, 2013.

[9] AUSTERLITZ, F., DAVID, O., SCHAEFFER, B., BLEAKLEY, K., OLTEANU, M., LEBLOIS, R., VEUILLE, M., AND LAREDO, C. DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics 10*, 14 (2009), S10.

[10] AVISE, J. C., ARNOLD, J., BALL, R. M., BERMINGHAM, E., LAMB, T., NEIGEL, J. E., REEB, C. A., AND SAUNDERS, N. C. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics 18*, 1 (1987), 489–522.

[11] BAKER, A., SENDRA TAVARES, E., AND ELBOURNE, R. Countering criticisms of single mitochondrial DNA gene barcoding in birds. *Molecular Ecology Resources* (2009), 257–268.

[12] BECKER, S., HANNER, R., AND STEINKE, D. Five years of FISH-BOL: brief status report. *Mitochondrial DNA 22*, sup1 (2011), 3–9.

[13] BENGTSSON, B. Genetic variation in organisms with sexual and asexual reproduction. *Journal of Evolutionary Biology 16* (2003), 189.

[14] BERGSTEN, J., BILTON, D. T., FUJISAWA, T., ELLIOTT, M., MONAGHAN, M. T., BALKE, M., HENDRICH, L., GEIJER, J., HERRMANN, J., FOSTER, G. N., ET AL. The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology* (2012), sys037.

[15] BERTOLAZZI, P., FELICI, G., AND WEITSCHEK, E. Learning to classify species with barcodes. *BMC Bioinformatics 10*, 14 (2009), S7.

[16] BEVILACQUA, S., UGLAND, K. I., PLICANTI, A., SCUDERI, D., AND TERLIZZI, A. An approach based on the total-species accumulation curve and higher taxon richness to estimate realistic upper limits in regional species richness. *Ecology and Evolution 8*, 1 (2017), 405–415.

[17] BICKEL, P., GÖTZE, F., AND VAN ZWET, W. Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica 7* (1997), 1–31.

[18] BICKEL, P., AND SAKOV, A. On the choice of m in the m-out-of-n bootstrap and confidence bounds for extrema. *Statistica Sinica 18* (2008), 967–985.

[19] BRAUKMANN, T., IVANOVA, N., PROSSER, S., ELBRECHT, V., STEINKE, D., RATNASINGHAM, S., DE WAARD, J., SONES, J., ZAKHAROV, E., AND HEBERT, P. Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources* (2019), 711–727.

[20] BROWN, S. D., COLLINS, R. A., BOYER, S., LEFORT, M.-C., MALUMBRES-OLARTE, J., VINK, C. J., AND CRUICKSHANK, R. H. Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources 12*, 3 (2012), 562–565.

[21] BUCKLIN, A., STEINKE, D., AND BLANCO-BERCIAL, L. DNA barcoding of marine metazoa. *Annual Review of Marine Science 3* (2011), 471–508.

[22] CAMERON, S., RUBINOFF, D., AND WILL, K. Who will actually use DNA barcoding and what will it cost? *Systematic Biology 55*, 5 (2006), 844–847.

[23] ČANDEK, K., AND KUNTNER, M. DNA barcoding gap: reliable species identification over morphological and geographical scales. *Molecular Ecology Resources 15*, 2 (2015), 268–277.

[24] CARSTENS, B., PELLETIER, T., REID, N., AND SATLER, J. How to fail at species delimitation. *Molecular Ecology 22* (2013), 4369–4383.

[25] CASELLA, G., AND BERGER, R. *Statistical Inference*. Duxbury Thomson Learning, 2002.

[26] CEBALLOS, G., EHRLICH, P. R., BARNOSKY, A. D., GARCÍA, A., PRINGLE, R. M., AND PALMER, T. M. Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances 1*, 5 (2015), e1400253.

[27] CHAO, A. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics 11* (1984), 265–270.

[28] CHEN, F., COATES, B., HE, K., ET AL. Effects of wolbachia on mitochondrial DNA variation in populations of Athetis lepigone (Lepidoptera: Noctuidae) in China. *Mitochondrial DNA Part A 28*, 6 (1984), 826–834.

[29] CHERNICK, M. *Bootstrap Methods: A Guide for Practitioners and Researchers*. Wiley, 2007.

[30] CHIARUCCI, A., BACARO, G., RICOTTA, C., PALMER, M., AND SCHEINER, S. Spatially constrained rarefaction: incorporating the autocorrelated structure of biological communities into sample-based rarefaction. *Community Ecology 10* (2009), 209–214.

[31] CLARE, E. L., LIM, B. K., FENTON, M. B., AND HEBERT, P. D. Neotropical bats: estimating species diversity with DNA barcodes. *PLOS ONE 6*, 7 (2011), e22648.

[32] CLEMENT, M., POSADA, D., AND CRANDALL, K. A. TCS: a computer program to estimate gene genealogies. *Molecular Ecology 9*, 10 (2000), 1657–1659.

[33] COHEN, J. Things I have learned (so far). *American Psychologist 45*, 12 (1990), 1304.

[34] COLLINS, R., BOYKIN, L., CRUICKSHANK, R., AND ARMSTRONG, K. Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods in Ecology and Evolution 3* (2012), 457–465.

[35] COLLINS, R., AND CRUICKSHANK, R. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources 13*, 6 (2013), 969–975.

[36] DA COSTA, L. S., CORNELEO, N. S., AND STEFENON, V. M. Conservation of Forest Biodiversity: how sample size affects the estimation of genetic parameters. *Anais da Academia Brasileira de Ciências 87*, 2 (2015), 1095–1100.

[37] DASMAHAPATRA, K. K., ELIAS, M., HILL, R. I., HOFFMAN, J. I., AND MALLET, J. Mitochondrial DNA barcoding detects some species that are real, and some that are not. *Molecular Ecology Resources 10*, 2 (2010), 264–273.

[38] DENGLER, J. Which function describes the species–area relationship best? a review and empirical evaluation. *Journal of Biogeography 36*, 4 (2009), 728–744.

[39] DESALLE, R., AND GOLDSTEIN, P. Review and interpretation of trends in DNA barcoding. *Frontiers in Ecology and Evolution 7*, 302 (2019), 1–11.

[40] DI STEFANO, J. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Functional Ecology 17*, 5 (2003), 707–709.

[41] DIXON, C. J. A means of estimating the completeness of haplotype sampling using the Stirling probability distribution. *Molecular Ecology Notes 6*, 3 (2006), 650–652.

[42] DOORENWEERD, C., SAN JOSE, M., BARR, N., LEBLANC, L., AND RUBINOFF, D. Highly variable COI haplotype diversity between three species of invasive pest fruit fly reflects remarkably incongruent demographic histories. *Scientific Reports 10*, 1 (2020), 1–10.

[43] DRUMMOND, A., AND RAMBAUT, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology 7*, 214 (2007), 1–8.

[44] EDDELBUETTEL, D., AND FRANÇOIS, R. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software 40*, 8 (2011), 1–18.

[45] EDDELBUETTEL, D., AND SANDERSON, C. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis 71* (2014), 1054–1063.

[46] EDGAR, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics 26*, 19 (2010), 2460–2461.

[47] EFRON, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* (1979), 1–26.

[48] EFRON, B., HALLORAN, E., AND HOLMES, S. Bootstrap confidence levels for phylogenetic trees. *PNAS 93*, 23 (1996), 13429–13434.

[49] ELBRECHT, V., VAMOS, E. E., STEINKE, D., AND LEESE, F. Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ 6* (2018), e4644.

[50] EWENS, W. J. The sampling theory of selectively neutral alleles. *Theoretical Population Biology 3*, 1 (1972), 87–112.

[51] EZARD, T., FUJISAWA, T., AND BARRACLOUGH, T. *splits: SPecies' LImits by Threshold Statistics*, 2017. R package version 1.0-19/r52.

[52] FELENSTEIN, J. The evolutionary advantage of recombination. *Genetics 78*, 2 (1974), 737–756.

[53] FELENSTEIN, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution 39*, 4 (1985), 783–791.

[54] FIEBERG, J., VITENSE, K., AND JOHNSON, D. Resampling-based methods for biologists. *PeerJ 8* (2020), e9089.

[55] FREEDMAN, D., AND DIACONIS, P. On the histogram as a density estimator: L_2 theory. *Z. Wahrscheinlichkeitstheorie verw. Gebiete 57* (1981), 453–476.

[56] FUJISAWA, T., ASWAD, A., AND BARRACLOUGH, T. A rapid and scalable method for multilocus species delimitation using Bayesian model comparison and rooted triplets. *Systematic Biology 65*, 5 (2016), 759?771.

[57] FUJISAWA, T., AND BARRACLOUGH, T. Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: a revised method and evaluation on simulated data sets. *Systematic Biology 62*, 5 (2013), 707–724.

[58] FUNK, D. J., AND OMLAND, K. E. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics 34*, 1 (2003), 397–423.

[59] GOOD, P., AND HARDIN, J. *Common Errors in Statistics (And How to Avoid Them)*. John Wiley & Sons, Inc., 2003.

[60] GOODALL-COPESTAKE, W., TARLING, G., AND MURPHY, E. On the comparison of population-level estimates of haplotype and nucleotide diversity: a case study using the gene cox1 in animals. *Heredity 109*, 1 (2012), 50–56.

[61] GOTELLI, N. J., AND COLWELL, R. K. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters 4*, 4 (2001), 379–391.

[62] GREWE, P. M., KRUEGER, C. C., AQUADRO, C. F., BERMINGHAM, E., KINCAID, H. L., AND MAY, B. Mitochondrial DNA variation among lake trout (*Salvelinus namaycush*) strains stocked into Lake Ontario. *Canadian Journal of Fisheries and Aquatic Sciences 50*, 11 (1993), 2397–2403.

[63] GRIMM, V., BERGER, U., BASTIANSEN, F., ELIASSEN, S., GINOT, V., GISKE, J., GOSS-CUSTARD, J., GRAND, T., HEINZ, S., HUSE, G., HUTH, A., JEPSEN, J., JØRGENSEN, C., MOOIJ, W., MÜLLER, B., PE'ER, G., PIOU, C., RAILSBACK, S., ROBBINS, A., ROBBINS, M., ROSSMANITH, E., RÜGER, N., STRAND, E., SOUISSI, S., STILLMAN, R., VABØ, R., VISSER, U., AND DEANGELIS, D. A standard protocol for describing individual-based and agent-based models. *Ecological Modelling 198*, 1 (2006), 115–126.

[64] GRIMM, V., BERGER, U., DEANGELIS, D., POLHILL, J., GISKE, J., AND RAILSBACK, S. The ODD protocol: A review and first update. *Ecological Modelling 221*, 23 (2010), 2760–2768.

[65] GWIAZDOWSKI, R. A., ELKINTON, J. S., DEWAARD, J. R., AND SREMAC, M. Phylogeographic diversity of the winter moths Operophtera brumata and O. bruceata (Lepidoptera: Geometridae) in Europe and North America. *Annals of the Entomological Society of America 106*, 2 (2013), 143–151.

[66] HAJIBABAEI, M., SINGER, G. A., HEBERT, P. D., AND HICKEY, D. A. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *TRENDS in Genetics 23*, 4 (2007), 167–172.

[67] HALE, M. L., BURG, T. M., AND STEEVES, T. E. Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PLOS ONE 7*, 9 (2012), e45170.

[68] HANNER, R. Data standards for BARCODE records in INSDC (BRIs).

[69] HANNER, R., BECKER, S., IVANOVA, N. V., AND STEINKE, D. FISH-BOL and seafood identification: Geographically dispersed case studies reveal systemic market substitution across Canada. *Mitochondrial DNA 22*, sup1 (2011), 106–122.

[70] HANNER, R., FLOYD, R., BERNARD, A., COLLETTE, B. B., AND SHIVJI, M. DNA barcoding of billfishes. *Mitochondrial DNA 22*, sup1 (2011), 27–36.

[71] HANNER, R. H., NAAUM, A. M., AND SHIVJI, M. S. Conclusion: DNA-Based Authentication of Shark Products and Implications for Conservation and Management. In *Seafood Authenticity and Traceability: A DNA-based Perspective*, A. M. Naaum and R. H. Hanner, Eds., 1 ed. Academic Press, 2016.

[72] HARRIS, D. J. Can you bank on GenBank? *Trends in Ecology & Evolution 18*, 7 (2003), 317–319.

[73] HART, M. W., AND SUNDAY, J. Things fall apart: biological species form unconnected parsimony networks. *Biology Letters 3*, 5 (2007), 509–512.

[74] HAUSMANN, A., GODFRAY, H. C. J., HUEMER, P., MUTANEN, M., ROUGERIE, R., VAN NIEUKERKEN, E. J., RATNASINGHAM, S., AND HEBERT, P. D. Genetic patterns in European geometrid moths revealed by the Barcode Index Number (BIN) system. *PLOS ONE 8*, 12 (2013), e84518.

[75] HEBERT, P. D., CYWINSKA, A., BALL, S. L., ET AL. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences 270*, 1512 (2003), 313–321.

[76] HEBERT, P. D., PENTON, E., BURNS, J., JANZEN, D., AND HALLWACHS, W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences 101*, 41 (2004), 14812–14817.

[77] HEBERT, P. D., RATNASINGHAM, S., AND DE WAARD, J. R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences 270*, Suppl 1 (2003), S96–S99.

[78] HEBERT, P. D., RATNASINGHAM, S., AND ZAKHAROV, E. Counting animal species with dna barcodes: Canadian insects. *Philosophical Transactions of the Royal Society B 371* (2016), 20150333.

[79] HEBERT, P. D., STOECKLE, M. Y., ZEMLAK, T. S., AND FRANCIS, C. M. Identification of birds through DNA barcodes. *PLOS Biology 2*, 10 (2004), e312.

[80] HICKERSON, M. J., MEYER, C. P., AND MORITZ, C. DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology 55*, 5 (2006), 729–739.

[81] HOBERN, D. BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. *Genome* (2020), 1–4.

[82] HOBERN, D., AND HEBERT, P. BIOSCAN - revealing eukaryote diversity, dynamics, and interactions. *Biodiversity Information Science and Standards 3* (2019), e37333.

[83] HOLT, J. A., STONEBERG HOLT, S. D., AND BUREŠ, P. Experimental design in intraspecific organelle DNA sequence studies III: statistical measures of sampling success. *Taxon 56*, 3 (2007), 847–856.

[84] HORTAL, J., AND LOBO, J. M. An ED–based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation 14*, 12 (2005), 2913–2947.

[85] HUBERT, N., AND HANNER, R. DNA Barcoding, species delineation and taxonomy: a historical perspective. *DNA Barcodes 3*, 1 (2015), 44–58.

[86] HUBERT, N., HANNER, R., HOLM, E., MANDRAK, N. E., TAYLOR, E., BURRIDGE, M., WATKINSON, D., DUMONT, P., CURRY, A., BENTZEN, P., ZHANG, J., APRIL, J., AND BERNATCHEZ, L. Identifying Canadian freshwater fishes through DNA barcodes. *PLOS ONE 3*, 6 (2008), e2490.

[87] HUELSENBECK, J., AND RONQUIST, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics 17*, 8 (2001), 754–755.

[88] HUEMER, P., MUTANEN, M., SEFC, K. M., AND HEBERT, P. D. Testing DNA barcode performance in 1000 species of European Lepidoptera: Large geographic distances have small genetic impacts. *PLOS ONE 9*, 12 (2014), e115774.

[89] HUNTER, M., OYLER-MCCANCE, S., DORAZIO, R., ET AL. Environmental DNA (eDNA) Sampling Improves Occurrence and Detection Estimates of Invasive Burmese Pythons. *PLOS ONE 10*, 4 (2015), e0121655.

[90] HYNDMAN, R. The problem with Sturges rule for constructing histograms. Unpublished, 1995.

[91] JIN, Q., HE, L.-J., AND ZHANG, A.-B. A simple 2D non-parametric resampling statistical approach to assess confidence in species identification in DNA barcoding–an alternative to Likelihood and Bayesian approaches. *PLOS ONE 7*, 12 (2012), e50831.

[92] JOLY, S., STEVENS, M. I., AND VAN VUUREN, B. J. Haplotype networks can be misleading in the presence of missing data. *Systematic Biology 56*, 5 (2007), 857–862.

[93] JONES, G. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of Mathematical Biology 74* (2017), 447–467.

[94] KEKKONEN, M., AND HEBERT, P. DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. *Molecular Ecology Resources 14* (2014), 706–714.

[95] KIMURA, M., AND WEISS, G. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics 49*, 4 (1964), 561–576.

[96] KINGMAN, J. F. C. The coalescent. *Stochastic Processes and their Applications 13*, 3 (1982), 235–248.

[97] KOLTER, A., AND GEMEINHOLZER, B. Plant DNA barcoding necessitates marker-specific efforts to establish more comprehensive reference databases. *Genome* (2020).

[98] KOROIVA, R., AND KVIST, S. Estimating the DNA barcoding gap in a global dataset of cox1 sequences for Odonata: Close, but no cigar. *Mitochondrial DNA 29*, 5 (2018), 765–771.

[99] KUMAR, S., STECHER, G., AND TAMURA, K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* (2016), 1870–1874.

[100] KVIST, S. Does a global barcoding gap exist in Annelida? *Mitochondrial DNA Part A* (2017), 2241–2252.

[101] LAVINIA, P., KERR, K., TUBARO, P., HEBERT, P., AND LIJTMAER, D. Calibrating the molecular clock beyond cytochrome b: assessing the evolutionary rate of COI in birds. *Journal of Avian Biology 47* (2016), 86–91.

[102] LAYTON, K., MARTEL, A., AND HEBERT, P. Patterns of DNA barcode variation in Canadian marine molluscs. *PLOS ONE 9*, 4 (2014), e95003.

[103] LEIGH, J. W., AND BRYANT, D. POPART: full-feature software for haplotype network construction. *Methods in Ecology and Evolution 6*, 9 (2015), 1110–1116.

[104] LENTH, R. V. Some practical guidelines for effective sample size determination. *The American Statistician 55*, 3 (2001), 187–193.

[105] LINDBLOM, L. Sample size and haplotype richness in population samples of the lichen-forming ascomycete Xanthoria parietina. *The Lichenologist 41*, 05 (2009), 529–535.

[106] LINDLEY, D. The philiosophy of statistics. *The Statistician 49*, 3 (2000), 293–337.

[107] LIU, J., PROVAN, J., GAO, L.-M., AND LI, D.-Z. Sampling strategy and potential utility of indels for DNA barcoding of closely related plant species: a case study in Taxus. *International Journal of Molecular Sciences 13*, 7 (2012), 8740–8751.

[108] LOHSE, K. Can mtDNA barcodes be used to delimit species? a response to Pons et al. (2006). *Systematic Biology 58*, 4 (2009), 439–442.

[109] LOU, M., AND GOLDING, G. Assigning sequences to species in the absence of large interspecific differences. *Molecular Phylogenetics and Evolution 58* (2010), 187–194.

[110] LOU, M., AND GOLDING, G. B. The effect of sampling from subdivided populations on species identification with DNA barcodes using a Bayesian statistical approach. *Molecular Phylogenetics and Evolution 65*, 2 (2012), 765–773.

[111] LUO, A., LAN, H., LING, C., ZHANG, A.-B., SHI, L., HO, S. Y., AND ZHU, C. A simulation study of sample size for DNA barcoding. *Ecology and Evolution 5*, 24 (2015), 5869–5879.

[112] LUO, A., LING, C., HO, S., AND ZHU, C.-D. Comparison of methods for molecular species delimitation across a range of speciation scenarios. *Systematic Biology 67*, 5 (2018), 830–846.

[113] MADDEN, M., YOUNG, R., BROWN, J., MILLER, S., FREWIN, A., AND HANNER, R. Using DNA barcoding to improve invasive pest identification at U.S. ports-of-entry. *PLOS ONE 14*, 9 (2019), e0222291.

[114] MARTIN, M., DANIËLS, P., D, E., AND SPOUGE, J. Figures of merit and statistics for detecting faulty species identification with DNA barcodes: A case study in Ramaria and related fungal genera. *PLOS ONE 15*, 8 (2020), e0237507.

[115] MATZ, M. V., AND NIELSEN, R. A likelihood ratio test for species membership based on DNA sequence data. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 360*, 1462 (2005), 1969–1974.

[116] MEIER, R., SHIYANG, K., VAIDYA, G., AND NG, P. DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology 55*, 5 (2006), 715–728.

[117] MEIER, R., ZHANG, G., AND ALI, F. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. *Systematic Biology 57*, 5 (2008), 809–813.

[118] MEYER, C. P., AND PAULAY, G. DNA barcoding: error rates based on comprehensive sampling. *PLOS Biology 3*, 12 (2005), e422.

[119] MIN, X., AND HICKEY, D. Assessing the effect of varying sequence length on DNA barcoding of fungi. *Molecular Ecology Notes 7* (2007), 365–373.

[120] MONAGHAN, M., WILD, R., ELLIOT, M., ET AL. Accelerated species inventory on madagascar using coalescent-based models of species delineation. *Systematic Biology 58*, 3 (2009), 298–311.

[121] MUIRHEAD, J. R., GRAY, D. K., KELLY, D. W., ELLIS, S. M., HEATH, D. D., AND MACISAAC, H. J. Identifying the source of species invasions: sampling intensity vs. genetic diversity. *Molecular Ecology 17*, 4 (2008), 1020–1035.

[122] MULLER, H. The relation of recombination to mutational advance. *Mutation Research 1*, 1 (1964), 2–9.

[123] MUTANEN, M., KIVELÄ, S. M., VOS, R. A., DOORENWEERD, C., RATNASINGHAM, S., HAUSMANN, A., HUEMER, P., DINCĂ, V., VAN NIEUKERKEN, E. J., LOPEZ-VAAMONDE, C., ET AL. Species-level para-and polyphyly in DNA barcode gene trees: strong operational bias in European Lepidoptera. *Systematic Biology 65*, 6 (2016), 1024–1040.

[124] NAAUM, A., SHEHATA, H., CHEN, S., LI, J., TABUJARA, N., AWMACK, D., LUTZE-WALLACE, C., AND HANNER, R. Complementary molecular methods detect undeclared species in sausage products at retail markets in Canada. *Food Control 84* (2018), 339–344.

[125] NAAUM, A. M., ST JAQUES, J., WARNER, K., SANTSCHI, L., IMONDI, R., AND HANNER, R. Standards for conducting a DNA barcoding market survey: Minimum information and best practices. *DNA Barcodes 3*, 1 (2015), 80–84.

[126] NAZARENO, A. G., BEMMELS, J. B., DICK, C. W., AND LOHMANN, L. G. Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Molecular Ecology Resources 17*, 6 (2017), 1136–1147.

[127] NEI, M. *Molecular Evolutionary Genetics*. Columbia University Press, 1987.

[128] NEI, M., AND LI, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences 76*, 10 (1979), 5269–5273.

[129] NIELSEN, R., AND MATZ, M. Statistical approaches for DNA barcoding. *Systematic Biology 55*, 1 (2006), 162–169.

[130] ONDREJICKA, D. A., LOCKE, S. A., MOREY, K., BORISENKO, A. V., AND HANNER, R. H. Status and prospects of DNA barcoding in medically important parasites and vectors. *Trends in Parasitology 30*, 12 (2014), 582–591.

[131] ONDREJICKA, D. A., MOREY, K., AND HANNER, R. H. DNA barcodes identify medically important tick species in Canada. *Genome 60*, 1 (2017), 74–84.

[132] OVERDYK, L. M., BRAID, H. E., CRAWFORD, S. S., AND HANNER, R. H. Extending DNA barcoding coverage for Lake Whitefish (Coregonus clupeaformis) across the three major basins of Lake Huron. *DNA Barcodes 3*, 1 (2015), 59–65.

[133] PANTE, E., PUILLANDRE, N., VIRICEL, A., ARNAUD-HAOND, S., AURELLE, D., CASTELIN, M., CHENUIL, A., DESTOMBE, C., FORCIOLI, D., VALERO, M., ET AL. Species are hypotheses: avoid connectivity assessments based on pillars of sand. *Molecular Ecology 24*, 3 (2015), 525–544.

[134] PAPADOPOULOU, A., MONAGHAN, M., BARRACLOUGH, T., ET AL. Sampling error does not invalidate the yule-coalescent model for species delimitation. a response to Lohse (2009). *Systematic Biology 58*, 4 (2009), 442–444.

[135] PARADIS, E. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics 26*, 3 (2010), 419–420.

[136] PARADIS, E., CLAUDE, J., AND STRIMMER, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics 20*, 2 (2004), 289–290.

[137] PARR, C. S., GURALNICK, R., CELLINESE, N., AND PAGE, R. D. Evolutionary informatics: unifying knowledge about the diversity of life. *Trends in Ecology & Evolution 27*, 2 (2012), 94–103.

[138] PARZEN, E. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics 33*, 3 (1962), 1065–1076.

[139] PEARSON, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A 185* (1894), 71–110.

[140] PENTINSAARI, M., HEBERT, P. D., AND MUTANEN, M. Barcoding beetles: a regional survey of 1872 species reveals high identification success and unusually deep interspecific divergences. *PLOS ONE 9*, 9 (2014), e108651.

[141] PFENNINGER, M., BÁLINT, M., AND PAULS, S. Methodological framework for projecting the potential loss of intraspecific genetic diversity due to global climate change. *BMC Evolutionary Biology 12*, 224 (2012), 1–13.

[142] PHILLIPS, J. D., GILLIS, D. J., AND HANNER, R. H. Incomplete estimates of genetic diversity within species: Implications for DNA barcoding. *Ecology and Evolution 9*, 5 (2019), 2996–3010.

[143] PHILLIPS, J. D., GILLIS, D. J., AND HANNER, R. H. `HACSim`: An R package to estimate intraspecific sample sizes for genetic diversity assessment using haplotype accumulation curves. *PeerJ Computer Science* (2020).

[144] PHILLIPS, J. D., GWIAZDOWSKI, R. A., ASHLOCK, D., AND HANNER, R. An exploration of sufficient sampling effort to describe intraspecific DNA barcode haplotype diversity: examples from the ray–finned fishes (Chordata: Actinopterygii). *DNA Barcodes 3*, 1 (2015), 66–73.

[145] POLITIS, D., ROMANO, J., AND WOLF, M. *Subsampling*. Springer, 1999.

[146] PONS, J., BARRACLOUGH, T., GOMEZ-ZURITA, ET AL. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology 55*, 4 (2006), 595–609.

[147] PRUETT, C., AND WINKER, K. The effects of sample size on population genetic diversity estimates in song sparrows Melospiza melodia. *Journal of Avian Biology 39* (2008), 252–256.

[148] PUILLANDRE, N., LAMBERT, A., AND BROUILLET, S. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology 21* (2011), 1864–1877.

[149] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[150] RATNASINGHAM, S., AND HEBERT, P. D. BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes 7*, 3 (2007), 355–364.

[151] RATNASINGHAM, S., AND HEBERT, P. D. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLOS ONE 8*, 7 (2013), e66213.

[152] ROGNES, T., FLOURI, T., NICHOLS, B., QUINCE, C., AND MAHÉ, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ 4* (2016), e2584.

[153] ROSENBERG, N., AND NORDBORG, M. Genealogical trees, coalecent theory and the analysis of genetic polymorphisms. *Nature Genetics Reviews 3* (2002), 380–390.

[154] ROSENBLATT, M. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics 27*, 3 (1956), 832–837.

[155] ROSS, H. A., MURUGAN, S., AND LI, W. L. S. Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology 57*, 2 (2008), 216–230.

[156] RYAN, K., AND CRAWFORD, S. Distribution and abundance of larval lake whitefish (Coregonus clupeaformis) in Stokes Bay, Lake Huron. *Journal of Great Lakes Research 40* (2014), 755–762.

[157] SCOTT, D. On optimal and data-based histograms. *Biometrika 66*, 3 (1979), 605–610.

[158] SCRUCCA, L., FOP, M., MURPHY, T., AND RAFTERY, A. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal 8* (2016), 205–233.

[159] SERRAO, N., STEINKE, D., AND HANNER, R. Calibrating snakehead diversity with DNA barcodes: Expanding taxonomic coverage to enable identification of potential and established invasive species. *PLOS ONE 9*, 6 (2014), e99546.

[160] SHEHATA, H., BOURQUE, D., STEINKE, D., CHEN, S., AND HANNER, R. Survey of mislabelling across finfish supply chain reveals mislabelling both outside and within Canada. *Food Research International 121* (2019), 723–729.

[161] SHEHATA, H., NAAUM, A., CHEN, S., MURPHY, T., LI, J., SHANNON, K., AWMACK, D., LOCAS, A., AND HANNER, R. Re-visiting the occurrence of undeclared species in sausage products sold in Canada. *Food Research International 122* (2019), 593–598.

[162] SHEHATA, H., NAAUM, A., GARDUNO, R., AND HANNER, R. DNA barcoding as a regulatory tool for seafood authentication in Canada. *Food Control 92* (2018), 147–153.

[163] SMITH, M., BERTRAND, C., CROSBY, K., ET AL. Wolbachia and DNA barcoding insects: Patterns, potential and problems. *PLOS ONE 7*, 5 (2012), e36514.

[164] SONET, G., JORDAENS, K., NAGY, Z. T., BREMAN, F. C., DE MEYER, M., BACKELJAU, T., AND VIRGILIO, M. Adhoc: an R package to calculate ad hoc distance thresholds for DNA barcoding identification. *Zookeys 365* (2013), 329–336.

[165] SPALL, J. C. Stochastic Optimization. In *Handbook of Computational Statistics: Concepts and Methods*, J. E. Gentle, W. K. Härdle, and Y. Mori, Eds., 2 ed. Springer, 2012.

[166] SPOUGE, J., AND MARIÑO-RAMIREZ, L. The practical evaluation of DNA barcode efficacy. In *DNA Barcodes: Methods and Protocols*, W. Kress and D. Erickson, Eds., 1 ed. 2012.

[167] SRIVATHSAN, A., AND MEIER, R. On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics 28* (2012), 190–194.

[168] STEIN, E. D., MARTINEZ, M. C., STILES, S., MILLER, P. E., AND ZAKHAROV, E. V. Is DNA barcoding actually cheaper and faster than traditional morphological methods: Results from a survey of freshwater bioassessment efforts in the United States? *PLOS ONE 9*, 4 (2014), e95525.

[169] STEINKE, D., BERNARD, A. M., HORN, R. L., HILTON, P., HANNER, R., AND SHIVJI, M. S. DNA analysis of traded shark fins and mobulid gill plates reveals a high proportion of species of conservation concern. *Scientific Reports 7*, 9505 (2017), 1–6.

[170] STEINKE, D., AND HANNER, R. The FISH-BOL collaborators' protocol. *Mitochondrial DNA 22*, sup1 (2011), 10–14.

[171] STEINKE, D., ZEMLAK, T. S., BOUTILLIER, J. A., AND HEBERT, P. D. DNA barcoding of Pacific Canada's fishes. *Marine Biology 156*, 12 (2009), 2641–2647.

[172] STEINKE, D., ZEMLAK, T. S., AND HEBERT, P. D. Barcoding Nemo: DNA-based identifications for the ornamental fish trade. *PLOS One 4*, 7 (2009), e6300.

[173] STOECKLE, M. Y., AND KERR, K. C. Frequency matrix approach demonstrates high sequence quality in avian BARCODEs and highlights cryptic pseudogenes. *PLOS ONE 7*, 8 (2012), e43992.

[174] STOECKLE, M. Y., AND THALER, D. S. DNA barcoding works in practice but not in (neutral) theory. *PLOS ONE 9*, 7 (2014), e100755.

[175] STROHM, J. H., GWIAZDOWSKI, R. A., AND HANNER, R. Mitogenome metadata: current trends and proposed standards. *Mitochondrial DNA Part A 27*, 5 (2016), 3263–3269.

[176] STURGES, H. The choice of a class interval. *Journal of the American Statistical Association 21* (1926), 65–66.

[177] TAJIMA, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics 105*, 2 (1983), 437–460.

[178] TAYLOR, H., AND HARRIS, W. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources 12*, 3 (2012), 377–388.

[179] TEMPLETON, A. R., CRANDALL, K. A., AND SING, C. F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics 132*, 2 (1992), 619–633.

[180] TERLIZZI, A., ANDERSON, M. J., BEVILACQUA, S., AND UGLAND, K. I. Species-accumulation curves and taxonomic surrogates: an integrated approach for estimation of regional species richnesss. *Diversity and Distributions 20* (2014), 356–368.

[181] TJØRVE, E. Shapes and functions of species–area curves: a review of possible models. *Journal of Biogeography 30*, 6 (2003), 827–835.

[182] TURON, X., ANTICH, A., PALACIN, C., PRÆBEL, K., AND WANGERSTEEN, O. From metabarcoding to metaphylogeography: separating the wheat from the chaff. *bioRxiv* (2019).

[183] WARD, R. D., HANNER, R., AND HEBERT, P. D. The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology 74*, 2 (2009), 329–356.

[184] WARD, R. D., ZEMLAK, T. S., INNES, B. H., LAST, P. R., AND HEBERT, P. D. DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 360*, 1462 (2005), 1847–1857.

[185] WARES, J. P., AND PAPPALARDO, P. Can Theory Improve the Scope of Quantitative Metazoan Metabarcoding? *Diversity 8*, 1 (2015), 1.

[186] WASSERSTEIN, R., SCHIRM, A., AND LAZR, N. Moving to a world beyond "p $<$ 0.05". *The American Statistician 73*, 51 (2019), 1–19.

[187] WATTERSON, G. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology 7*, 2 (1975), 256–276.

[188] WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[189] WIEMERS, M., AND FIEDLER, K. Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology 4*, 8 (2007).

[190] WILKINSON, M., SZABO, C., FORD, C., YAROM, Y., CROXFORD, A., CAMP, A., AND GOODING, P. Replacing Sanger with Next Generation Sequencing to improve coverage and quality of reference DNA barcodes for plants. *Scientific Reports 7* (2017), 46040.

[191] WILLIAMS, P., BYVALTSEV, A., CEDERBERG, B., BEREZIN, M., ØDEGAARD, F., RASMUSSEN, C., RICHARDSON, L., HUANG, J., SHEFFIELD, C., AND WILLIAMS, S. Genes suggest ancestral colour polymorphisms are shared across morphologically cryptic species in arctic bumblebees. *PLOS ONE 10*, 12 (2015), e0144544.

[192] WILLIAMS, P. H., HUANG, J., RASMONT, P., AND AN, J. Early-diverging bumblebees from across the roof of the world: the high-mountain subgenus Mendacibombus revised from species gene coalescents and morphology (Hymenoptera, Apidae). *Zootaxa 4204*, 1 (2016), 1–72.

[193] WONG, E. H.-K., SHIVJI, M. S., AND HANNER, R. H. Identifying sharks with DNA barcodes: assessing the utility of a nucleotide diagnostic approach. *Molecular Ecology Resources 9*, s1 (2009), 243–256.

[194] WRIGHT, S. The genetical structure of populations. *Annals of Eugenics 15*, 1 (1951), 323–354.

[195] YAO, PENG-CHENG, G., HAI-YAN, W., YA-NAN, ZHANG, J.-H., CHEN, X.-Y., AND LI, H.-Q. Evaluating sampling strategy for DNA barcoding study of coastal and inland halo-tolerant Poaceae and Chenopodiaceae: A case study for increased sample size. *PLOS ONE 12*, 9 (2017), e0185311.

[196] YOUNG, M. R., BEHAN-PELLETIER, V. M., AND HEBERT, P. D. Revealing the hyperdiverse mite fauna of subarctic Canada through DNA barcoding. *PLOS ONE 7*, 11 (2012), e48755.

[197] YOUNG, R., ABOTT, C., THERRIAULT, T., AND ADAMOWICZ, S. Barcode-based species delimitation in the marine realm: a test using Hexanauplia (Multicrustacea: Thecostraca and Copepoda). *Genome 60*, 2 (2017), 169–182.

[198] ZHANG, A., HAO, M., YANG, C., AND SHI, Z. BarcodingR: an integrated R package for species identification using DNA barcodes. *Methods in Ecology and Evolution* (2016), 1–8.

[199] ZHANG, A.-B., HE, L.-J., CROZIER, R. H., MUSTER, C., AND ZHU, C.-D. Estimating sample sizes for DNA barcoding. *Molecular Phylogenetics and Evolution 54*, 3 (2010), 1035–1039.

[200] ZHANG, A.-B., MUSTER, C., ZHU, C.-D., CROZIER, R., WAN, P., FENG, J., AND WARD, R. A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Molecular Ecology 21*, 8 (2012), 1848–1863.

[201] ZHANG, H.-G., LV, M.-H., YI, W.-B., ZHU, W.-B., AND BU, W.-J. Species diversity can be overestimated by a fixed empirical threshold: insights from DNA barcoding of the genus Cletus (Hemiptera: Coreidae) and the meta-analysis of COI data from previous phylogeographical studies. *Molecular Ecology Resources 17* (2017), 314–323.

[202] ZHANG, J., KAPLI, P., PAVLIDIS, P., AND STAMATAKIS, P. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics 29*, 22 (2013), 2869–2876.