Marine Mammal Science

MARINE MAMMAL SCIENCE, 33(SPECIAL ISSUE): 101–131 (June 2017)
Published 2017. This article is a U.S. Government work and is in the public domain in the U.S.A.
DOI: 10.1111/mms.12414

Diagnosability of mtDNA with Random Forests: Using sequence data to delimit subspecies

This is the fifth of six papers forming a special issue of Marine Mammal Science (Vol. 33, Special Issue) on delimiting cetacean subspecies using primarily genetic data. An introduction to the special issue and brief summaries of all papers it contains is presented in Taylor et al. (2017b). Together, these papers lead to a proposed set of guidelines that identify informational needs and quantitative standards (Taylor et al. 2017a) intended to promote consistency, objectivity, and transparency in the classification of cetaceans. The guidelines are broadly applicable across data types. The quantitative standards are based on the marker currently available across a sufficiently broad number of cetacean taxa: mitochondrial DNA control region sequence data. They are intended as "living" standards that should be revised as new types of data (particularly nuclear data) become available.

Frederick I. Archer,[1] **Karen K. Martien, Barbara L. Taylor,** Marine Mammal and Turtle Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 8901 La Jolla Shores Drive, La Jolla, California 92037, U.S.A.

Abstract

We examine the use of an ensemble method, Random Forests, to delimit subspecies using mitochondrial DNA (mtDNA) sequences. Diagnosability, a measure of the ability to correctly determine the taxon of a specimen of unknown origin, has historically been used to delimit subspecies, but few studies have explored how to estimate it from DNA sequences. Using simulated and empirical data sets, we demonstrate that Random Forests produces classification models that perform well for diagnosing subspecies and species. Populations with strong social structure and relatively low abundances (*e.g.*, killer whales, *Orcinus orca*) were found to be as diagnosable as species. Conversely, comparisons involving subspecies that are abundant (*e.g.*, spinner and spotted dolphins, *Stenella longirostris* and *S. attenuata*), are only as diagnosable as many population comparisons. Estimates of diagnosability reported in subspecies and species descriptions should include confidence intervals, which are influenced by the sample sizes of the training data. We also stress the importance of reporting the certainty with which individuals in the training data are classified in order to communicate the strength of the classification model and diagnosability estimate. Guidance as to ideal minimum diagnosability thresholds for subspecies will improve with more comprehensive analyses; however, values in the range of 80%–90% are considered appropriate.

Key words: taxonomy, subspecies, mtDNA, random forests, machine learning, species, population genetics, systematics, classification.

In the beginning of his book outlining the principles of systematics and taxonomy, Ernst Mayr (1969) defines taxonomy as, "the theory and practice of classifying

[1]Corresponding author (e-mail: eric.archer@noaa.gov).

organisms." This emphasis on classification at the root of one of the central fields of biology reflects the way we naturally conceptualize the world around us. Our tendency to group items based on shared features is a practice that facilitates information retrieval, comparisons and contrasts, and provides a basis for future hypotheses (Castro and Toro 1995). Taxonomy is merely a formalization of this natural process of making sense of our surroundings.

Species are the basic units of systematics and taxonomy, and there is a rich history of ways to define, describe, and delimit them, which have been organized into formal species concepts (Zink and Davis 1999, Lee 2003, Sites and Marshall 2004, de Queiroz 2007). Diagnosability is at the heart of many of these concepts and is also, practically speaking, often an operational necessity for species delimitation (Mayr 1969, Li *et al.* 2006, Brambilla *et al.* 2009). The Phylogenetic Species Concept (PSC), which defines a species as "the smallest diagnosable cluster of individual organisms within which there is a parental pattern of ancestry and descent" (Cracraft 1983) is one of the more commonly used species concepts, especially when using genetic data for taxonomy. Under the PSC, there is an expectation that all members of a species are fully diagnosable (Baum and Donoghue 1995, Wheeler 1999, Helbig 2002), meaning that in the suite of distinguishing characteristics used for the diagnoses (*e.g.*, morphological, genetic, *etc*.), there can be no overlap with other species. In other words, assignments for species must be unambiguous and without error.[2]

The above requirement for complete diagnosability of species is in contrast to the use of diagnosability with regard to subspecies, where it is also a central and defining characteristic (Mayr 1942, Amadon 1949, Barrowclough 1982, Remsen 2010).[3] Subspecies reside immediately below species and are the smallest named taxonomic unit. Taylor *et al.* (2017*b*) define a *subspecies* as a "population, or collection of populations, that *appears to be* a separately evolving lineage with discontinuities resulting from geography, ecological specialization, or other forces that restrict gene flow to the point that the population or collection of populations is diagnosably distinct." A part of this definition is intended to recognize that although gene flow has been restricted, it may still be occurring at low levels, and as such, some degree of character overlap is expected (Amadon 1949, Patten and Unitt 2002, Patten 2010). Unlike species, subspecies can be partially diagnosable. Thus, important questions for subspecies delimitation are what level of diagnosability is sufficient, and how it is best measured (Patten 2010).

The primary quantitative guidance for diagnosability of subspecies has been the "75% rule" (Amadon 1949, Patten and Unitt 2002). The rule as described by Amadon (1949), requires that for two putative subspecies to be considered diagnosable, at least 75% of the distribution of a given character for each must lie outside of at least 99% of the distribution of the other (Fig. 1A). Amadon (1949) also demonstrated that if the character being used for the diagnosis could be described by a Normal distribution, this is equivalent to requiring that 97% of one distribution lies outside of 97% of the other. The point where the two distributions overlap identifies a threshold value for the character, or the point at which an individual would have an equal chance of belonging to either subspecies (Fig. 1B). It should be kept in mind that

[2]Diagnostic characters for species need not apply to all members of a species at any life stage. Most species concepts recognize that diagnostic characters may be different for different ages, stages, or sexes, or nonexistent altogether in some classes (Helbig *et al*. 2002).

[3]Strict adherents to the Phylogenetic Species Concept do not recognize the taxonomic rank of subspecies, because taxa are either fully diagnosable, and thus elevated to species, or not diagnosable and have no status (McKitrick and Zink 1988, Cracraft 1992).

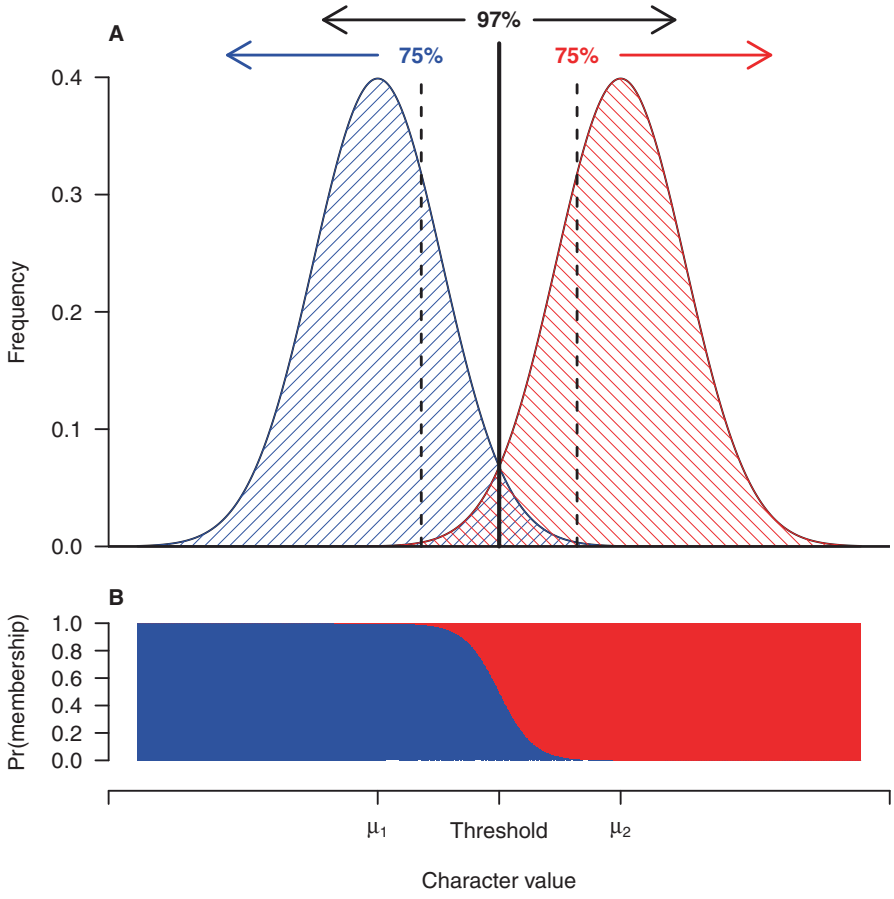

Figure 1. (A) Distribution of a hypothetical character for two putative subspecies (red and blue) demonstrating minimum overlap necessary to satisfy 75% rule of Amadon (1949). Character is continuous on the *x*-axis. Dashed lines indicate the point at which 75% of each distribution is outside of 99%+ of the other. Solid line indicates point of overlap where 97% of both distributions are outside one another. (B) Probability of membership to subspecies for specimens having values along the character axis. Probability is based on the ratio of the distribution frequencies at each point along the *x*-axis, with a 50:50 probability occurring at the threshold point.

both formulations are different ways of describing the minimum acceptable overlap between the two distributions.

Diagnosability in general, and the 75% rule in particular, has been interpreted in multiple ways by various authors (Wilson and Brown 1953, Patten and Unitt 2002, Remsen 2010). Most of these interpretations can be categorized in one of two ways. The first is an evaluation of the degree of overlap in the distributions of the putative subspecies for a particular character. Representing the second form of interpretation, Remsen (2010) argues that diagnosability should have the individual, rather than the population, as its unit of analysis. Under this paradigm, focus is placed on estimating how distinct individuals are rather than the degree of differentiation among groups

of individuals. In this usage, diagnosability is satisfied if some minimum percentage of the individuals can be correctly classified. We believe that this conceptualization best fits the standard usage of diagnosability and the systematic goals for delimiting subspecies as being more discrete than populations (see Taylor *et al*. (2017*b*) and Martien *et al.* (2017) for further discussion). We thus propose a general definition of diagnosability as "a measure of the ability to correctly determine the taxon of a specimen of unknown origin based on a set of distinguishing characteristics."

We note that there are two related, but different terms—"diagnosability" and "diagnosable." The first, diagnosability, is the measure that results from an analytical method on a particular data set. For example, with a given data set, one algorithm might estimate that 75% of unknown individuals are diagnosable. That is, one is able to correctly classify 75% of the individuals presented to the model to their original stratum. On the other hand, another algorithm applied to the same data might estimate that 98% are diagnosable. The difference between the two diagnosability estimates comes from the power of each analytical method to correctly classify individuals. In contrast, to say that a given subspecies is "diagnosable" is to state that given a particular algorithm and data set, the estimated degree of diagnosability exceeds a threshold value above which one is comfortable assigning subspecies status. So, if a threshold of diagnosability were set at 97%, then in the previous example, the first analysis, where diagnosability was 75%, would not make for a sufficiently diagnosable subspecies, while the latter analysis on the same data (diagnosability = 98%) would meet the threshold.

As in any study, sample size is also a critical factor that should be carefully considered when evaluating diagnosability (Taylor *et al.* 2017*a*). For the purposes of creating a classification algorithm, small sample sizes are often not able to adequately characterize the diversity in the strata under comparison. More importantly, small sample sizes also directly affect the uncertainty of diagnosability estimates, a topic that is rarely, if ever, discussed in the literature. It should be kept in mind that classification probabilities are only estimates and as such have some amount of variability. As sample size increases, these estimates are expected to become more accurate and precise. In light of the fact that thresholds for delimiting subspecies based on diagnosability are arbitrary by nature (Amadon 1949, Wilson and Brown 1953, Johnson *et al.* 1999, Patten and Unitt 2002, Remsen 2010), the effect of sample size on the uncertainty around these diagnosability estimates must be taken into account as well.

Traditionally, morphometric characters formed the basis for most taxonomic descriptions. However, in many fields, genetic markers are increasingly being used to augment morphological data, or even as the sole source of data for taxonomic studies (Cronin 1993, Bradley and Baker 2001, Rosel *et al.* 2017*b*). Due to its nonrecombining nature, combination of hypervariable and conserved sites, and high copy number, mitochondrial DNA (mtDNA) has become the marker of choice for many studies (Zink and Barrowclough 2008, Rosel *et al.* 2017*b*, Martien *et al.* 2017). There are a host of assignment, clustering, and tree-building methods for genetic data. However, few are explicitly designed to assess diagnosability (see Patten 2010 and Martien *et al.* 2017 for reviews). As one example, phylogenetic trees are frequently used to assess diagnosability of species (Brambilla *et al.* 2009). However, using phylogenetic trees for diagnosing subspecies is fundamentally flawed, as trees can only assess monophyly, or the absence thereof, of the haplotypes under examination, and subspecies are not expected to be monophyletic (Patten 2010). It is a well-known limitation that for recently diverged taxa, lineage sorting will be delayed in neutral sites like mtDNA, such that the gene tree will lag behind the species tree (Eckert and Carstens

2008, Remsen 2010). In addition, because low levels of gene flow are expected between subspecies, migration between areas with fairly diverged haplotypes can result in polyphyletic patterns with no shared haplotypes (Archer *et al.* 2013). More importantly, few of these methods allow for the diagnosis of individuals possessing novel sequences. Although mtDNA haplotypes can be assigned to strata based on their frequencies, one cannot assign new haplotypes that have not been seen before.

To bridge this gap, a method is needed that uses genetic sequence data to produce a classification algorithm that can predict the subspecies or species to which unknown specimens belong. One such method, Random Forests (Breiman 2001), is specifically designed to build unbiased classification models and overcomes many of the shortcomings described above. Random Forests has been increasing in popularity as a technique to uncover patterns in large, complex data sets (Cutler *et al.* 2007, Winham *et al.* 2012, Touw *et al.* 2012). In a review of ensemble methods, Berk (2006) demonstrated that Random Forests outperformed and was more robust than methods such as Classification and Regression Trees (CART), bagging, and boosting (all of which can be considered to be special cases of Random Forests). Random Forests has several characteristics that make it an ideal tool for quantifying diagnosability:

- Creates an algorithm to classify individuals of unknown origin.
- Is internally validated and produces strata-specific estimates of classification error.
- Produces individual-specific estimates of classification uncertainty.
- Identifies diagnostic characters.
- Is nonparametric.
- Uses all types of data (continuous/discrete, ordered/unordered).
- Allows weighting of classification probabilities based on prior knowledge.
- Permits tuning to balance classification errors.

In our proposed implementation of Random Forests, the raw data comprise a set of individuals represented by aligned mtDNA sequences, each grouped into their *a priori* defined taxa (*e.g.*, putative subspecies). All individuals in the data set are used, rather than reducing the sequences to unique haplotypes, so that the frequencies of haplotypes, and hence, that of their constituent nucleotide substitutions are properly represented. Additionally, sequences can be reduced to just variable sites so the analysis does not waste time evaluating conserved sites that have no classification information. Because Random Forests operates on the level of the individual nucleotides rather than haplotypes, substitutions that have arisen within strata (synapomorphies) provide the strongest signal for the algorithm. This difference in the way Random Forests treats sequences means that unique and rare haplotypes can contribute useful information to the classification algorithm in a way in which they cannot in standard frequency-based population genetics methods (Martien *et al.* 2017).

In this study, we examine the use of Random Forests for the taxonomic diagnosis of mtDNA sequences. Given that genetic differentiation among diverging strata is influenced by multiple factors such as population size, divergence time, and migration and mutation rates, we generated a series of simulated data sets to model how these factors affect the development of diagnostic sites. These models provide a context and insight towards evaluating the performance of Random Forests on real-world data sets. We then examine the diagnosability of just such an empirical set of mtDNA sequences from recognized populations, subspecies, and species of cetaceans culled from an extensive survey of literature and unpublished studies (Rosel *et al.*

2017*a*). This data set was collected as part of a larger project aimed at bringing more rigor and guidance to the field of cetacean taxonomy, especially in the delimitation of subspecies, which are likely sorely under described (Taylor *et al.* 2017*a*). Finally, we discuss the benefits and limitations of using Random Forests on sequence data to quantify diagnosability for taxonomy, and propose some guidelines for its use.

Materials and Methods

Simulated Data

We simulated mtDNA D-loop data sets generated to represent a range of potential cetacean populations, subspecies, and species using the coalescent-based program *fastsimcoal* (Excoffier and Foll 2011). Each coalescent model was based on a single population that split into two (both with effective population size *Ne*), T generations in the past. After divergence, individuals in each population had a probability m of migrating to the other population each generation. Each simulated individual possessed a 450 base pair (bp) sequence with a mutation rate of μ substitutions/bp/generation. For each simulated data set, we sampled n individuals per stratum.

Preliminary analyses indicated that diagnosability was strongly influenced by the absolute number of migrants (N_em) and the population mutation rate parameter, $\theta = 4N_e\mu$ (Watterson 1975). Therefore, we stratified parameter sampling and modeling in two ways to ensure a uniform distribution of random draws across the parameter space. Additionally, these preliminary analyses indicated that the effect of all parameters except sample size was multiplicative, thus most parameters were sampled from a Uniform distribution in $\log_{10}$ space. For the first model (M1), we first drew 300,000 parameter samples from the following distributions:

Migration rate (m) ~ $10^{\text{Uniform}(-10,\ 0)}$
Mutation rate (μ) ~ $10^{\text{Uniform}(-9,\ -5)}$
Effective population size (N_e) ~ $10^{\text{Uniform}(1.699,\ 5.699)}$
Sample size (n) ~ Gamma (shape = 0.84, rate = 0.0045)
Divergence time (T) ~ $10^{\text{Uniform}(1,\ 6)}$

For the second model (M2), we drew another 300,000 parameter samples from the following distributions, and calculated m and μ:

Number of migrants (N_em) ~ $10^{\text{Uniform}(-5,\ 5)}$
Theta (θ) ~ $10^{\text{Uniform}(-7,\ 1)}$
Effective population size (N_e) ~ $10^{\text{Uniform}(1.699,\ 5.699)}$
Migration rate (m) = N_em/N_e
Mutation rate (μ) = $\theta/4N_e$
Sample size (n) ~ Gamma (shape = 0.84, rate = 0.0045)
Divergence time (T) ~ $10^{\text{Uniform}(1,\ 6)}$

Effective population size (N_e) was sampled to range between 50 and 500,000, covering potential values of cetacean species. The shape and rate parameters of the Gamma distribution for n are derived from a fit to the distribution of individual sizes from the empirical cetacean data described below (mean = 186, median = 80, minimum = 8, maximum = 1,424; Rosel *et al.* 2017*a*). The range of mutation rates for M1 was selected to represent those found in the mtDNA control region (Hoelzel *et al.* 1991, Hayano *et al.* 2004, Jackson *et al.* 2009). In M2, because migration rate

was a calculated value, we censored parameter draws to ensure that $m < 1$. In both models, because N_e and n were sampled independently, we also censored parameter draws to only those where $n < N_e$, and $30 \leq n \leq 300$, making the range of sample sizes comparable to those in the empirical data and similar to what would be encountered in a real study.

For the final simulation data sets, we subdivided these censored parameters to create sets of biologically meaningful categories and ensure an equal number of draws in each. To represent populations in the process of becoming species, we selected 3,000 random parameter draws each from M1 and M2 for which we set m equal to zero. From the remaining M1 parameters, we chose another 3,000 random draws to capture the effect of the range of migration rates sampled. Given that in ideal populations, approximately one migrant per generation is sufficient to prevent differentiation due to genetic drift (Mills and Allendorf 1996, Wang 2004), we drew 3,000 random parameters from M2 (not previously selected for $M2_{m=0}$) where $N_em < 1$ to represent diverging populations, and another 3,000 random parameters where $N_em \geq 1$, representing populations with homogenizing levels of dispersal.

For the purposes of this study, between-individual variability was more important than within-individual variability due to stochasticity in the coalescent, so only one *fastsimcoal* simulation replicate was run for each of the final 3,000 parameter draws from M1 and 9,000 parameter draws from M2.

Modeling Effect of Simulation Parameters on Classification Accuracy

We modeled the relationship between overall classification accuracy and the simulation parameters with a series of logistic Generalized Additive Models (GAM; Wood 2006). The models were constructed to predict the percent of all individuals correctly classified. Overall accuracy was chosen as the response measure because both strata had equal individual sizes, which make it a stable measure of classification ability across all simulations. Because the parameters were measured on different units, all models used a tensor-plate spline function (*te*) with a maximum basis of four. Preliminary analyses indicated that the effects were multiplicative, thus $\log_{10}$-transforms of all parameters were used as predictors in the models. For each subset of parameter draws from the M1 and M2 parameter sets, we fit the following models:

$$M1_{m=0} : p(k|n) \sim \mathrm{te}[\log_{10}(N_e)] + \mathrm{te}[\log_{10}(\mu)] + \mathrm{te}[\log_{10}(T)]$$

$$M1_{m>0} : p(k|n) \sim \mathrm{te}[\log_{10}(N_e)] + \mathrm{te}[\log_{10}(\mu)] + \mathrm{te}[\log_{10}(m)] + \mathrm{te}[\log_{10}(T)]$$

$$M2_{m=0} : p(k|n) \sim \mathrm{te}[\log_{10}(\theta)] + \mathrm{te}[\log_{10}(T)]$$

$$M2_{m>0} : p(k|n) \sim \mathrm{te}[\log_{10}(\theta)] + \mathrm{te}[\log_{10}(N_em)] + \mathrm{te}[\log_{10}(T)],$$

where N_e is effective population size, μ is the mutation rate, T is the divergence time in generations, m is the migration rate, and $\theta = 4N_e\mu$. For $M2_{m>0}$, three models were fit: one for all data where $m > 0$, a second for the $N_em < 1$ subset, and the third for the $N_em \geq 1$ subset.

Stratification Errors

Misstratification is incorrectly placing an individual within a stratum, which could happen when animals stray into atypical geographic regions or mix with other

groupings during certain seasons. This error could be particularly problematic for genetic data primarily obtained from biopsy samples where other independent lines of evidence, like morphology, are not possible to obtain. In order to examine the effect of errors in individual stratification, we conducted a sensitivity test with the simulated data. In the test, for each of the 9,000 M2 data sets in which more than one haplotype was present, individuals were assigned to the incorrect stratum based on misstratification probabilities of 0.001, 0.005, 0.01, 0.05, and 0.1. We then ran the same Random Forests analysis on these misstratified data sets as described below and recorded the decrease in diagnosability resulting from a given level of misstratification error.

Empirical Data

We also analyzed a set of mitochondrial control region sequences (280–961 bp) compiled from literature and unpublished data as described in Rosel *et al*. (2017*a*) (Table 1). These empirical comparisons were selected based on the following factors: (1) a consensus as to their taxonomic level (population, subspecies, or species); (2) ensuring that as many subspecies were represented as possible; and (3) ensuring the presence of cases that would represent difficult to assess comparisons due to issues like large variability due to large population sizes, or low variability due to highly structured social systems. The *a priori* assignment of individuals to strata was taken from the original authors' designations for published data sets or based on geography or morphological features for unpublished data sets. Further rationale and details of comparisons chosen are given in Rosel *et al.* (2017*a*) and Supporting Information therein.

Random Forests

Random Forests is an ensemble-based classification algorithm that extends the more familiar method of CART by adding several layers of stochasticity to the tree growing process. This permits the algorithm to fully explore the predictive capability of all variables, as well as producing an internally validated classifier. The process of building a Random Forests model is illustrated in Figure 2. For each tree in the forest, the first step is to select a random set of sequences that are used as the training set for the tree. Those sequences not selected (the out-of-bag or OOB sequences) are set aside for cross-validation of the tree's prediction accuracy. The tree is then grown in the following iterative manner:

(1) Choose a random subset of nucleotide sites from all available sites.
(2) For each site chosen, create a splitting rule that divides the sequences into two groups with the greatest purity of the *a priori* designated groups (lowest Gini index).
(3) Choose the site that produces the best split and create two daughter nodes of sequences based on that split.
(4) For each of these daughter nodes, return to step 1 and repeat until all nodes contain a single sequence.

The OOB sequences are then sent through the decision tree based on its splitting rules and classified to the stratum of the individual in the final node they end up in. In this manner, a tree produces a single "vote" to a given stratum for each OOB sequence. Steps 1 through 4 are repeated multiple times to produce many trees (the

Table 1. List of pairwise comparisons from empirical mtDNA D-loop data sets, categorized by type (population, subspecies, or species). Further details of comparison selection provided in Rosel *et al.* (2017*a*).

Comparison type	Comparison label	Species	Strata (sample size)	Fixed differences	Sequence length
Population	Bmys.1	*Balaena mysticetus*	Atlantic (279) *vs.* BCB (343)	0	349
	Bmys.2	*Balaena mysticetus*	BCB (343) *vs.* Okhotsk (20)	0	397
	Bphy.1	*Balaenoptera physalus*	Gulf of California (33) *vs.* northeastern Pacific (313)	0	411
	Mno*vs*.1	*Megaptera novaengliae*	HI (59) *vs.* MX-ML (60)	0	454
	Mno*vs*.2	*Megaptera novaengliae*	NGOA (59) *vs.* SEA (60)	0	455
	Oorc.1	*Orcinus orca*	EAL.TRI (65) *vs.* WAL.RUS (95)	1	919
	Oorc.2	*Orcinus orca*	NR (13) *vs.* SR (25)	1	919
	Pcra	*Pseudorca crassidens*	HI Insular (96) *vs.* Pacific pelagic (69)	0	945
	Pdal	*Phocoenoides dalli*	America (33) *vs.* Gyre (15)	0	379
	Pmac.1	*Physeter macrocephalus*	Ca.Current (52) *vs.* ETP (114)	0	399
	Pmac.2	*Physeter macrocephalus*	NAtl (293) *vs.* NPac (194)	0	399
	Ppho.1	*Phocoena phocoena*	Gulf.of.Maine (80) *vs.* Newfoundland (42)	0	342
	Ppho.2	*Phocoena phocoena*	Monterey (92) *vs.* San Juan Islands (20)	0	393
	Satt.1	*Stenella attenuata*	Central.America (36) *vs.* Costa.Rica (32)	0	421
	Satt.2	*Stenella attenuata*	Ecuador (33) *vs.* Northern.Mexico (34)	0	421
	Satt.3	*Stenella attenuata*	four.islands (27) *vs.* Oahu (27)	0	535
	Slon.1	*Stenella longirostris*	Maui (59) *vs.* Oahu (39)	0	415
	Ttru.1	*Tursiops truncatus*	ATL.coastal (100) *vs.* GOM.coastal (72)	0	354
	Ttru.2	*Tursiops truncatus*	four.islands (26) *vs.* Oahu (30)	0	401
	Ttru.3	*Tursiops truncatus*	Kauai (41) *vs.* Offshore (69)	0	402

(Continued)

Table 1. (Continued)

Comparison type	Comparison label	Species	Strata (sample size)	Fixed differences	Sequence length
Subspecies	Bphy.2	*Balaenoptera physalus*	North Atlantic (33) *vs*. Southern Hemisphere (48)	1	410
	Ccom	*Cephalorhynchus commersonii*	commersonii (196) *vs*. kerguelensis (11)	0	423
	Chec	*Cephalorhynchus hectori*	hectori (318) *vs*. maui (70)	1	340
	Gmel	*Globicephala melas*	G..melas.edwardii (573) *vs*. G..melas.melas (70)	0	345
	Lobs.1	*Lagenorhynchus obscurus*	L.o.fitzroyi (14) *vs*. L.o.obscurus (21)	0	591
	Lobs.2	*Lagenorhynchus obscurus*	L.o.fitzroyi (14) *vs*. L.o.posidonia (118)	0	591
	Lobs.3	*Lagenorhynchus obscurus*	L.o.obscurus (21) *vs*. L.o.posidonia (118)	2	591
	Ppho.3	*Phocoena phocoena*	Black Sea (113) *vs*. North Atlantic (560)	0	338
	Satt.4	*Stenella attenuata*	Coastal (135) *vs*. Offshore (90)	0	421
	Slon.2	*Stenella longirostris*	Slonlon (116) *vs*. Slonori (87)	0	400
	Ttru.4	*Tursiops truncatus*	T.t.ponticus (43) *vs*. T.t.truncatus (E.Atl and E.Med) (88)	0	415
Species	Bmys *vs*. Egla	*Balaena mysticetus* *Eubalaena glacialis*	Bmys (642) *vs*. Egla (430)	8	388
	Bphy *vs*. Bmus	*Balaenoptera physalus* *Balaenoptera musculus*	Bmus (295) *vs*. Bphy (427)	23	347
	Bphy *vs*. Mnov	*Balaenoptera physalus* *Megaptera novaengliae*	Bphy (427) *vs*. Mnov (1424)	14	388
	Ccom *vs*. Chec	*Cephalorhynchus commersonii* *Cephalorhynchus hectori*	Ccom (207) *vs*. Chec (388)	10	340
	Dleu *vs*. Mmon	*Delphinapterus leucas Monodon monoceros*	Dleu (122) *vs*. Mmon (421)	25	291
	Eaus *vs*. Ejap	*Eubalaena australis Eubalaena japonica*	Eaus (637) *vs*. Ejap (23)	7	380

(Continued)

Table 1. (Continued)

Comparison type	Comparison label	Species	Strata (sample size)	Fixed differences	Sequence length
	Egla *vs*. Eaus	*Eubalaena glacialis* *Eubalaena australis*	Eaus (637) *vs*. Egla (430)	8	426
	Egla *vs*. Ejap	*Eubalaena glacialis* *Eubalaena japonica*	Egla (430) *vs*. Ejap (23)	8	399
	Erob *vs*. Bphy	*Eschrichtius robustus* *Balaenoptera physalus*	Bphy (427) *vs*. Erob (262)	31	417
	Erob *vs*. Mnov	*Eschrichtius robustus* *Megaptera novaengliae*	Erob (262) *vs*. Mnov (1424)	21	543
	Fatt *vs*. Pcra	*Feresa attenuata* *Pseudorca crassidens*	Fatt (54) *vs*. Pcra (202)	38	955
	Gmac *vs*. Gmel	*Globicephala macrorhynchus* *Globicephala melas*	Gmac (891) *vs*. Gmel (232)	4	304
	Gmac *vs*. Pele	*Globicephala macrorhynchus* *Peponocephala electra*	Gmac (891) *vs*. Pele (177)	12	963
	Kbre *vs*. Ksim	*Kogia breviceps* *Kogia sima*	Kbre (258) *vs*. Ksim (90)	4	402
	Lobs *vs*. Lobl	*Lagenorhynchus obscurus* *Lagenorhynchus obliquidens*	Lobl (59) *vs*. Lobs (153)	6	520
	Npho *vs*. Nasi	*Neophocoena phocoenoides* *Neophocoena asiaorientalis*	Nasi (10) *vs*. Npho (27)	0	294
	Obre *vs*. Ohei	*Orcaella breviceps* *Orcaella heinsonii*	Obre (46) *vs*. Ohei (14)	17	403
	Pele *vs*. Fatt	*Peponocephala electra* *Feresa attenuata*	Fatt (54) *vs*. Pele (177)	25	961
	Ppho *vs*. Pdal	*Phocoena phocoena* *Phocoenoides dalli*	Pdal (239) *vs*. Ppho (431)	13	391
	Psin *vs*. Pspi	*Phocoena sinus* *Phocoena spinnipinis*	Psin (43) *vs*. Pspi (29)	27	398
	Satt *vs*. Slon	*Stenella attenuata* *Stenella longirostris*	Satt (325) *vs*. Slon (735)	4	390
	Sflu *vs*. Sgui	*Sotalia fluviatalis* *Sotalia guianensis*	Sflu (21) *vs*. Sgui (55)	6	280

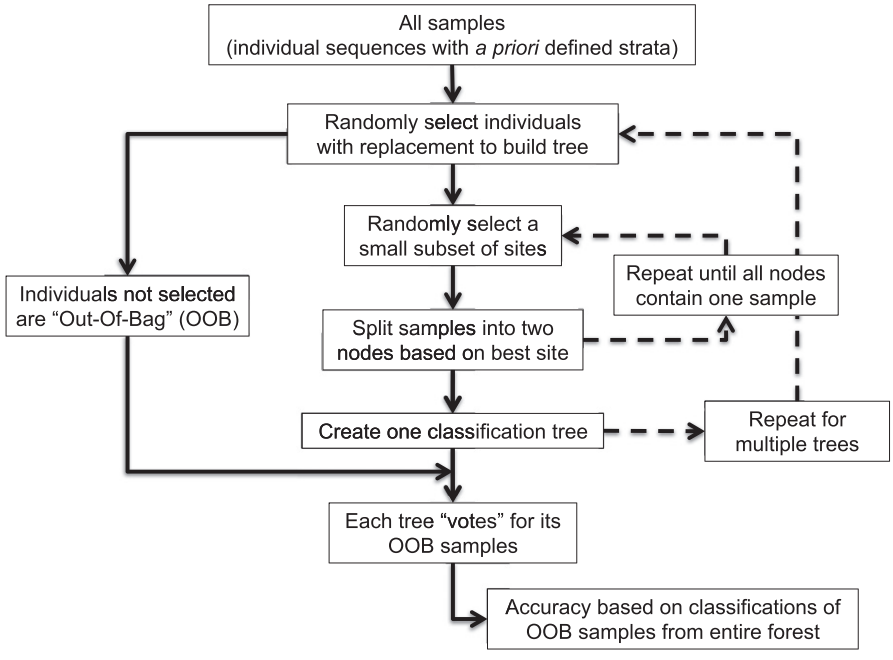

Figure 2. Illustration of steps in constructing a Random Forests ensemble of classification trees.

"forest"), each of which votes for the strata of their own respective OOB sequences. The probability (P) that an individual is classified to a given stratum is the fraction of trees voting for that stratum in the subset of trees in the forest where the individual was OOB. Thus, a sequence is predicted to belong to the stratum with the largest P. In the simple case of two strata, this would be the stratum for which $P > 0.5$.

We ran Random Forests on all data using the *randomForest* package (Liaw and Wiener 2002) in R v3.0.2 (R Core Team 2015). For each comparison, only variable sites were used as predictors in the Random Forests analysis. Sites that were variable as a result of a substitution in a single individual were also excluded because they do not add useful information for validation of the classification model with the OOB individuals. Insertion/deletions (indels) were treated as unique substitutions equivalent to other nucleotides. For all simulated and empirical comparisons 10,000 trees were built for each forest, which were found to produce stable classification models for all empirical comparisons. The number of individuals chosen to build each tree was set to that of half of the smallest stratum, and sampling was done without replacement, ensuring that OOB individuals would be available from each strata for cross-validation. All other *randomForest* parameters were left to their defaults, and individual models were not optimized in order to produce comparable, unbiased results.

Because all Random Forests analyses were made between two strata, individuals were assigned to the stratum for which more than 50% of the trees voted for them when they were OOB. In this paper, we refer to the total fraction of individuals correctly classified in this manner as the percent diagnosable-50 (PD_{50}), which is equal

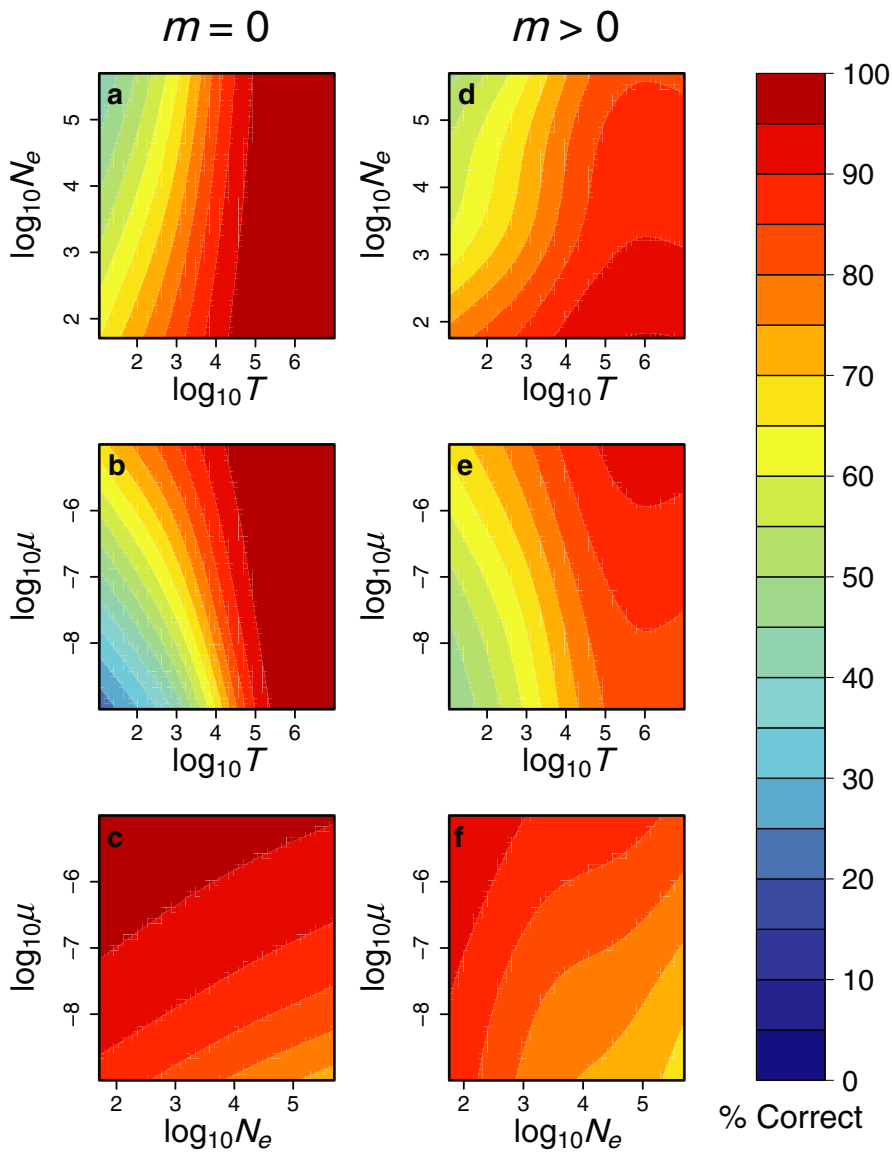

Figure 3. Two-dimensional GAM fits of effective population size (N_e), divergence time in generations (T), and mutation rate (μ) from Model 1 simulated data. Results from models without migration to the left and those with migration to the right. Colors indicate model prediction of percent correctly classified.

to 1 minus the strata-specific OOB error rate output by *randomForest*. In order to characterize the distribution of individual classification probabilities (fraction of trees voting for each individual), we also report PD_{95}, which is the fraction of individuals in a stratum with classification probabilities >95% to that stratum. As described above,

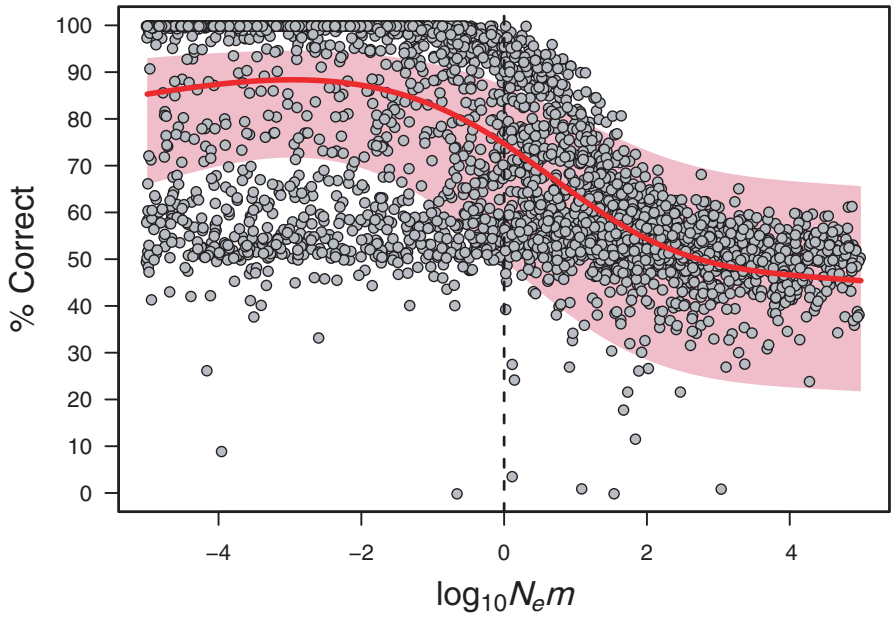

Figure 4. GAM fit of number of migrants (N_em) from Model 2 parameters. Solid line shows median value of predicted percent correctly classified, and shaded area shows 95% CI. The switch from bimodal distribution to a normal distribution occurs at $N_em = 1$ ($\log_{10}N_em = 0$).

because PD_{50} and PD_{95} are stratum-specific, we use the smallest PD_{50} and its associated PD_{95} as the overall measures of diagnosability and individual classification certainty for a comparison. Thus the estimate of diagnosability for a comparison is a conservative one. Uncertainty of the PD_{50} diagnosability estimate was determined using a standard binomial distribution, from which we present the central 95% credibility interval (CI).

We also present two other measures to help place the performance of the Random Forests models into a comparable context with one another. The first is the degree of diagnosability one would expect by random assignment of individuals, or the "prior" model accuracy for the least diagnosable stratum (PD_{prior}). This is simply the proportion of all individuals in a comparison represented by that stratum. For example, if a stratum contains 25% of the samples in a comparison, one would expect that by random chance alone, one would be able to classify 25% of the samples from that stratum correctly. For a model with two strata and equal sample sizes, PD_{prior} would simply be 50%.

The second measure is the maximum diagnosability the model could achieve (PD_{max}), based on classifying individuals using only the frequencies of their haplotypes in each stratum. Individuals with unique haplotypes are not considered. In cases where a haplotype is shared among strata, Random Forests will classify all individuals with that haplotype to the stratum in which it is at the highest frequency, putting an upper limit on the maximum diagnosability achievable. Thus, while it is possible for diagnosability as measured by PD_{50} to be less than PD_{prior} (a sign of very low classification ability and poor model performance), it cannot be greater than PD_{max}.

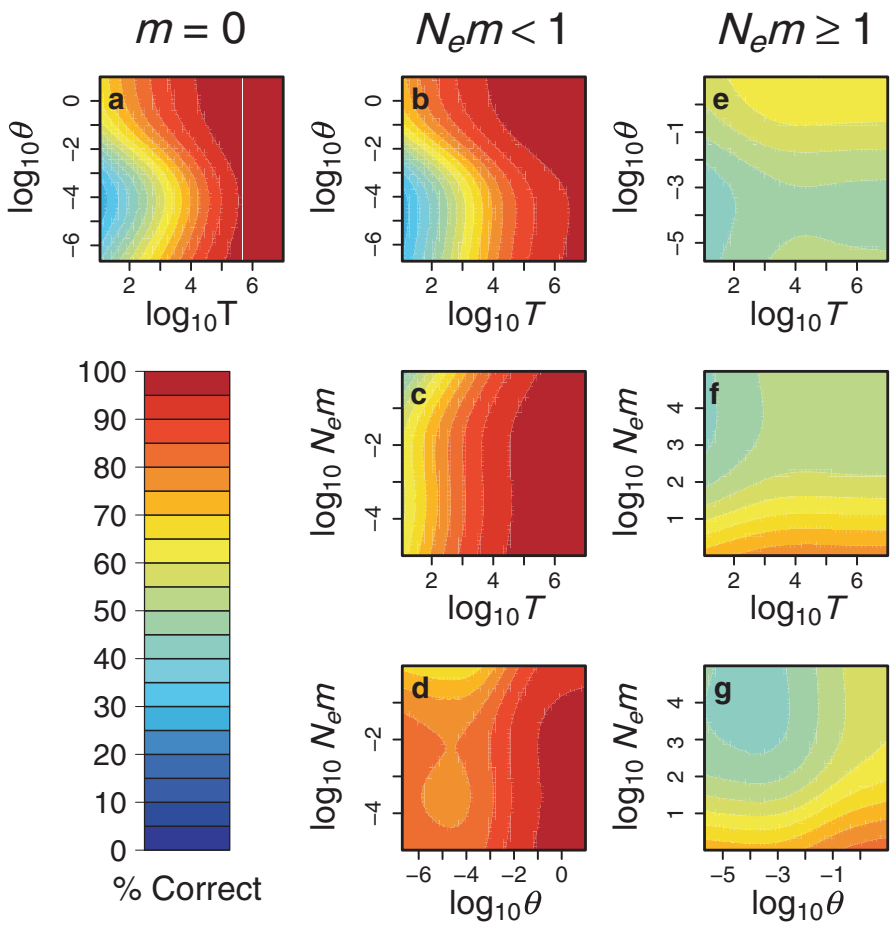

Figure 5. Two-dimensional GAM fits of theta (θ), number of migrants (N_em), and divergence time in generations (T) from Model 2 simulated data. From left to right, columns show results from models without migration ($m = 0$), with migration and $N_em < 1$, and $N_em \geq 1$. Colors indicate model prediction of percent correctly classified.

Results

Simulated Data

The effects of the simulation parameters on predictions of overall classification accuracy from Random Forests models are well illustrated by the GAM model fits (Fig. 3–5). Given that the simulations were coalescent-based, the general relationships are easily interpreted from basic population genetics theory. Below, we highlight those parts of the parameter space leading to the development of diagnostic sites as evinced by particularly large estimated classification accuracy from the models.

As expected, classification accuracy was generally lower in the M1 data set with migration than in the one without. All individuals were correctly classified in 1,293 of the 3,000 simulations (43%) where $m = 0$ as compared with 621 of the 3,000

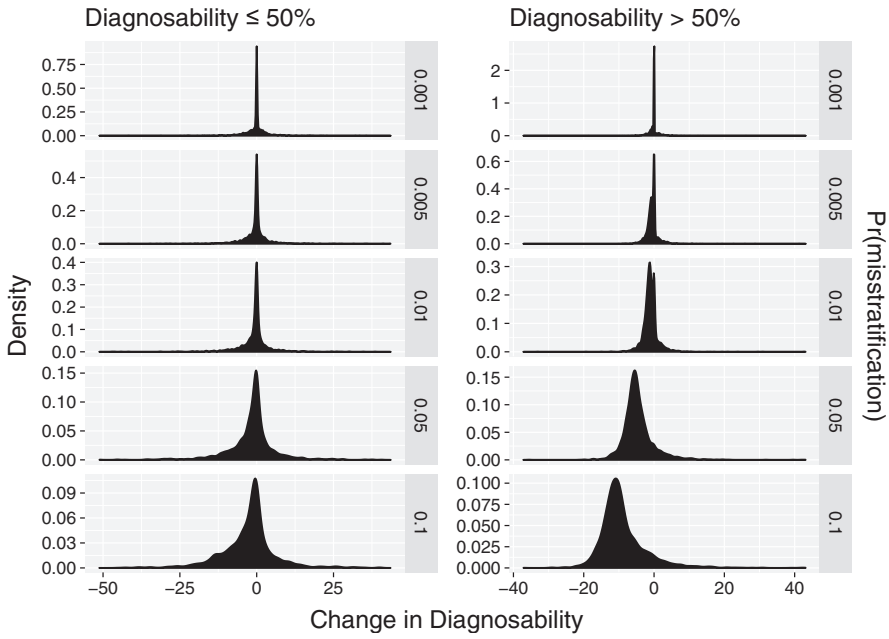

Figure 6. Frequency distributions of the change in observed diagnosability (*x*-axis) in the simulated data for increasing levels of the probability of misstratification (vertical panels). Figures on the left and right columns are censored by data sets for original diagnosability ≤50% and >50%, respectively.

simulations (21%) where $m > 0$. In the GAM fits to the $M1_{m=0}$ data, divergence time (T) was the strongest predictor of overall accuracy (Fig. 3a, b), reaching perfect classification at approximately 10^5–10^6 generations across all values of effective population size (N_e) and most values of mutation rate (μ). Predicted classification rates decreased with decreasing mutation rate, but were more strongly affected by divergence time (Fig. 3b). For example, classification rates greater than approximately 70% are not predicted for mutation rates $<10^{-7}$ unless divergence time is greater than approximately 1,000 generations. Finally, although classification rates tended to be higher with smaller effective population size, this effect was primarily evident for short divergence times (Fig. 3a), indicating that effective population size was not a strong predictor across much of the parameter space. With migration present in the $M1_{m>0}$ data (Fig. 3d–f), the patterns were similar to those seen in the $M1_{m=0}$ fits; large correct classification rates (>90%) were only predicted with long divergence times ($>10^5$ generations), small effective population sizes (<1,000 individuals), and high mutation rates ($>10^{-6}$/bp/generation). Divergence time was still a very strong predictor, but overall accuracy was affected more prominently by effective population size (Fig. 3d, f).

The M2 parameter sets were selected to explore the effect of the effective population size-normalized parameters N_em and θ. Similar to the case with the M1 parameters, 1,037 of the 3,000 $M2_{m=0}$ simulations (35%) had perfect classification. There is a stark difference between the $M2_{m>0}$ set of parameters where $N_em < 1$ and $N_em \geq 1$ (Fig. 4). In the 3,000 simulations where $N_em < 1$, 803 had perfect classification of all individuals (27%) while only 2/3,000 (0.06%) where $N_em \geq 1$ showed perfect classification. Simulations in the $N_em < 1$ data set formed two modes, one with large

classification accuracy (>80%), and another centered between 45% and 55%. The latter mode was primarily composed of simulations in which only one haplotype was present among all individuals in both strata. Simulations from the $N_em \geq 1$ data set converged on a classification accuracy of 50%.

In the $M2_{m=0}$ model, where the only parameters were θ and divergence time (T), overall accuracy increases with increasing T, predicted to exceed 0.9 when divergence time is greater than approximately 10^5 generations (Fig. 5a). However, this level of accuracy can also be reached at divergence times closer to 10^4 generations when θ is less than approximately 10^{-3}. When migration is present, but $N_em < 1$, there is a stronger effect of θ in relation to generation time when $\theta > 10^{-3}$ implying that when mutation is reduced past this level, migration will be a stronger homogenizing force, even at long divergence times (Fig. 5b). For values of N_em less than approximately 0.1, there is little difference in estimated classification accuracy across values of divergence time and mutation rate (Fig. 5c, d). In contrast, when $N_em \geq 1$, there is little effect of divergence time on classification accuracy (Fig. 5e, f). In this model, the highest estimated accuracies were between 70% and 85%, and were only predicted for θ greater than approximately 0.1 and N_em less than approximately 10 individuals per generation (Fig. 5f, g).

Stratification Errors

The misstratification sensitivity test revealed that increasing the probability that individuals were misstratified tended to cause a decrease in diagnosability, as expected. This tendency was more strongly seen when the original diagnosability was >50%, (Fig. 6). However, even at the largest stratification error rate examined (P = 0.1), the central 95th percentile of the distribution of the change in diagnosability still spanned zero (Table 2). If this distribution is further censored to data sets with diagnosability >60%, the median for this error rate becomes −10.9, and the central 95th percentile increases to −0.19–18.51. With a misstratification rate of P = 0.01, the median change in diagnosability is only −0.61 and −1.12 for the diagnosability ≤50% and >50% data sets, respectively.

Empirical Data

A summary of the Random Forests analyses for the empirical comparisons is given in Figure 7, which shows the range of diagnosability estimates (smallest PD_{50}), their associated 95% confidence intervals, the PD_{95} for the same strata as a measure of individual classification certainty, and the range of PD_{prior} to PD_{max} as a measure of the model performance. For example, in one of the two *Physeter macrocephalus* population

Table 2. Summary of stratification error sensitivity test on M2 simulated data where diagnosability >50%.

		Change in diagnosability		
Pr(misstratification)	*n*	Median	Central 95%	Percent > 0
0.001	3,537	0	−3.04–3.33	0.27
0.005	3,520	−0.46	−3.95–3.62	0.57
0.01	3,527	−1.12	−5.41–4.11	0.72
0.05	3,485	−5.2	−11.66–6.65	0.89
0.1	3,388	−10.24	−18.32–8.12	0.91

comparisons (Pmac.2: North Atlantic *vs*. North Pacific), the estimated diagnosability is 72% (95% CI = 66%–77%). This is better than the prior of 60% for this particular stratum, but based on the distribution of haplotypes shared between the two strata in this comparison, 72% is the best diagnosability obtainable with these data. Thus, for this comparison, there is evidence of some diagnostic signal in the mtDNA sequences, but diagnosability is limited.

The other *Physeter macrocephalus* population comparison (Pmac.1: California Current *vs*. eastern tropical Pacific) provides a contrasting example. In this comparison, if there were diagnostic sites, the best possible diagnosability would be 33%. However, the actual diagnosability is 27% (95% CI = 15%–41%), which is even less than the 31% that would be expected by random chance. This poor performance is likely related to an imbalance in the sample sizes, in which the smaller population (California Current, $n = 52$) performed much worse than the larger (eastern tropical Pacific, $n = 114$). This poor performance is also reflected in the individual classification uncertainty for the California Current strata, in which none of the individuals were classified with greater than 95% probability ($PD_{95} = 0$).

As expected, all comparisons with fixed differences among strata (21/22 species, 3/11 subspecies, and 2/20 populations) had all individuals correctly classified. The only species comparison without fixed differences (*Neophocaena phocoenoides vs. N. asiaorientalis*) had PD_{50} values of 100% (*N. phocoenoides*) and 80% (*N. asiaorientalis*). The 95% CI for the *Neophocoena* diagnosability estimate was also the widest of all comparisons at 44%–97%. For all species comparisons except *Neophocoeana,* PD_{95} values were >90%, indicating large confidence in individual assignments.

Three subspecies comparisons (*Balaenoptera physalus* [Bphy.2], *Commersoni hectori* [Chec], and *Lagenorhynchus obscurus* [Lobs.3]) had fixed differences and diagnosabilities of 100%. The *Phocoena phocoena* subspecies comparison (Ppho.3) also had a diagnosability of 100%, even in the absence of fixed differences. The remaining seven subspecies comparisons had diagnosabilities ranging from 50% to 99%. The two worst performing subspecies comparisons were for *Stenella longirostris* (Slon.2), and *S. attenuata* (Satt.4) with diagnosabilities less than 60%. Eight of the eleven subspecies comparisons had diagnosabilities equal to PD_{max}, demonstrating very good model performance in these cases. Two of the four subspecies comparisons with diagnosability less than PD_{max} (Lobs.1 and Ttru.4) had 95% CIs very close to or exceeding PD_{max}.

The two population comparisons with fixed differences were both killer whale comparisons in which each population is characterized by a single haplotype. The two other population comparisons with diagnosabilities >85% were *Pseudorca crassidens* (Pcra), and one of the *Tursiops truncatus* comparisons (Ttru.3). The remainder of the population comparisons had diagnosabilities <81%. In only six of the 20 population comparisons were diagnosabilities seen that were lower than the PD_{prior} values indicating that there was little diagnostic information in the sequences in these comparisons. Additionally, most individuals in many population comparisons (13/20) were not classified to their population of origin with great certainty as evidenced by the PD_{95} values <0.5.

Discussion

Factors Affecting Diagnosability

The development of diagnostic sites in the continuum from populations to full species is the result of the interplay of multiple factors. The most obvious factor

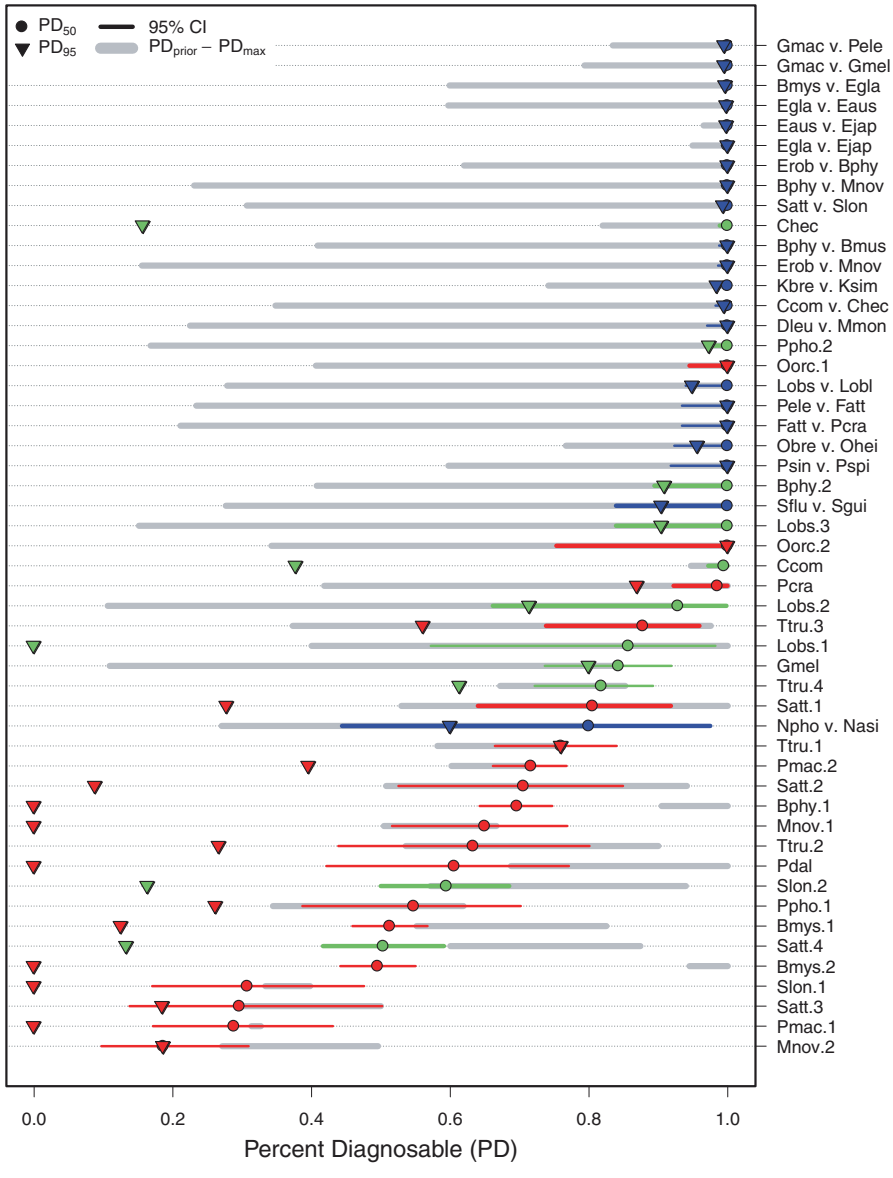

Figure 7. Summary of Random Forests classifications for each empirical comparison. Each row shows results from the stratum with the smallest fraction of individuals correctly classified, with comparisons labeled by their taxonomic codes as listed in Table 2. Colors identify comparison type as species (blue), subspecies (green), and populations (red). Points show the fraction of individuals correctly classified with probabilities > 50% (PD_{50}, circles), and > 95% (PD_{95}, triangles). Thin colored lines show 95% confidence intervals (CI) around PD_{50} estimates. Gray bars show range of *a priori* random classification rates based on individual size (left) to maximum possible classification rates based on shared haplotypes (right).

promoting speciation is a decrease in dispersal among diverging populations, which allows for the independent accrual of mutations and fixation of haplotypes resulting from genetic drift. Our simulation results strongly support previous findings that differentiation occurs when dispersal is less than approximately one migrant per generation (Mills and Allendorf 1996, Wang 2004). Less than this threshold, there is still a small part of the parameter space (low θ and low divergence time) where classification can be low. However, outside of this part of the parameter space, classification accuracy tended to be greater than approximately 80%, indicating that dispersal was low enough that diagnostic sites could develop independently.

The simulations also highlighted how strongly divergence time affects classification accuracy. On average, it took a minimum of approximately 10^4 generations for genetic drift to produce high (>0.9) classification accuracy between two diverging species. Given cetacean generation lengths of approximately 20 yr (Taylor *et al.* 2007), our models suggest that perfect diagnosability between species would take on the order of two million years to develop. This estimate is consistent with published estimates of divergence times for many closely related cetacean species (Caballero *et al.* 2007, McGowen *et al.* 2009, Vilstrup *et al.* 2011).

It is well known that haplotypes drift to fixation more rapidly in populations with relatively small N_e (Frankham 1996, Charlesworth 2009), which should lead to higher classification accuracy. Although this effect was seen in the simulated data, the relationship with N_e is not as strong as that of other factors. It is most evident at short divergence times and low mutation rates, where diversity and differentiation would be expected to be low on average, and therefore amplified by these increased rates of genetic drift. The effective population size of a given stratum will reflect its life history characteristics and depend on the taxonomic level under examination. Coastal species with limited geographic ranges will tend to have smaller N_e than pelagic species with transoceanic distributions. Likewise, N_e of a subspecies will always be larger than that of the populations that comprise it, as will the effective population size of a species relative to its subspecies.

Finally, mutation rate was also seen to be an important predictor of classification accuracy. Across the range of mutation rates we examined (10^{-9}–10^{-5}/bp/generation), the lowest values were insufficient for producing any variability except when divergence times were relatively long. Mutation rate estimates from pedigree studies in humans for the mtDNA control region are on the order of 10^{-5}–10^{-6}/bp/generation (Sigurðardóttir *et al.* 2000, Heyer *et al.* 2001). Rates for cetaceans based on fossil calibrations are on the order of 10^{-7}–10^{-8}, with baleen whales tending to have slower mutation rates than toothed whales (Hoelzel *et al.* 1991, Harlin *et al.* 2003, Hayano *et al.* 2004, Jackson *et al.* 2009). Thus, while the development of diagnostic sites will likely be slower for loci with more conserved regions, it is also likely to occur at a lower rate in some taxa as compared to others.

Empirical Performance of Random Forest

Full diagnosability of species was observed in our empirical data set of cetacean mtDNA control region sequences in all of the species comparisons but one (*Neophocaena phocaenoides vs. N. asiaorientalis*). The description of *N. asiorientalis* as a new species is recent (Wang *et al.* 2010, Jefferson and Wang 2011) and was based primarily on diagnostic morphological characters. The lack of fixed differences in the control region sequence for this species pair is explained by the very recent estimate of divergence time (18,000 yr), leading to incomplete lineage sorting (Jefferson and Wang

2011). As mentioned earlier, the choices of which species pairs to include in Rosel *et al.* (2017*a*) was not random but specifically focused on cases likely to be problematic for developing quantitative standards for genetic taxonomic delimitation. The *Neophocaena* data set was chosen to aide in this process knowing that the shared haplotypes would not allow full diagnosability with this marker.

When used for subspecies delimitation, diagnosability does not mean perfect classification, but rather something close to it (Patten and Unitt 2002, Remsen 2010), making the boundary between subspecies and populations subjective (Wilson and Brown 1953, Martien *et al.* 2017, Taylor *et al.* 2017*b*). In our analysis, although subspecies tended to be more diagnosable than populations, there were no absolute boundaries between the two types of comparisons. Subspecies tended to have a larger proportion of individuals correctly classified with high certainty, while many population comparisons performed worse than would be expected by random chance ($PD_{50} \ll PD_{prior}$). While a special algorithm like Random Forests is not necessary with fixed differences, these patterns in the absence of fixed differences are evidence that Random Forests is making use of diagnostic information from a combination of sites that has arisen during population divergence. Lack of concordance among multiple sites is not a confounding issue for Random Forests as it is with standard morphological methods (Remsen 2010).

In interpreting these results from the empirical data analysis, we re-emphasize that the set of comparisons we used should not be construed as a random selection of all possible comparisons at each taxonomic level. As described in Rosel *et al.* (2017*a*), these comparisons were selected to be (1) as comprehensive as possible of the available data for subspecies level comparisons; (2) representative of "difficult" cases (*i.e.*, comparisons that were expected to be near the population/subspecies or subspecies/species boundaries; and (3) comparisons for which there was relative consensus as to their taxonomic level. The desire for consensus necessarily limited the number of possible comparisons because the objective of the series of papers in this volume is to move towards reducing the number of likely errors in cetacean taxonomy.

For example, multiple ecotypes of killer whales are found in every ocean basin in which they have been studied (Hoelzel *et al.* 2007, Morin *et al.* 2010, Foote *et al.* 2011). Because there is ongoing debate about whether these ecotypes are populations, subspecies, or species, we did not conduct comparisons between ecotypes. Instead, we elected to compare neighboring populations within the ecotype known as "North Pacific resident killer whales." These comparisons were deliberately chosen to examine potential errors using mtDNA for species with exceptionally low effective population size. These populations also have very low diversity, wherein each is fixed for one of three haplotypes, causing each to be fully diagnosable. Similar to killer whales, Hawaii insular and Pacific pelagic populations of false killer whales (*Pseudorca crassidens*) have strong social structures and relatively low effective population sizes (Martien *et al.* 2014). Although there were no fixed differences or shared haplotypes between these two strata, they were also highly diagnosable in the Random Forests analysis ($PD_{50} = 98\%$, $PD_{95} = 88\%$), making them more similar to other subspecies comparisons than to the other population comparisons. However, as of this writing, there has not been a formal examination of the taxonomic status of false killer whales.

In contrast to the *Orcinus* and *Pseudorca* comparisons (which were chosen to reveal potential classification issues with very low N_e), those comparisons with low diagnosability tended to be pelagic delphinid populations and subspecies (which were chosen in part to reveal potential classification issues with high N_e). As seen in the simulation models, the increase in diversity and related decrease in the rate of genetic drift

for populations with large N_e leads to low diagnosability with a marker like the control region. This is especially true in recently diverged taxa and when sampling is low relative to population size, as is likely the case for the *Stenella longirostris* and *S. attenuata* population and subspecies comparisons. For these strata, sample sizes were between 20 and 40 individuals, while population sizes as recently as the 1960s are on the order of several millions (Wade *et al.* 2007).

There is also the potential for misstratification of individuals in these and similar data sets. In all studies, the original assignment of individuals to strata is based on a measure independent of the genetic data, usually morphology in the case of species, or geographic location in the case of populations. For subspecies, there can be overlap of morphological characters, as well as geographic overlap of individuals. If individuals are taken from the region of sympatry, *a priori* assignment of individuals for the purposes of training a classification model can be problematic and can carry with it an air of circularity (Remsen 2010). In their re-analysis of sage sparrow (*Amphispiza belli*) data, Cicero and Johnson (2006) have shown that errors of assignment can strongly affect estimates of diagnosability and hence inferences of subspecies status. In our sensitivity test, we saw relatively small changes in diagnosability with error rates $\leq 1\%$, which should be achievable in most studies if attention is paid to sample provenance and the necessary corroborative data for subspecies assignment can be obtained.

Overclassification

Clearly, more sites provide more information for classification, thus longer sequences would be predicted to improve diagnosability. In this era of high-throughput sequencing, one can expect that data sets composed of the full mitogenome as well as kilobases of nuclear sequences will soon become the norm (Davey *et al.* 2011). Given enough loci, one would expect to be able to distinguish clusters of closely related individuals or family groups as they are more likely to share unique characters than randomly selected individuals within a population. If, either as an artifact of sampling or natural patterns of distribution, individuals have been stratified such that they are largely composed of family groups, there is a potential that increasing the amount of data will increase the chance of incorrectly assigning subspecific status, thus making an overclassification error.

This concern overlooks a key feature of subspecies that distinguishes them from family groups. Most definitions of subspecies emphasize their geographic distinctiveness and diagnosability (Amadon 1949, Wilson and Brown 1953, Patten 2010, Winker 2010). As outlined above, this cannot be considered a sufficient criterion by itself. However, subspecies are also recognized as entities along an evolutionary continuum (Haig *et al.* 2006). As well articulated by Patten (2010), although not all subspecies become species, all species theoretically had to go through a subspecies stage. In light of this, the definition of subspecies set forth by Taylor *et al.* (2017*b*) requires that subspecies are both diagnosable and appear to be on separate evolutionarily lineages. This reflects the expectation that subspecies should demonstrate a degree of divergence greater than expected between populations or related individuals within a population.

This is illustrated in the relationship between diagnosability and divergence seen in the empirical cetacean data (Fig. 8). While the two metrics are correlated, there is considerable variability in both, suggesting that each is measuring a different aspect of the divergence process. In the region of overlap for population and subspecies

divergence estimates (1.51×10^{-4} – 1.08×10^{-2}), populations tend to be less diagnosable. Populations plateau at diagnosability less than one, unlike subspecies. This pattern is consistent with the results from the simulations showing that even after long divergence times, diagnosability was capped in the presence of dispersal.

In response to concerns of making overclassification errors because of too many characters, Helbig (2002) advised limiting the number of characters used in diagnosing species to "two or three." On the contrary, rather than applying an arbitrary limit to the number of characters used in a diagnosis, we suggest that the goal of a study should be to use all available characters to produce the best validated model without overfitting the data. For some taxa, the mitochondrial control region will be a sufficient marker for demonstrating diagnosability. For others, especially those with extremely large effective population sizes, at least the entire mitogenome will be required if not other nuclear markers, especially if individual size is relatively small. Diagnosability should then be assessed in conjunction with evidence of evolutionary

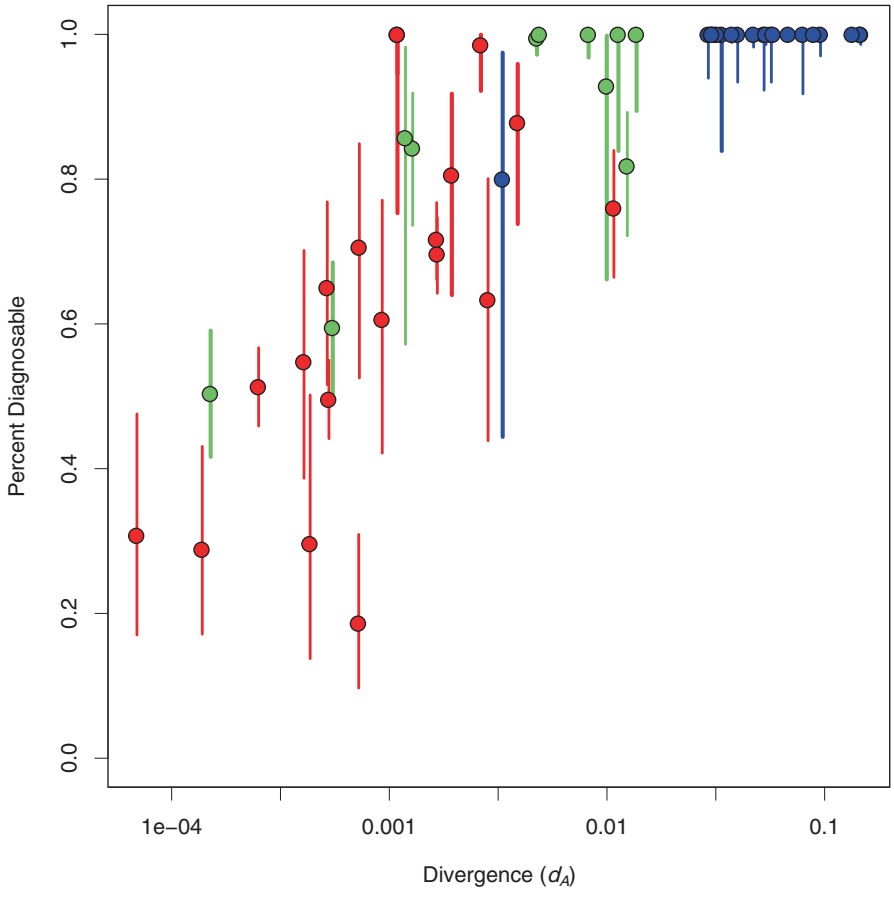

Figure 8. Relationship of diagnosability to Nei's net nucleotide divergence (d_A, Nei and Kumar 2000). Colors indicate taxonomic level of comparison: species (blue), subspecies (green), and populations (red). Vertical lines indicate the binomial 95% CI around diagnosability. Note that x-axis for d_A is $\log_{10}$-scaled.

divergence, as it is the combination of these two metrics that best define the boundary between populations and subspecies.

Effects of Sample Size

As seen in the simulated data, the number of individuals necessary for a given study will likely be case-specific, as it is a factor of both the effective population size and the degree of divergence among the strata under consideration. However, as sample size decreases, uncertainty in the diagnosability estimate increases, which can affect interpretation, especially if it is near the threshold. For example, if we had a model that correctly classified 59 of 70 individuals in a stratum, the diagnosability of that stratum would be 84%. Were the diagnosability threshold for subspecies status set at 90%, we may not believe the bar had been cleared. Based on a standard binomial distribution, the 95% confidence interval for this estimate is 74%–91%, and the probability that the true diagnosability is ≥90% is 0.046. Although there is a small probability that the threshold value is the true value, it might be close enough to warrant further investigation. However, if we hypothetically double the sample size and consider 118 correctly classified out of 140 individuals, the 95% CI becomes 77%–89%, and the probability that the true diagnosability is ≥90% decreases to 0.015. With this larger sample size, our certainty that the observed diagnosability does not meet the threshold increases.

In most genetic analyses, it is the haplotypic diversity that is of prime importance. This is also true for Random Forests, but in a slightly different manner. Because Random Forests operates on the level of the individual nucleotides rather than haplotypes, substitutions that have arisen within strata (synapomorphies) provide the strongest signal for the algorithm. Capturing this signal requires a sample size sufficiently large and broad to ensure that the data contain as many of these closely related haplotypes as possible. This difference in the way Random Forests treats sequences means that unique and rare haplotypes can contribute useful information to the classification algorithm in a way in which they cannot in standard frequency-based population genetics methods (Martien *et al.* 2017).

The most influential factor in interpreting the performance of a Random Forests model is the use of strata of extremely different individual sizes (Berk 2006). This issue is not unique to Random Forests, but rather affects many classification algorithms by creating an inherently biased classifier in which the larger stratum will tend to produce better classifications. This occurs for two reasons: (1) the presence of more individuals in a stratum can either provide more information for the classification algorithm to build a good model for the larger stratum, or give the appearance of lower variability in the smaller stratum; and (2) even in the absence of useful information for classification, individuals will tend to be classified to the larger stratum by chance alone.

The effect of the first issue cannot be quantified ahead of time given that it is influenced by the amount of overlap among strata as well as the distribution of diagnostic characters within strata. On the other hand, the effect of the second is simply related to the ratio of the individual sizes. For example, in the case of two strata, one (A) containing 75 individuals, and the other (B) containing 25 individuals, a random classifier will correctly classify 75% of the A's and 25% of the B's. But, because there are more A's than B's, the result is an overall percent correctly classified of 62.5% ($0.75 \times 75 + 0.25 \times 25$), larger than the 50% that one might expect based only on the fact that there were two strata. This deviation from 50% increases as the imbalance in

individual size increases (*e.g.*, a 95/5 split = 90.5% correct), potentially leading one to believe that the classifier is performing well, when in fact it is performing no better than could be expected by chance alone. The apparent improvement in overall performance comes at the cost of poorer classification for the smaller class. This imbalance in the classification rates is another sign of bias and a poorly trained model.

Recommendations

We recommend that estimates of diagnosability report both how well individuals can be assigned, and how certain one is of those assignments. Because diagnosability is an individual-based rather than population-based metric (Philips 1982), thresholds should be marker-independent. Thresholds based on the degree of accuracy one believes are required can be set *a priori*. Citing Amadon (1949), a 75% threshold is most frequently used. We have previously discussed what was intended by this value, and as an absolute threshold, we feel that it is not sufficiently high to definitively say that a subspecies is diagnosable. Being able to correctly classify three out of four individuals is not considerably different from random chance (50%). On the other end of the spectrum, a diagnosability of 95% has also been suggested (Patten and Unitt 2002, Patten 2010, Remsen 2010), usually in conformation with the traditional critical α of 0.05 used in frequentist statistics and similar to the 97% threshold actually described by Amadon (1949). Given the standard sample sizes used in most studies, we feel that a 95% threshold is actually too high. For example, with 60 individuals, one would have to have 58 individuals correctly classified to meet this bar. With two more misclassifications, the estimated diagnosability would be 93%. For a relatively well-sampled stratum, that means the difference between being diagnosable and not is four misclassifications. Thus, in our view, a threshold for minimal diagnosability of subspecies between 80% and 90% seems most appropriate. However, we do not advocate for any one threshold in this study. To properly do so requires a detailed examination of multiple standards and guidelines for delimiting subspecies, which is undertaken by (Taylor *et al.* 2017*a*).

Having an adequate number of individuals in a study will always be a critical factor for the multiple reasons that we have previously discussed. In the absence of fixed differences, a minimum of 30 individuals per stratum is likely necessary to begin to have reasonable estimates of diagnosability as each misclassification represents an approximately 3% decrease in diagnosability. However, even with 30 samples, the uncertainty around a diagnosability estimate can be quite large, and the only way to decrease it would be to add more individuals. In some cases, especially for endangered or otherwise rarely encountered species, collecting more samples may be unlikely to happen in a reasonable amount of time.

We were unable to find any discussions in the literature concerning how certain one should be about individual classifications when evaluating diagnosability. However, although not explicitly stated, these probabilities are implicit in the original formulation of the 75% rule. As illustrated in Figure 1B, because the two distributions have the same likelihood at the point where they cross and have equivalent overlap (97% of one lies outside of 97% of the other), the classification probability would be 50% to either distribution for an individual with a value of the diagnostic character at this threshold point. As one moves away from this point, the classification probabilities change proportional to the ratio of the likelihoods. The rate at which this occurs is dependent on the variance of the two distributions. Because sequence data are not continuously distributed, the Random Forests classification probabilities,

as defined by the distribution of tree votes, are not distributed in a predictably parametric manner, making comparisons with the 75% rule inappropriate. It is entirely possible to have a scenario in which all individuals in a stratum are correctly classified, but they are classified with only a plurality of trees, say 55%–65%, voting for the correct stratum. One would have less reliability in the classifications from that model than if the same individuals were being correctly classified in >90% of the trees.

One should take care to remove classification bias and achieve better parity among error rates when building the models. The suggested way for Random Forests is to have the algorithm build trees using strata of equal sizes (Berk 2006), which is the approach we have taken in building the models in this study. The effects of strata imbalance can also be addressed by setting the relative costs of the various misclassification errors (Berk 2008), but this becomes very difficult with more than two strata. One should also evaluate the performance of the model by comparing the results to what would be expected based on a simple random classification of individuals (PD_{prior}). This will avoid misinterpretations of high diagnosability that are likely only due to relatively large sample sizes.

In this study, each data set we examined was composed of only two strata. Remsen (2010) advocates that rules for diagnosability should only apply to pairwise comparisons such as these rather than simultaneous multiple comparisons of three or more strata where possible. Whether this is appropriate does not have a straightforward answer and may depend on the particular analysis. In a simultaneous multiple comparison analysis, the algorithm can better define classification rules for each individual stratum in the presence of all of the others. This use of all of the data at once puts diagnosability estimates of all putative strata on the same level. Additionally, misclassifications in these models can be used to help identify strata that do not represent valid taxonomic units, or strata that have not been adequately sampled and perhaps should not be included.

However, there are potential difficulties in interpreting the results of multiple pairwise comparisons. For example, if strata A and B are found to be reciprocally diagnosable as are strata B and C, but strata A and C are not, how does one determine the taxonomic status of the three? That is, features that distinguish two strata might not be as strongly diagnostic in the presence of a third stratum. Some comparisons in a multiple-strata analysis may not be necessary as in the case where strata are arranged in a stepping-stone pattern. If contact among strata at the ends is implausible, the degree of diagnosability of these seemingly allopatric units may not be important if the question is one of subspecific delimitation. The presence of these unlikely misassignments may decrease diagnosability estimates in strata of interest.

Future Directions

The benefit of a tool like Random Forests extends past its utility for estimating diagnosability for the taxonomic purpose of defining subspecies. The same classification engine developed for the delimitation of taxa can be used in forensic settings for the diagnosis of unknown specimens. Random Forests has been shown to perform well and is comparable to other methods for genetic barcoding (Austerlitz *et al.* 2009). However, unlike many other algorithms such as tree-building or those based on genetic distances, Random Forests also provides detailed estimates of classification uncertainty, and the ability to identify and rank diagnostic sites. Assessments of diagnosability need not be restricted to Random Forests (Bazinet and Cummings 2012).

Other algorithms such as Support Vector Machines are available and should be explored (Austerlitz *et al*. 2009, Seo 2010, Bazinet and Cummings 2012).

Although mtDNA has proven to be a very good marker for estimating diagnosability in this study, given the rapidly increasing role that genomic data are playing in population genetic and taxonomic studies, we fully expect it to be supplemented by large numbers of nuclear loci. The use of Random Forest for estimating diagnosability is not limited to mtDNA and should be readily transferrable to nucleotide sequences from any locus. Single nucleotide polymorphisms (SNP) can also be used by coding genotypes at each locus as three-state characters (two homozygotes and one heterozygote). Because it is becoming common to generate several thousand SNP loci for a study, they are rapidly becoming a popular source of diagnostic markers both for taxonomic studies as well as management of units of conservation concern (Kalinowski *et al.* 2011, Funk *et al.* 2012, Sousa and Hey 2013, Cronin *et al.* 2015).

Finally, because Random Forests is not based on an underlying evolutionary or population genetics model, loci that are neutrally evolving as well as those under selection can both be used side-by-side in the same analysis, as suggested by Funk *et al.* (2012). An integral feature of Random Forests is the ability to identify and rank predictor variables that contribute the most diagnostic information to the classification model. This is done by permuting the values in the predictor variables and observing the decrease in classification accuracy. Predictors that result in a large decrease in classification accuracy when permuted are considered to be more "important" to the classifier (Liaw and Wiener 2002). With whole genome data sets, these importance measures should prove to be useful tools for identifying loci that are candidates for understanding local adaptation and drivers of the speciation process.

Acknowledgments

The authors wish to acknowledge the efforts of Brittany Hancock-Hanser, Kelly Robertson, and Victoria Pease in assembling the sequences used. We also would like to thank Brittany Hancock-Hanser, Cleridy Lennert-Cody, Philip Morin, Patricia Rosel, and three anonymous reviewers for helpful comments.

Literature Cited

Amadon, D. 1949. The seventy-five per cent rule for subspecies. The Condor 51:250–258.

Archer, F. I., P. A. Morin, B. L. Hancock-Hanser, *et al*. 2013. Mitogenomic phylogenetics of fin whales (*Balaenoptera physalus* spp.): Genetic evidence for revision of subspecies. PLoS ONE 8:e63396.

Austerlitz, F., O. David, B. Schaeffer, *et al.* 2009. DNA barcode analysis: A comparison of phylogenetic and statistical classification methods. BMC Bioinformatics 10:S10.

Barrowclough, G. F. 1982. Geographic variation, predictiveness, and subspecies. Auk 99:601–603.

Baum, D. A., and M. J. Donoghue. 1995. Choosing among alternative "phylogenetic" species concepts. Systematic Botany 20:560–573.

Bazinet, A. L., and M. P. Cummings. 2012. A comparative evaluation of sequence classification programs. BMC Bioinformatics 13:92.

Berk, R. 2006. An introduction to ensemble methods for data analysis. Sociological Methods and Research 34:263–295.

Berk, R. A. 2008. Statistical learning from a regression perspective. Springer, New York, NY.

Bradley, R. D., and R. J. Baker. 2001. A test of the genetic species concept: Cytochrome-b sequences and mammals. Journal of Mammalogy 82:960–973.

Brambilla, M., S. Vitulano, A. Ferri, F. Spina, E. Fabbri and E. Randi. 2009. What are we dealing with? An explicit test reveals different levels of taxonomical diagnosability in the *Sylvia cantillans* species complex. Journal of Ornithology 151:309–315.

Breiman, L. 2001. Random forests. Machine Learning 45:5–32.

Caballero, S., F. Trujillo, J. A. Vianna, *et al*. 2007. Taxonomic status of the genus *Sotalia*: Species level ranking for "tucuxi" (*Sotalia fluviatilis*) and "costero" (*Sotalia guianensis*) dolphins. Marine Mammal Science 23:358–386.

Castro, L., and M. A. Toro. 1995. Human evolution and the capacity to categorize. Journal of Social and Evolutionary Systems 18:55–66.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics 10:195–205.

Cicero, C., and N. K. Johnson. 2006. Diagnosability of subspecies: Lessons from sage sparrows (*Amphispiza belli*) for analysis of geographic variation in birds. The Auk 123:266–274.

Cracraft, J. 1983. Species concepts and speciation analysis. Current Ornithology 1:159–187.

Cronin, M. A. 1993. Mitochondrial DNA in wildlife taxonomy and conservation biology: Cautionary notes. Wildlife Society Bulletin 21:339–348.

Cronin, M. A., A. Cánovas, D. L. Bannasch, M. Oberbauer and J. F. Medrano. 2015. Single nucleotide polymorphism (SNP) variation of wolves (*Canis lupus*) in southeast Alaska and comparison with wolves, dogs, and coyotes in North America. Journal of Heredity 106:26–36.

Cutler, D. R., T. C. J. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen and M. L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12:499–510.

de Queiroz, K. 2007. Species concepts and species delimitation. Systematic Biology 56:879–886.

Eckert, A. J., and B. C. Carstens. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Molecular Phylogenetics and Evolution 49:832–842.

Excoffier, L., and M. Foll. 2011. Fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27:1332–1334.

Foote, A. D., P. A. Morin, J. W. Durban, E. Willerslev, L. Orlando and M. T. Gilbert. 2011. Out of the Pacific and back again: Insights into the matrilineal history of Pacific killer whale ecotypes. PLoS ONE 6:e24980.

Frankham, R. 1996. Relationship of genetic variation to population size in wildlife. Conservation Biology 10:1500–1508.

Funk, W. C., J. K. McKay, P. A. Hohenlohe and F. W. Allendorf. 2012. Harnessing genomics for delineating conservation units. Trends in Ecology & Evolution 27:489–496.

Haig, S. M., E. A. Beever, S. M. Chambers, *et al*. 2006. Taxonomic considerations in listing subspecies under the U.S. Endangered Species Act. Conservation Biology 20:1584–1594.

Harlin, A. D., T. Markowitz, C. S. Baker, B. Würsig and R. L. Honeycutt. 2003. Genetic structure, diversity, and historical demography of New Zealand's dusky dolphin (*Lagenorhynchus obscurus*). Journal of Mammalogy 84:702–717.

Hayano, A., M. Yoshika, M. Tanaka and M. Amano. 2004. Population differentiation in the Pacific white-sided dolphin *Lagenorhynchus obliquidens* inferred from mitochondrial DNA and microsatellite analyses. Zoological Science 21:989–999.

Helbig, A. J., A. G. Knox, D. T. Parkin, G. Sangster and M. Collinson. 2002. Guidelines for assigning species rank. Ibis 144:518–525.

Heyer, E., E. Zietkiewicz, A. Rochowski, V. Yotova, J. Puymirat and D. Labuda. 2001. Phylogenetic and familial estimates of mitochondrial substitution rates: Study of control region mutations in deep-rooting pedigrees. American Journal of Human Genetics 69:1113–1126.

Hoelzel, A. R., J. M. Hancock and G. A. Dover. 1991. Evolution of the cetacean mitochondrial D-loop region. Molecular Biology and Evolution 8:475–493.

Hoelzel, A. R., J. Hey, M. E. Dahlheim, C. Nicholson, V. Burkanov and N. Black. 2007. Evolution of population structure in a highly social top predator, the killer whale. Molecular Biology and Evolution 24:1407–1415.

Jackson, J. A., C. S. Baker, M. Vant, D. J. Steel, L. Medrano-Gonzalez and S. R. Palumbi. 2009. Big and slow: Phylogenetic estimates of molecular evolution in baleen whales (suborder Mysticeti). Molecular Biology and Evolution 26:2427–2440.

Jefferson, T. A., and J. Y. Wang. 2011. Revision of the taxonomy of finless porpoises (genus *Neophocaena*): The existence of two species. Journal of Marine Animals and Their Ecology 4:3–16.

Johnson, N. K., J. V. J. Remsen and C. Cicero. 1999. Resolution of the debate over species concepts in ornithology: A new comprehensive biologic species concept. Pages 1470–1482 *in* 22nd Ornithological Congress. BirdLife South Africa, Durban, South Africa.

Kalinowski, S., B. J. Novak, D. P. Drinan, R. D. Jennings and N. V. Vu. 2011. Diagnostic single nucleotide polymorphisms for identifying westslope cutthroat trout (*Oncorhynchus clarki lewisi*), Yelowstone cutthroat trout (*Oncorhynchus clarkii bouvieri*) and rainbow trout (*Oncorhynchus mykiss*). Molecular Ecology Resources 11:389–393.

Lee, M. S. Y. 2003. Species concepts and species reality: Salvaging a Linnaen rank. Journal of Evolutionary Biology 16:179–188.

Li, S.-H., J.-W. Li, L.-X. Han, C.-T. Yao, H. Shi, F.-M. Lei and C. Yen. 2006. Species delimitation in the Hwamei *Garrualx canorus*. Ibis 148:698–706.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2 (3):18–22.

Martien, K. K., S. J. Chivers, R. W. Baird, *et al*. 2014. Nuclear and mitochondrial patterns of population structure in North Pacific false killer whales (*Pseudorca crassidens*). Journal of Heredity 105:611–626.

Martien, K. K., M. S. Leslie, B. L. Taylor, *et al*. 2017. Analytical approaches to subspecies delimitation with genetic data. Marine Mammal Science 33(Special Issue):27–55.

Mayr, E. 1942. Systematics and the origin of species from the viewpoint of a zoologist. Columbia University Press, New York, NY.

Mayr, E. 1969. Principles of systematic zoology. McGraw-Hill, New York, NY.

McGowan, M. R., M. Spaulding and J. Gatesy. 2009. Divergence date estimation and a comprehensive molecular tree of extant cetaceans. Molecular Phylogenetics and Evolution 53:891–906.

Mills, L. S., and F. W. Allendorf. 1996. The one-migrant-per-generation rule in conservation and management. Conservation Biology 10:1509–1518.

Morin, P. A., F. I. Archer, A. D. Foote, *et al*. 2010. Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. Genome Research 20:908–916.

Patten, M. A. 2010. Null expectations in subspecies diagnosis. Ornithological Monographs 67:35–41.

Patten, M. A., and P. Unitt. 2002. Diagnosability versus mean differences of sage sparrow subspecies. Auk 119:26–35.

Philips, A. R. 1982. Subspecies and species: Fundamentals, needs, and obstacles. The Auk 99:612–615.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Remsen, J. V. J. 2010. Subspecies as a meaningful taxonomic rank in avian classification. Ornithological Monographs 67:62–78.

Rosel, P. E., B. L. Hancock-Hanser, F. I. Archer, *et al*. 2017*a*. Examining metrics and magnitudes of genetic differentiation used to delimit cetacean subspecies based on mitochondrial DNA control region sequences. Marine Mammal Science 33(Special Issue):76–100.

Rosel, P. E., B. L. Taylor, B. L. Hancock-Hanser, *et al*. 2017*b*. A review of genetic markers and analytical approaches that have been used for delimiting marine mammal subspecies and species. Marine Mammal Science 33(Special Issue):56–75.

Seo, T. 2010. Classification of nucleotide sequences using support vector machines. Journal of Molecular Evolution 71:250–267.

Sigurðardóttir, S., A. Helgason, J. R. Glcher, K. Stefansson and P. Donnelly. 2000. The mutation rate in the human mtDNA control region. American Journal of Human Genetics 66:1599–1609.

Sites, J. W., and J. C. Marshall. 2004. Operational criteria for delimiting species. Annual Review of Ecology Evolution and Systematics 35:199–227.

Sousa, V., and J. Hey. 2013. Understanding the origin of species with genome-scale data: Modelling gene flow. Nature Reviews: Genetics 14:404–414.

Taylor, B. L., S. J. Chivers, J. Larese and W. F. Perrin. 2007. Generation length and percent mature estimates for IUCN assessments of cetaceans. NOAA Administrative Report LJ-07-01:24.

Taylor, B. L., F. I. Archer, K. K. Martien, *et al*. 2017*a*. Guidelines and quantitative standards to improve consistency in cetacean subspecies and species delimitation relying on molecular genetic data. Marine Mammal Science 33(Special Issue):132–155.

Taylor, B. L., W. F. Perrin, R. R. Reeves, *et al*. 2017*b*. Why we should develop guidelines and quantitative standards for using genetic data to delimit subspecies for data-poor organisms like cetaceans. Marine Mammal Science 33(Special Issue):12–26.

Touw, W. G., J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels and S. a. F. T. Van Hijum. 2012. Data mining in the life sciences with random forest: A walk in the park or lost in the jungle? Briefings in Bioinformatics 14:315–326.

Vilstrup, J. T., S. Y. Ho, A. D. Foote, *et al*. 2011. Mitogenomic phylogenetic analyses of the Delphinidae with an emphasis on the Globicephalinae. BMC Evolutionary Biology 11:65.

Wade, P. R., G. M. Watters, T. Gerrodette and S. B. Reilly. 2007. Depletion of spotted and spinner dolphins in the eastern tropical Pacific: Modeling hypotheses for their lack of recovery. Marine Ecology Progress Series 343:1–14.

Wang, J. 2004. Application of the one-migrant-per-generation rule to conservation and management. Conservation Biology 18:332–343.

Wang, J. Y., S. C. Yang, B. J. Wang and L. S. Wang. 2010. Distinguishing btween two species of finless porpoises (*Neophocaena phocaenoides* and *N. asiaorientalis*) in areas of sympatry. Mammalia 74:305–310.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7:256–276.

Wheeler, Q. D. 1999. Why the phyl2ogenetic species concept?—Elementary. Journal of Nematology 31:134–141.

Wilson, E. O. and W. L. J. Brown. 1953. The subspecies concept and its taxonomic application. Systematic Zoology 2:92–111.

Winham, S. J., C. L. Colby, R. R. Freimuth, X. Wang, M. De Andrade, M. Huebner and J. M. Biernacka. 2012. SNP interaction detection with Random Forests in high-dimensional genetic data. BMC Bioinformatics 13:164.

Winker, K. 2010. Subspecies represent geographically partitioned variation, a gold mine of evolution biology, and a challenge for conservation. Ornithological Monographs 67:6–23.

Wood, S. 2006. Generalized additive models: An introduction with R. CRC Press, Boca Raton, FL.

Zink, R. M., and G. F. Barrowclough. 2008. Mitochondrial DNA under siege in avian phylogeography. Molecular Ecology 17:2107–2121.

Zink, R. M., and J. I. Davis. 1999. New perspectives on the nature of species. Pages 1505–1518 *in* N. J. Adams, and R. H. Slotow, eds. Proceedings 22nd International Ornithological Congress, Durban. BirdLife South Africa, Johannesburg, South Africa.

Received: 25 March 2015
Accepted: 10 January 2017