Issues and Perspectives in Species Delimitation using Phenotypic Data: Atlantean Evolution in Darwin’s Finches

CARLOS DANIEL CADENA1,*, FELIPE ZAPATA2, and IVÁN JIMÉNEZ3

1Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá, Colombia; 2Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA, USA; and 3Center for Conservation and Sustainable Development, Missouri Botanical Garden, St. Louis, MO, USA

*Correspondence to be sent to: Departamento de Ciencias Biológicas, Universidad de los Andes, Carrera 1 no. 18A-10, Bogotá, Colombia;
E-mail: ccadena@uniandes.edu.co.

Received 5 April 2017; reviews returned 22 August 2017; accepted 30 August 2017
Associate Editor: John E. McCormack

Abstract.—Progress in the development and use of methods for species delimitation employing phenotypic data lags behind conceptual and practical advances in molecular genetic approaches. The basic evolutionary model underlying the use of phenotypic data to delimit species assumes random mating and quantitative polygenic traits, so that phenotypic distributions within a species should be approximately normal for individuals of the same sex and age. Accordingly, two or more distinct normal distributions of phenotype traits suggest the existence of multiple species. In light of this model, most analytical approaches employed in taxonomic studies using phenotypic data are often compromised by three issues: 1) reliance on graphical analyses that convey little information on phenotype frequencies; 2) exclusion of characters potentially important for species delimitation following reduction of data dimensionality; and 3) use of measures of central tendency for species delimitation (Luckow 1995; McDade 1995; Sangster 2014; Almon 2016), the use of objective criteria for species diagnosis based on phenotypic characters has a long tradition in taxonomy, rooted in evolutionary theory (Wiens and Servedio 2000; Zapata and Jiménez 2012; Futuyama 2013). The basic evolutionary model for the distribution of a continuous quantitative character within a species (Fisher 1918) assumes polygenic inheritance and random mating; under these assumptions, gene frequencies would be close to Hardy–Weinberg equilibrium; two or more loci would be near linkage equilibrium, and phenotypic variation among individuals of a single species would tend to be normally distributed (Templeton 2006). On the other hand, phenotypic variation may be best described by two or more distinct normal distributions (i.e., distributions differing in means, variances, or covariances); in this latter case, one may conclude that there is more than one species in a sample of individuals (Coyne and Orr 2004). Thus, we show that the assumption that parameter differences between normal distributions do not reflect genetic

Systematic biology seeks to discover and describe species, and to establish phylogenetic relationships among them and among clades at higher levels. Given these two main goals of the field, reviews published over a decade ago noted that the literature on theory and methods of phylogenetic inference and on theory of species concepts was extensive, whereas methods for delimiting species had received much less attention (Sites Jr. and Marshall 2003; Sites Jr. and Marshall 2004). Over the past few years, this imbalance has been partly overcome with considerable development, application, and integration of methods for species delimitation (Padial et al. 2010; Camargo and Sites Jr. 2013). Largely driven by increased availability of multilocus data sets brought about by advances in DNA sequencing technology, however, much recent progress has focused on probabilistic methods for analyses of molecular data (reviewed by Fujita et al. 2012; Carstens et al. 2013), whereas relatively little effort has been devoted to approaches using phenotypic data to delimit species (Wiens and Servedio 2000; Ezard et al. 2010; Guillot et al. 2012; Zapata and Jiménez 2012; Edwards and Knowles 2014; Solís-Lemus et al. 2014). Yet, because most fossil and living species have been discovered and named based on phenotypic distinctiveness (Luckow 1995; Mallet 2013; Miller 2016), and because genomic-based species delimitation approaches are no substitute for judicious assessments of other sources of information (Sukumaran and Knowles 2017), the theory and practice of delimiting species using phenotypic data remain central to modern systematics.

Although species descriptions employing phenotypic data are often nonquantitative and although systematists may often not be explicit about the rationale they follow to delimit species (Luckow 1995; McDade 1995; Sangster 2014; Almon 2016), the use of objective criteria for species diagnosis based on phenotypic characters has a long tradition in taxonomy, rooted in evolutionary theory (Wiens and Servedio 2000; Zapata and Jiménez 2012; Futuyama 2013). The basic evolutionary model for the distribution of a continuous quantitative character within a species (Fisher 1918) assumes polygenic inheritance and random mating; under these assumptions, gene frequencies would be close to Hardy–Weinberg equilibrium; two or more loci would be near linkage equilibrium, and phenotypic variation among individuals of a single species would tend to be normally distributed (Templeton 2006). On the other hand, phenotypic variation may be best described by two or more distinct normal distributions (i.e., distributions differing in means, variances, or covariances); in this latter case, one may conclude that there is more than one species in a sample of individuals (Coyne and Orr 2004). Thus, we show that the assumption that parameter differences between normal distributions do not reflect genetic
polymorphisms (e.g., sex-related variation), ontogenetic variation, or phenotypic plasticity. Therefore, differences between normal distributions caused by few loci of large effect (e.g., Smith 1993) or largely driven by environmental factors (e.g., Moczek and Emlen 1999) do not constitute evidence of more than one species. While this Fisherian model readily applies to (polygenic) continuous traits and not to qualitative traits, we note that phenotypic variation commonly regarded as qualitative is actually continuous (Stevens 1991), including variation in the shape (e.g., Leaché et al. 2009) and color (e.g., McKay et al. 2014) of morphological structures.

We stress that distinct phenotypic distributions may represent evidence of species boundaries given a variety of species definitions (sensu de Queiroz 1998). For instance, distinct phenotypic distributions may serve as a species criterion (i.e., as a standard to judge whether a group of organisms qualifies as a species) under species definitions that emphasize phenotypic and genotypic clusters (e.g., Ezard et al. 2010), interbreeding (e.g., Mayr 1992), phenotypic cohesion (e.g., Bond and Stockman 2008), or diagnosability (e.g., Crisp and Weston 1995). The same is true under species definitions that are alternative descriptions of the general lineage concept of de Queiroz (1998), including the evolutionary species definitions of Simpson (1951) and Wiley (1978). Therefore, the Fisherian model described above serves as a conceptual basis to infer species limits given diverse views regarding the nature of species.

Despite the long tradition of the basic model for species delimitation based on quantitative phenotypic characters, statistical tools for its formal application to empirical data were fairly limited until recently. Procedures allowing one to fit combinations of normal distributions to phenotypic variation among specimens, without a priori knowledge of species limits, were initially developed in the late XIX century (Pearson 1894). However, practical application only became possible following computational advances in the 1970s (i.e., the expectation–maximization algorithm; McLachlan and Peel 2000) and software development from the late XX century into the present (e.g., Frayley and Raftery 2002, Fraley et al. 2012). Because these statistical approaches entered the literature on species delimitation only a few years ago (Ezard et al. 2010; see also Haußdorf and Hennig 2010 for an application to molecular data), it is not surprising that even recent studies do not employ them when analyzing phenotypic data to delimit species. Instead, systematists frequently infer species limits examining phenotypic variation based on visual inspection of scatter plots defined by a few axes that account for most phenotypic variance, often derived from principal components analysis (PCA). In addition, systematists often delimit species based on differences between groups of specimens in the central tendency of phenotypes. This is true of work on living plants and animals (reviewed by Rieseberg et al. 2006), as well as in studies of extinct taxa in the fossil record (reviewed by Allmon 2016).

Here, we show that the way in which analytical approaches are commonly employed to examine phenotypic data in taxonomic studies is often inadequate in light of the evolutionary model underlying species delimitation described above. It follows that if species delimited by inadequate statistical approaches are used as units for subsequent analyses, then any mistakes may carry on and influence views in other areas of inquiry, such as speciation research. Focusing on Darwin’s finches from the Galapagos Islands, an iconic group for the study of natural selection, speciation, and adaptive radiation (Lack 1947; Bowman 1961; Grant 1999; Grant and Grant 2008; Grant and Grant 2014), we provide an example of how employing statistical approaches explicitly related to the basic evolutionary model underlying the use of phenotypic data in species delimitation may enhance assessments of species limits and thus our understanding of evolutionary processes.

**Sisyphean Evolution in Darwin’s Finches?**

Among Darwin’s finches, the many studies of ground-finches in the genus *Geospiza* have been especially productive in terms of insights into species formation and the role of geographic isolation, natural selection, and hybridization in microevolutionary processes that may scale up to macroevolutionary patterns (reviewed by Grant 1999; Grant and Grant 2008; Grant and Grant 2014). There has been considerable disagreement in the literature about the number of species in the group (reviewed by McKay and Zink 2015), but most modern taxonomic treatments have recognized six species of ground-finches (Lack 1947; Rising et al. 2011). However, based on genomic evidence (Lamichhaney et al. 2015) and some vocal and behavioral data, three subspecies were recently elevated to species rank, bringing the total number of recognized species to nine (Remsen Jr. et al. 2017).

In a provocative recent paper, McKay and Zink (2015) offered an intriguing alternative perspective on the taxonomy and evolution of ground-finches (see also Zink 2002). These authors boldly argued that morphological evidence for the existence of multiple species of *Geospiza* is lacking and they presented the iconoclastic argument that different phenotypes should be considered transient ecomorphs within a single species. Furthermore, according to these authors, ground-finches are an appropriate model to study forces involved in geographic variation and local adaptation, but not to demonstrate the workings of speciation because in their view speciation in the group has not occurred. Instead, incipient speciation has been repeatedly stalled or reversed owing to shifting conditions affecting the strength and direction of natural selection and to ongoing gene flow (McKay and Zink 2015). Because speciation is initiated but never completed, McKay and Zink (2015) described evolution in ground-finches as “Sisyphean”
in reference to Sisyphus, a character in Greek mythology condemned by the gods to ceaselessly push a boulder up a mountain, only to watch it roll back down, repeating this task eternally. Because of its originality in challenging “entrenched orthodoxy regarding speciation in Darwin’s Finches,” the study by McKay and Zink (2015) was duly recognized with an award by a major ornithological organization (Cooper Ornithological Society 2016).

A central premise of the arguments by McKay and Zink (2015) was their assertion that phenotypic discontinuities do not exist among recognized species of ground-finches (contra Lack 1947; Grant et al. 1985). This claim seems particularly important in light of recent work indicating that some pairs of species (i.e., Geospiza fortis vs. G. acutirostris, G. scandens vs. G. propinqua; G. magnirostris vs. G. conirostris) appear not to be clearly distinguishable from each other with genomic data (Fig. 1b in Lamichhaney et al. 2015).

Although McKay and Zink (2015) rightly noted that given currently available phenotypic data “the real test of species limits is determining the extent to which specimens form multiple morphological clusters when a priori specimen identifications are ignored,” they did not formally conduct such a test. Instead, their approach illustrates three problematic issues in analyses of phenotypic data for species delimitation. In the next section, we describe these issues and outline possible solutions afforded by statistical tools directly related to the basic evolutionary model underlying the use of phenotypic data in species delimitation. We then implement these solutions in a reanalysis of the morphological data on Geospiza ground-finches to revisit the question of whether morphological evidence supports the hypothesis that there are several species in the group.

THREE FREQUENT ISSUES IN ANALYSES OF PHENOTYPIC DATA FOR SPECIES DELIMITATION

Graphical analyses may convey little information on phenotype frequencies crucial to assess evidence for multiple species

Many species delimitation studies rely on visual inspection of bivariate (rarely trivariate) scatter plots of phenotypic space to detect discontinuities and thus define phenotypic groups (e.g., Fig. 1 in McKay and Zink 2015). These scatter plots may offer only limited insight into the structure of character variation because visual cluttering and record overplotting hinder perception of phenotype frequencies crucial to identify groups (McLachlan 2004). We illustrate this problem with a hypothetical example in which specimens from a given locality seem to reveal no phenotypic discontinuities, with intermediate phenotypes across the range of morphological trait.

![Figure 1](https://example.com/figure1.png)

**FIGURE 1.** Visual inspection of phenotypic data may yield limited insight regarding species limits. A) Sample of 200 museum specimens (triangles) arranged according to a morphological phenotype (triangle size), from small in the upper left to large in the lower right. The specimens appear to form a smooth gradient with no morphological gaps. B) Plot of specimen measurements along a single continuous axis representing the size of the morphological trait in (A). At the resolution of the measurements, extreme values in the sample seem to be gradually connected by intermediate phenotypes throughout. Thus, there seem to be no obvious morphological gap, suggesting the specimens correspond to a single variable species. C) Two distinct normal distributions are revealed by examining the frequency (gray bars) of specimen phenotypes in the sample, suggesting that specimens may correspond to two species. In fact, the sample was drawn from a mixture of two normal distributions (continuous black lines).
variation (Fig. 1a); accordingly, a univariate scatter plot fails to reveal evidence of distinct phenotypic groups (Fig. 1b). The problem with scatter plots concealing crucial information (also common in 2D and 3D scatter plots) is revealed by a histogram of phenotype frequencies employing the same data, which reveals two distinct normal distributions (Fig. 1c). Following the model for species delimitation based on continuous phenotypic characters described above, this histogram suggests the existence of two species.

Graphical analysis of phenotype frequencies (e.g., Fig. 1c) may be effective to detect groups when few characters are relevant (but see McLachlan and Peel 2000, page 9). However, it may be difficult to detect distinct normal distributions in phenotypic spaces defined by more than two dimensions, where complex covariance structures are likely (McLachlan 2004). Moreover, graphical analysis may suggest the existence of several distinct distributions when, in fact, all variation derives from a single normal distribution (Day 1969; McLachlan and Peel 2000, page 17; Supplementary Material Appendix 1 at http://dx.doi.org/10.5061/dryad.9g9r0). In general, then, detection of phenotypic groups exclusively based on graphical analysis is potentially highly subjective and difficult to replicate, or, as stated by Pearson (1894) over a century ago: “To throw the solution on the judgment of the eye in examining the graphical results is, I feel certain, quite futile.” Therefore, graphical analysis of phenotype frequencies is a useful but limited tool for species delimitation.

Recent statistical developments allow systematists to go beyond graphical analysis by using normal mixture models (NMMs, McLachlan and Peel 2000) as a formal approach to test for the existence of distinct species based on multivariate phenotypic data (Ezard et al. 2010; Guillot et al. 2012; Edwards and Knowles 2014; Kleindorfer et al. 2014). These models conceptualize phenotypic variation as a combination (i.e., a mixture) of distinct normal distributions; a mixture may include one or more distinct normal distributions, representing the hypothesis of one or more species, respectively. The parameters of a NMM specifying a particular hypothesis include the means and variance–covariance matrices describing the Gaussian phenotypic distribution of each species. These parameters can be estimated using maximum likelihood from data on phenotypic measurements, without a priori knowledge of species limits, employing the expectation–maximization algorithm (McLachlan and Krishnan 2008). Comparison of empirical support among models representing different hypotheses is often based on the Bayesian Information Criterion (BIC; Schwarz 1978), which evaluates the likelihood of each model while adjusting for model complexity (Fraley and Raftery 2002).

Reduction of dimensionality via PCA may exclude important characters for species delimitation

Species delimitation studies often begin analyses by reducing the dimensionality of phenotypic space, typically via principal component analysis (PCA) or related procedures (McLachlan 2004; Ezard et al. 2010), and then focusing attention on few principal components that account for most of the variation in the data. For example, McKay and Zink (2015) focused on three principal components explaining >99% of the variation in six morphological characters of Geospiza ground-finches (see their Fig. 1). This use of PCA and related procedures in taxonomy was suggested decades ago (Sneath and Sokal 1973) and is still prescribed nowadays (e.g., Ezard et al. 2010). However, there is no reason to believe that principal components accounting for most of the variation in a data set are most useful for group discrimination (Chang 1983).

To illustrate the problem of reducing dimensionality to the principal components accounting for most of the variation, we use a hypothetical example based on two phenotypically distinct species, each represented by a bivariate normal distribution (Fig. 2a). The first principal component of the mixture of these two distributions explains >99% of the variation and, yet, it is useless to distinguish the two species (Fig. 2b). In contrast, the second principal component accounts for <1% of the variation and readily discriminates species (Fig. 2c). This example is bivariate for simplicity, but the statistical principle applies to mixtures of two normal distributions in any number of dimensions (Chang 1983). We stress that the problem at hand is not rotation of the data using PCA or related procedures, because such rotation may serve a number of useful purposes; rather, the problem is using the amount of phenotypic variance explained by each principal component as a proxy for its usefulness to distinguish phenotypic groups (Chang 1983).

Although alternatives to PCA and related approaches for dimensionality reduction should be regularly considered in analysis aiming to detect groups in multivariate space (McLachlan and Peel 2000; McLachlan 2004), they are rarely implemented in species delimitation studies. For example, one may reduce dimensionality based on a priori considerations about which set of characters may be best to diagnose particular species and then use those characters in analyses based on NMMs. In particular, when a priori information about specific traits separating species is available (e.g., original species descriptions), one should favor analyzing variation in such traits; far from being circular (McKay and Zink 2015), it is only natural that one should critically examine evidence for species limits precisely in the dimensions in which such limits are hypothesized to exist (see also Rensen Jr. 2010; Patten and Rensen Jr. 2017). Alternatively, one may use methods that aim to find the set of variables (phenotypic traits) that best discriminates groups in a NMM, with no a priori information about groups (Raftery and Dean 2006; Maugis et al. 2009a; Maugis et al. 2009b).
Differences in central tendency are not evidence of distinct phenotypic distributions

Distinct normal distributions in quantitative characters constitute evidence for the existence of distinct species, but differences in central tendency between groups of individuals defined a priori do not. This issue has been pointed out previously (e.g., Mayr et al. 1953; Luckow 1995; Patten and Unitt 2002), but seems to be ignored when statistical procedures to investigate differences in central tendency (e.g., t-tests, analysis of variance, Cohen's $d$) are advanced as potentially valid tools to evaluate species boundaries (e.g., Simpson 1951; Henderson 2006; Tobias et al. 2010). McKay and Zink (2015, page 695 and their Fig. 2) have done as much by suggesting that statistical differences in average phenotypes between allopatric island populations of ground-finches could be equated to distinct morphological groups which, in turn, would have to be recognized as species.

Because this issue appears to commonly afflict assessments of species limits between allopatric forms (e.g., Tobias et al. 2010), we illustrate it with a hypothetical example likely encountered in many studies of species delimitation: spatially separated populations of a species exhibiting geographic variation in phenotype. Following the analysis of McKay and Zink (2015, their Fig. 2), we imagine researchers sampling individuals of a species in two islands and then documenting significant differences in the central tendency of phenotypes between the two island populations (Fig. 3a). This difference in central tendency, however, does not constitute evidence that phenotypic variation is best described by two (or more) distinct normal distributions. In fact, the relevant NMM analysis indicates that a single normal distribution best explains phenotypic variation across individuals from the two island populations (Fig. 3b), consistent with the Fisherian model for a single species. Therefore, there is no evidence for more than one species in the sample of specimens regardless of differences in average phenotypes. This illustration focuses on a priori groups of specimens defined by geography (i.e., spatially separated populations), but the issue may affect comparisons involving groups of specimens defined by time (i.e., allochronic populations; Simpson 1951) or by any other criterion. In general, phenotypic variation conforming to a single normal distribution can be arbitrarily split into parts that differ in central tendency (Supplementary Material Appendix 2 available on Dryad). Therefore, on their own, differences in central tendency between groups of specimens cannot be regarded as evidence for distinct normal distributions.

The solution to the above problem is simple: do not treat phenotypic differences in central tendency as evidence for the existence of distinct phenotypic groups and, therefore, distinct species. No matter how statistically significant, even very large effect sizes are not germane in light of the basic model for species delimitation based on quantitative phenotypic characters. In light of this model, the focus of analysis should be on determining the number of normal distributions needed to describe phenotypic variation among specimens, as well as on estimating the parameters of those distributions (e.g., means and standard deviations).
FIGURE 3. Differences in central tendency are not evidence of distinct phenotypic distributions. A) Specimens from two island populations a and b (striped and gray histogram bars, respectively) differ markedly in the interquartile range (box), whiskers extending to the most extreme values within ±1.5 interquartile ranges from the box, and outliers. The means of the two groups are 1.717 and 2.707, and their standard deviations are 0.3528 and 0.3867, respectively. The difference between means is statistically significant (0.99, 95% CI: 0.89–0.99). Cohen’s d = 2.66, which is generally regarded as a large effect size. B) Despite the difference in central tendency, a single normal distribution (continuous black line) describes phenotypic variation across specimens from both islands (gray bars) better than two normal distributions. In particular, empirical support for a normal mixture model assuming that populations a and b constitute two distinct normal distributions is substantially lower than that for a model specifying a single normal distribution: a difference of 10 in BIC. Thus, in light of the basic model for species delimitation based on quantitative phenotypic characters, there are no grounds to suggest that the specimens represent two distinct species despite marked differences in central tendency.

ARE THERE PHENOTYPICALLY DISTINCT GROUPS OF GROUND-FINCHES?

We examined phenotypic variation among Geospiza ground-finches by analyzing data from six morphological measurements of museum specimens (wing length, tail length, tarsus length, bill length, bill width, and bill depth) taken on adult males by H. S. Swarth for his monographic revision of the birds of the Galapagos (Swarth 1931). These were the same data employed by McKay and Zink (2015), which we use here with permission from the California Academy of Sciences; our sample sizes differ from those of the earlier study (486 vs. 501 male individuals) because we excluded a few individuals that were duplicated in the original data set. The data we employed and the R code used to conduct the analyses described below are available as Supplementary Material Appendices 3 and 4 available on Dryad, respectively. We emphasize that the purpose of our analysis was not to establish species limits among ground-finches nor to provide species diagnoses. Rather, we sought to revisit the question of whether the morphological data analyzed by McKay and Zink (2015) are consistent with the existence of multiple species and thereby illustrate an empirical application of approaches involving variable selection and NMMs for species delimitation. As we argue below, the morphological data we analyzed are one of multiple sources of evidence that researchers may use for subsequent assessments of species limits under a variety of species definitions (sensu de Queiroz 1998). We asked how many distinct groups of ground-finches exist in the Galapagos using morphological data from specimens collected across the archipelago (total 18 islands). To define the morphological space for this analysis, we followed McKay and Zink (2015) and used PCA on the covariance matrix of log-transformed data. Rather than examining evidence for species limits using only the first three principal components accounting for >99% of the variation (McKay and Zink 2015), we used the R package clustvarsel (Scrucca et al. 2016) to fit a wide range of NMMs. At one extreme, NMMs assuming one morphological group represented the Sisyphean evolution hypothesis that there is a single species of ground-finch...
Fig. 2 available on Dryad). For example, in the best
1947; Remsen Jr. et al. 2017) were partially consistent
according to taxonomic treatments of
models specifying six or nine morphological groups
Fig. 5, Supplementary Material Fig. 1 available on
McKay and Zink (2015) morphologically distinct groups
supported hypotheses of several (but not dozens or 30;
Fig. 4), considering differences in BIC scores greater
recognizing six or nine species were weakly supported
hypotheses that there are several species in the group (Lack 1947;
Lamichhaney et al. 2015; Remsen Jr. et al. 2017).

We found the first four principal components to be most useful for group discrimination; NMMs
ignoring the fourth principal component, although it explained only 0.6% of the morphological variance, had
substantially less empirical support (ΔBIC > 55) than those including it. Therefore, in contrast to McKay and
Zink (2015), we did not discard the fourth principal component for analysis. The models specifying seven
and eight distinct morphological groups of ground-finches received the strongest support (ΔBIC ≤ 1.26).
Support for all “dozens of cluster species,” or “1 or 6 or 30 species” (page 695) was negligible (ΔBIC in all cases >20; Fig. 4). In turn, the model with the
lowest support represented the Sisyphian evolution hypothesis proposing no distinct morphological groups
of ground-finches (i.e., that there is a single group; McKay and Zink 2015), which had a 500 BIC difference
to the second-worse model and >821 BIC difference to the two best models. Support for hypotheses consistent with
the alternative scenarios that there might be “dozens of cluster species” or 30 species (McKay and Zink 2015;
page 695) was also poor. Relative to the best models, models specifying groupings consistent with taxonomy
recognizing six or nine species were weakly supported (Fig. 4), considering differences in BIC scores greater
than six are typically regarded as strong or very strong evidence against models with lower support (Kass and
Raftery 1995). In sum, the data provided poor empirical support for the hypothesis that ground-finches consist
of only one species (McKay and Zink 2015) and strongly
supported hypotheses of several (but not dozens or 30; McKay and Zink 2015) morphologically distinct groups
(Fig. 5, Supplementary Material Fig. 1 available on Dryad). However, those groups did not exactly align
with existing taxonomic treatments of Geospiza (Fig. 6, Supplementary Material Fig. 2 available on Dryad).

Despite their comparatively low empirical support, models specifying six or nine morphological groups
according to taxonomic treatments of Geospiza (Lack 1947; Remsen Jr. et al. 2017) were partially consistent
with the best models (Fig. 6 and Supplementary Material Fig. 2 available on Dryad). For example, in the best models, all specimens of two of the nine currently

( McKay and Zink 2015). Toward the opposite end, NMMs assuming up to 30 distinct morphological
groups represented hypotheses alluded to by McKay and
Zink (2015) when they suggested ground-finches may consist of “dozens of cluster species,” or “1 or 6 or 30 species” (page 695). We also fitted NMMs specifying the six
(Lack 1947) or nine (Lamichhaney et al. 2015; Remsen Jr. et al. 2017) species recognized by alternative
taxonomic treatments of Geospiza, using the original specimen identifications in Swarth’s data updated to
reflect changes in nomenclature. We used the Bayesian
Information Criterion (BIC; Schwarz 1978) to measure empirical support for different NMMs (Fraley and
Raftery 2002) and thereby explicitly evaluated the
hypothesis that there is only one species of ground-
finch (McKay and Zink 2015) relative to hypotheses
that there are several species in the group (Lack 1947;
Lamichhaney et al. 2015; Remsen Jr. et al. 2017).

Recognized species (G. scandens, G. septentrionalis) were assigned to two respective morphological groups
which included few or no specimens of other species (Fig. 6; Supplementary Material Fig. 2 available on Dryad). Discrepancies between our analysis and current
taxonomy were most evident in cases such as those
recognizing six species (Lack 1947; Rising et al. 2011) or nine species (Remsen Jr. et al. 2017). Empirical
support was measured as difference in BIC relative to the best model (ΔABC). The two models with highest-empirical support assumed seven
eight distinct morphological groups. Empirical support for the model corresponding to the Sisyphian evolution hypothesis positing
recognizing six species (Lack 1947; McKay and Zink 2015) was
negligible (ΔABC >820).

FIGURE 4. Analysis of morphological data strongly supported hypotheses that there are multiple distinct groups of Geospiza
ground-finches. The plot shows the empirical support (ordinate) for normal mixture models assuming 1–30 distinct morphological
groups (abscissa), and for the two models specifying groupings of specimens reflecting taxonomic treatments recognizing six species (Lack 1947;
Rising et al. 2011) or nine species (Remsen Jr. et al. 2017). Empirical support was measured as difference in BIC relative to the best model (ΔABC). The two models with highest-empirical support assumed seven
and eight distinct morphological groups. Empirical support for the model corresponding to the Sisyphian evolution hypothesis positing there is a single species of ground-finch (i.e., a single morphological group, McKay and Zink 2015) was negligible (ΔABC >820).
species delimitation was not based on morphology, but rather resulted from recent genomic analyses revealing that phenotypically similar populations are distantly related (Lamichhaney et al. 2015). This likely explains why our analysis did not fully discriminate some species pairs in the morphological space we examined (G. conirostris vs. G. propinqua and G. difficilis vs. G. septentrionalis), although they may be more distinct in other phenotypic spaces including bill profile and song as well as behavior (Grant et al. 2000; Grant and Grant 2002). Also, we assumed that specimen identifications in the data set we analyzed were faultless; thus, part of the apparent mismatch between morphological groups detected in our analyses and taxonomy may reflect identification errors. Evaluating this possibility would require detailed examinations of individual specimens beyond the scope of our work.

Geographic context is an important consideration in assessments of species limits using phenotypic traits. Under a wide range of species definitions (sensu de Queiroz 1996), distinct phenotypic groups among sympatric individuals are readily accepted as evidence for the existence of distinct species (Mayr 1992; Mallet 2008). However, distinct phenotypic groups corresponding to nonsympatric populations may be less readily accepted as evidence of distinct species under species definitions that emphasize “intrinsic” over “extrinsic” barriers to gene exchange (Harrison 1998) because such groups may reflect ephemeral within-species differentiation due to geographic isolation or local adaptation (Zapata and Jiménez 2012). The morphological groups of ground-fiches we detected (Fig. 5, Supplementary Material Fig. 1 available on Dryad) cannot be interpreted to reflect within-species,
Current taxonomy (Remsen Jr. et al. 2017)

G. acutirostris

G. fortis

G. propinqua

G. scandens

G. difficilis

G. magnirostris

G. septentrionalis

Previous taxonomy (Lack 1947; Rising et al. 2011)

G. conirostris

G. fuliginosa

G. scandens

G. difficilis

G. magnirostris

G. septentrionalis

Morphological group

FIGURE 6. Eight morphological groups of Geospiza ground-finches in the Galapagos Archipelago identified by one of the best normal mixture models partially correspond to the nine species recognized by current taxonomy (Remsen Jr. et al. 2017) and to the six species recognized by previous taxonomy (Lack 1947; Rising et al. 2011). Each histogram shows, for each recognized species, the number of specimens assigned to each of the eight morphological groups. Groups are colored according to the scheme in Fig. 5.

TABLE 1. Number of islands in the Galapagos Archipelago where each of the eight morphological groups of Geospiza ground-finches identified by one of the best normal mixture models were found to occur (diagonal) and co-occur with other groups (off diagonal).

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
<th>Group 5</th>
<th>Group 6</th>
<th>Group 7</th>
<th>Group 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. acutirostris</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>G. conirostris</td>
<td>---</td>
<td>9</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>G. propinqua</td>
<td>---</td>
<td>---</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G. scandens</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>11</td>
<td>2</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>G. difficile</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>G. magnirostris</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>G. septentrionalis</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>14</td>
</tr>
<tr>
<td>G. fortis</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

Note: All groups co-occurred with each other in at least one island, except for cases involving Group 3, which did not co-occur with three other groups. Note, however, that Group 3 was not recovered as distinct in the other best model, which identified only seven groups (Supplementary Table 1 available on Dryad).
Thus, we refrain from additional evidence (e.g., genomics; Lamichhaney et al. 2015) of historical morphological data with contemporary discussions about species limits involving comparisons of all morphological groups in multiple islands. Nonetheless, our analyses serve to demonstrate that if such morphs were treated as species, then one would need to recognize dozens of species in the group; our analyses suggest this is not the case given the occurrence of distinct morphological groups in potential sympatry.

This is because over the past century, ground-finches exist within islands and argued that the burden of proof for systematists proposing to lump ground-finches into a single species based on morphological data is on showing that distinct groups do not exist.

**CONCLUSIONS; OR, ATLANTEAN EVOLUTION IN DARWIN’S FINCHES**

Our reanalysis of morphological data pointed strongly to the existence of several groups of phenotypically distinct *Geospiza* ground-finches based on six linear morphological measurements. In addition, we found evidence of distinct phenotypes in geographic scenarios (i.e., sympathy within islands) where one should not expect them if populations had not achieved evolutionary independence. Specifically, because the variation in quantitative morphological traits we examined is polygenic (Grant and Grant 1999; Grant and Grant 1994; Lamichhaney et al. 2015) and not caused by differences in sex or age (we restricted analyses to adult males), the existence of distinct phenotypic groups in areas where populations come into contact implies there are likely several species of ground-finches. Therefore, we contend that ground-finches are indeed present and not an example of Sisyphian evolution (McKay and Zink 2015), a term that could well apply to other systems in nature (Seehausen 2006; Nosil et al. 2009; Rudman and Schluter 2016). Instead, evolutionary forces maintaining populations of ground-finches are likely in place, just as in Greek mythology Atlas prevents the merging of the Earth and the sky with his shoulders. Ground-finches thus likely represent an example of what one might call “Atlantean evolution.” One, of course, does not need a new term to refer to speciation, but thinking of Atlas brings to mind atlas, a collection of maps, which reminds one of the central role of geography in speciation (Price 2008) and in the basic model underlying species delimitation based on phenotypic variation.

The question of exactly how many species of Darwin’s ground-finches are there remains open and requires further attention to morphology, including careful scrutiny of discrepancies between morphological variation and taxonomy (e.g., Fig. 6). In addition, morphological variation should be further examined in light of biological factors including additional phenotypic characters, ecological niches, mating behavior, population dynamics, and patterns of genetic and genomic variation among populations (Grant 1999; Huber et al. 2007; Grant and Grant 2008; Farrington et al. 2014; Grant and Grant 2014; Lamichhaney et al. 2015; McKay and Zink 2015). Fruitful discussions about species limits in the group would likely start by addressing some of the additional thought-provoking arguments advanced by McKay and Zink (2015) that we did not touch on and which are beyond the scope of our work (e.g., the extent to which morphological differentiation due to geographic isolation or local adaptation.
Table 2. Summary of three frequent issues in analysis of phenotypic data for species delimitation, possible solutions afforded by normal mixture models (NMMs), and associated challenges

<table>
<thead>
<tr>
<th>Issues</th>
<th>Potential pitfalls</th>
<th>Proposed solutions</th>
<th>Some challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Graphical analyses may convey little information on phenotype frequencies</td>
<td>Failure to detect distinct phenotypic groups (Fig. 1)</td>
<td>Use NMMs to model phenotypic variation according to evolutionary theory</td>
<td>Improve approaches to estimate the number of distinct normal distributions (or “components”) in NMMs</td>
</tr>
<tr>
<td></td>
<td>Infer the existence of several phenotypic groups when, in fact, only one exists (Supplementary Material Appendices 1 available on Dryad)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assessment can be highly subjective and difficult to replicate, particularly in analyses including ≥ 2 phenotypic traits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Reduction of dimensionality via PCA (or related approaches) may exclude important characters</td>
<td>Failure to detect distinct phenotypic groups by considering only principal components that explain a high proportion of the overall phenotypic variance (Fig. 2)</td>
<td>Use variable-selection approaches to reduce the number of dimensions (i.e., traits or principal components) to those that best discriminate groups in NMMs, using no a priori group information</td>
<td>Improve variable selection techniques for NMMs</td>
</tr>
<tr>
<td></td>
<td>Use NMMs to determine the number of distinct normal distributions (i.e., phenotypic groups) and estimate the parameters describing such distributions (means and variance-covariance matrices)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Differences in central tendency are not evidence of distinct phenotypic distributions</td>
<td>Misinterpret differences in mean phenotypes as evidence of distinct groups (Fig. 3, Supplementary Material Appendix 1 available on Dryad)</td>
<td>Use NMMs to determine the number of distinct normal distributions (i.e., phenotypic groups) and estimate the parameters describing such distributions (means and variance-covariance matrices)</td>
<td></td>
</tr>
</tbody>
</table>

Outlook

We have described three issues that commonly affect species delimitation using phenotypic data, potential pitfalls that may derive from not considering such issues, and some proposed solutions (Table 2). However, we also highlight that the approaches used here to analyze phenotypic data for species delimitation are not free of problems. Issues such as estimation of the number of groups in NMMs (McLachlan and Peel 2000; McLachlan and Rathnayake 2014) or how to select variables for NMM analyses of multidimensional data sets (Poon et al. 2013) are critical areas of active research in statistics in which progress remains to be made (Table 2). Despite these issues, however, we argue that the statistical tools we used are appropriate because they are directly related to the basic evolutionary model underlying species delimitation using phenotypic data (Fisher 1918). Moreover, these tools allow systematists to go beyond fairly limited graphical analysis, and to break free from problems resulting from reduction of dimensionality using PCA or related approaches and from comparisons of measurements of central tendency. The value of embracing approaches with a solid theoretical basis despite limitations in their implementation in systematics is clear considering other developments in the field in which theory predated robust methodologies that subsequently blossomed. Such developments include the use of statistical methods to study species limits among fossil populations (Newell 1956), the application of probabilistic models to infer phylogenetic trees (Felsenstein 1981), time-calibration of molecular phylogenies (Kishino and Hasegawa 1990), and the estimation of species trees from gene trees (Maddison 1997).

Practical approaches to fit NMMs without a priori information about species limits offer a fresh perspective in inferences available to systematists and bring into question conventional reliance on two criteria often employed to establish species limits: lack of overlap in
phenotypic ranges and gaps in phenotypic distributions. In the absence of NMMs, it seemed reasonable to argue that species limits should be based on fixed phenotypic differences because continuous variation could only be subdivided using subjective criteria (Cracraft 1989; Davis 1997). Accordingly, overlap of phenotypic ranges has been conventionally stressed as a criterion to suggest samples of individuals are conspecific (e.g., Simpson 1951; Davis and Heywood 1963; Zink 2002; McKay and Zink 2015). However, under the framework offered by NMMs, overlap in phenotypic ranges is not relevant for species delimitation for two reasons. First, one may find strong empirical support for models in which the phenotypic ranges of distinct normal distributions overlap (e.g., Figs. 1 and 5), indicating that range overlap does not imply absence of evidence of species limits. Second, because phenotypic variation conforming to a single normal distribution can be arbitrarily split into parts with nonoverlapping ranges (Supplementary Material Appendix 1 available on Dryad), absence of range overlap does not imply strong empirical support for models with more than a single species.

In addition to lack of overlap of phenotypic ranges, gaps in phenotypic distributions have been conventionally used as a species criterion (Mallet 2013). Such gaps may be defined as phenotypic regions with low frequency of individuals and therefore do not necessarily imply lack of overlap in phenotypic ranges; indeed, a gap may exist between two distinct phenotypic distributions that overlap in their extremes (e.g., Fig. 1). Although true phenotypic gaps (along with multimodality in phenotypic distributions) are sufficient to suggest species boundaries (Zapata and Jiménez 2012; Mallet 2013; but see Day 1969; McLachlan and Peel 2000, page 17), they are not necessary to demonstrate such boundaries exist because NMMs specifying more than one species may be strongly supported in the absence of phenotypic gaps. An example of support for more than one normal distribution in the absence of phenotypic gaps was provided at the inception of NMMs: Karl Pearson inferred two groups among specimens of the shore crab (Carcinus maenas) from the Bay of Naples, even though the mixture of the groups was not bimodal and therefore they were not separated by a gap (Pearson 1894). Moreover, Pearson examined the possibility of inferring the existence of groups with different phenotypic variances but identical phenotypic means, which are by definition not separated by a gap (Supplementary Material Appendix 2 available on Dryad).

To conclude, we note that the criteria for species delimitation discussed above are relevant in the context of ideas about the reality of species. In particular, it has been argued that if the hypothesis that species are real entities in nature is correct, then biological diversity should be a patchwork of phenotypic clusters delineated by gaps (Coyne and Orr 2004; Barrachough and Humphreys 2015). This prediction, however, would not necessarily follow from the hypothesis that species are real if, as we argue, phenotypically distinct species need not be separated by gaps. In other words, species may be real, phenotypically distinct entities in nature even if phenotypic gaps are not major elements structuring biological diversity. Because statistical approaches now allow systematists to make unprecedented formal inferences about the existence of species even in the absence of phenotypic gaps, they constitute particularly useful tools to describe the structure of biological diversity, a necessary step to understand the evolutionary processes that generated it.

**Supplementary Material**

Data available from the Dryad Digital Repository: [http://dx.doi.org/10.5061/dryad.9gh90](http://dx.doi.org/10.5061/dryad.9gh90)

**Acknowledgements**

We thank the California Academy of Sciences (Jack Dumbacher) for allowing us to use and publish morphological data from H. S. Swarth’s archive. Bailey McKay kindly provided tabulated data. We thank Peter Grant, James Mallet, Bailey McKay, Van Remsen Jr, Sara Ruane, Elizabeth Spriggs, Peter Stevens, members of C.D. Cadena’s laboratory group, anonymous reviewers, and associate editor John McCormack for discussion and helpful comments on the manuscript. C.D. Cadena thanks the Universidad de los Andes for allowing a sabbatical leave that enabled him to complete this project.

**References**


Hausdorf B., Hennig C. 2010. Species delimitation using dominant and
Grant P. R., Grant B.R. 2014. 40 years of evolution: Darwin’s Finches on
Grant P. R., Grant B.R. 2008. How and why species multiply: the
Grant B.R., Grant P. R. 2002. Simulating secondary contact in allopatric
Grant B.R., Grant P. R. 1989. Evolutionary dynamics of a natural
Fraley C., Raftery A.E., Murphy T.B., Scrucca L. 2012. mclust Version
Fisher R.A. 1918. The correlation between relatives on the supposition
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a
de Queiroz K. 1998. The general lineage concept of species, species
de León L.F., Raeymaekers J.A.M., Bermingham E., Podos J., Herrel
Davis J.I. 1997. Evolution, evidence and the role of species concepts in
de Queiroz K. 1998. The general lineage concept of species, species
Edwards D.L., Knowles L.L. 2014. Species detection and individual
assignment in species delimitation: can integrative data increase
Department of Statistics, University of Washington.
Coalescent-based species delimitation in an integrative taxonomy.
Grant B.R., Grant P.R. 1989. Evolutionary dynamics of a natural
Grant B.R., Grant P.R. 2003. Simulating secondary contact in allopatric
Grant P.R., Grant B.R. 1994. Phenotypic and genetic effects of
Grant P.R., Grant B.R. 2008. How and why species multiply: the
Grant P.R., Grant B.R. 2014. 40 years of evolution: Darwin’s Finches on
Grant P.R., Grant B.R., Petren K. 2000. The allopatric phase of
speciation: the sharp-beaked ground finch (Geospiza difficilis) on the
Henderson A. 2006. Traditional morphtopometrics in plant systematics
Luckow M. 1995. Species concepts: assumptions, methods, and
Mallet J. 2008. Hybridization, ecological races and the nature of species:
Mayr E., Linley E.G., Usinger R.L. 1953. Methods and principles of
Science 350:993–996.
Miller W. 2016. The species problem: concepts, conflicts, and


