Syst. Biol. 64(6):900–908, 2015

DOI:10.1093/sysbio/syu130
Advance Access publication January 19, 2015

Delimiting Species-Poor Data Sets using Single Molecular Markers: A Study of Barcode Gaps, Haplowebs and GMYC

SIMON DELLICOUR[1] AND JEAN-FRANÇOIS FLOT[2,*]

[1]*Department of Zoology, University of Oxford, Oxford OX1 3PS, UK and* [2]*Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK*

[*]*Correspondence to be sent to: Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK; E-mail: j.flot@ucl.ac.uk.*

Received 27 November 2014; reviews returned 9 December 2014; accepted 18 December 2014
Associate Editor: Bryan Carstens

Abstract.—Most single-locus molecular approaches to species delimitation available to date have been designed and tested on data sets comprising at least tens of species, whereas the opposite case (species-poor data sets for which the hypothesis that all individuals are conspecific cannot by rejected beforehand) has rarely been the focus of such attempts. Here we compare the performance of barcode gap detection, haplowebs and generalized mixed Yule–coalescent (GMYC) models to delineate chimpanzees and bonobos using nuclear sequence markers, then apply these single-locus species delimitation methods to data sets of one, three, or six species simulated under a wide range of population sizes, speciation rates, mutation rates and sampling efforts. Our results show that barcode gap detection and GMYC models are unable to delineate species properly in data sets composed of one or two species, two situations in which haplowebs outperform them. For data sets composed of three or six species, bGMYC and haplowebs outperform the single-threshold and multiple-threshold versions of GMYC, whereas a clear barcode gap is only observed when population sizes and speciation rates are both small. The latter conditions represent a "sweet spot" for molecular taxonomy where all the single-locus approaches tested work well; however, the performance of these methods decreases strongly when population sizes and speciation rates are high, suggesting that multilocus approaches may be necessary to tackle such cases. [Barcode gap; generalized mixed Yule–coalescent models; haplowebs; heterozygosity; molecular markers; species delimitation; systematics; taxonomy.]

Molecular approaches to species delimitation have been undergoing rapid development for the last 15 years, yielding a wide diversity of methods. Faced with this plurality of views, some researchers advocate using several approaches (Carstens et al. 2013) and taking into account non-molecular lines of evidence such as morphology, behavior or ecology (Dayrat 2005). This holistic approach that places human expertise at the center of the species delimitation process contrasts with the high degree of optimism initially vested in some simplistic automatic methods, such as using an *a priori* sequence divergence threshold to delineate and identify species ("DNA barcoding"; Hebert et al. 2003a; 2003b; 2004). The quest for an efficient and accurate automatic approach continues, however, as the availability of taxonomic experts appears to be the most significant bottleneck to the description of the Earth's biodiversity.

The most popular single-locus methods currently used to delineate species are barcode gap detection (Lefébure et al. 2006; Puillandre et al. 2012) and the generalized mixed Yule–coalescent model (GMYC; Pons et al. 2006; Monaghan et al. 2009; reviewed in Fujisawa and Barraclough 2013). Although these methods are applicable to both haploid (e.g., mitochondrial) and diploid (e.g., nuclear) data sets, they are usually performed on cytochrome *c* oxidase (COI) mitochondrial sequence alignments as this marker presents the advantages of having a small effective population size (Moore 1995), being easily sequenced using universal primers (Folmer et al. 1994) and being very variable in most animal groups. A different approach to delineate species using diploid nuclear markers was proposed recently (Flot et al. 2010). In this approach, termed "haplowebs", the co-occurrence of haplotypes in heterozygous individuals is used to delineate gene pools that are reproductively isolated from one another, each of which corresponds to a putative species. Following an earlier proposal by Doyle (1995), himself inspired by Carson (1957), these putative species are called "fields for recombination". Because this approach requires a dense sampling comprising several individuals per species, it appears well suited to delineate species in species-poor data sets, and indeed most applications published so far dealt with data sets comprising 1 (Flot et al. 2013; Adjeroud et al. 2014), 2 (Flot et al. 2010), 3 (Flot et al. 2011), 5 (Li 2012) or 10 species (Miralles and Vences 2013). However, a rigorous test of the domain of validity of haplowebs and of their performance compared with barcode gap detection and GMYC is still wanting.

To meet this aim, we analyzed a test data set of chimpanzees and bonobos (Hey 2010) using all three types of approaches and conducted mono- and multispecies simulations using different population sizes, speciation rates, mutation rates and sampling efforts (i.e., the number of gene copies sequenced). In addition to the single-threshold (ST-GMYC; Pons et al. 2006) and multiple-threshold (MT-GMYC; Monaghan et al. 2009) variants of GMYC, we included in our comparison a recently proposed Bayesian approach using the same logic (bGMYC; Reid and Carstens 2012). The inclusion of bGMYC (that returns probabilities of conspecificity for each pair of sequences in the data set) motivated us to quantify the accuracy of each method by calculating the percentage of pairs of conspecific sequences mistaken as heterospecific (%oversplitting) and the percentage of pairs of heterospecific sequences

mistaken as conspecific (%overlumping), a significant shift from the methodologies adopted in previous simulation studies aimed at assessing the performance of GMYC (Papadopoulou et al. 2008; Lohse 2009; Esselstyn et al. 2012; Reid and Carstens 2012; Fujisawa and Barraclough 2013; Zhang et al. 2013; Tang et al. 2014).

Materials and Methods

Empirical Data Set

To start investigating the performance of barcode gap detection, haplowebs, ST-GMYC-ST, MT-GMYC-MT and bGMYC when dealing with species-poor data sets, we applied these methods to a data set of nuclear markers of bonobos and chimpanzees (Hey 2010). Among the 73 markers in this data set, we selected the 16 for which the same set of individuals (3 populations of 10 chimps each and 1 population of 9 bonobos) had been consistently sequenced. Four markers were subsequently excluded as they caused errors in the downstream GMYC or bGMYC analyses, resulting in a final set of 12 markers (Fischer regions 11, 13, 14, 15, 16, 21, 22, 25, 26, 28, 29 and 30; Fischer et al. 2006) ranging in size from 486 (region 26) to 803 base pairs (region 13).

Barcode gaps were studied by computing for each marker the mismatch distribution and Sarle's bimodality coefficient (Ellison 1987; Pfister et al. 2013). Sarle's coefficient is a measure of bimodality that varies between 0 (perfectly unimodal data) and 1 (perfectly bimodal data). Values between 5/9 and 1 indicate a bi- or multimodal distribution, whereas values below 5/9 indicate unimodality. Mismatch distributions and Sarle's coefficients were respectively computed using the R packages *pegas* (Paradis 2010) and *moments* (Komsta and Novomestky 2012). We also calculated separately the distributions of mismatches for intraspecific pairs and for interspecific pairs of sequences.

After removing duplicate sequences since they may cause problems with downstream GMYC analyses (Fujisawa and Barraclough 2013), the best-fitting substitution model for each marker was chosen with the help of jModelTest (Posada 2008) following the Akaike information criterion (Akaike 1974) corrected for small samples sizes (Hurvich and Tsai 1989). HKY (Hasegawa et al. 1985) was identified as the best-fitting model for all markers, except regions 13 and 15 for which the best model was GTR (Lanave et al. 1984; Tavaré 1986), regions 25 and 26 for which the best model was F81 (Felsenstein 1981), and region 30 for which the best model was K80 (Kimura 1980). Phylogenetic analyses were performed using BEAST 1.7.4 (Drummond and Rambaut 2007) following a model characterized by a constant population size and a strict clock (length of the Markov chain Monte Carlo (MCMC): 1,000,000,000 generations following a 100,000,000-generation burn-in; sampling every 1,000,000 generations). For each sampled population of 100 trees, a consensus was built with TreeAnnotator 1.7.4 (from the BEAST package; Drummond and Rambaut 2007) using the maximum clade credibility method and setting the posterior probability limit to 0; ST-GMYC and MT-GMYC analyses were conducted on this consensus tree using the R package *splits* (Ezard et al. 2013). bGMYC analyses were performed by running the eponymous R package (Reid and Carstens 2012) on the 100 trees sampled during the MCMC; this was done over 110,000 generations, discarding the first 10,000 as burn-in and sampling every 100 generations afterwards. In the case of bGMYC analyses, the convergence of the MCMC was assessed by checking the evolution graph of the posterior probability against the number of generations, as advised in the bGMYC tutorial.

Haplowebs were constructed as in Flot et al. (2010): briefly, median-joining networks (Bandelt et al. 1999) were drawn for each marker using the program Network (Fluxus Technologies) and exported as PDFs using Network Publisher. Connecting curves between the haplotypes found co-occurring in heterozygous individuals were subsequently added on each network using Inkscape (Bah 2011).

Simulated Data Sets

We used the program SIMCOAL (Excoffier et al. 2000) to simulate 10 replicate populations of 10 and 100 gene copies of a locus of 1000 base pairs, for three population sizes (10^4, 10^5 and 10^6 haploid individuals, corresponding to 5×10^3, 5×10^4, and 5×10^5 diploid individuals respectively) and three mutation rates (10^{-5}, 10^{-4} and 10^{-3} mutations per generation and per locus, hence 10^{-8}, 10^{-7} and 10^{-6} mutations per base per generation). Multispecies simulations were performed using the Python package DendroPy 3.12.0 (Sukumaran and Holder 2010): we first simulated trees of either three or six species using a Yule model with three different birth rates (0.1, 1 and 10 births per lineage per million generations), then used standard coalescent conditions with population sizes of 10^4, 10^5 and 10^6 to simulate 10 replicate genealogies of 10 and 100 gene copies within each species; each genealogy was finally turned into three alignments of 1000 base pairs using mutation rates of 10^{-5}, 10^{-4} and 10^{-3} mutations per generation per sequence. Following the approach of Esselstyn et al. (2012), we used directly the simulated gene genealogies as inputs for the GMYC approaches, hence removing one possible source of uncertainty (the inference of phylogenetic trees from simulated sequence data) and avoiding the most computationally expensive step in the process. As described in Esselstyn et al. (2012), the selected birth rate values correspond to the range of speciation rates reported for a variety of organisms (Baldwin and Sanderson 1998; Mendelson and Shaw 2005; Phillimore and Price 2008; Moyle et al. 2009; Rowe et al. 2011) and the effective sizes to a range of biologically relevant values (Burgess and Yang 2008; Russell et al. 2011). The lowest per-site mutation rate we considered (10^{-8} mutations per generation) was the rate previously reported for the human nuclear genome (Roach et al.

2010); we also considered mutation rates two orders of magnitude higher to account for the preferential use of highly variable markers (such as mitochondrial genes and nuclear introns) in species delimitation studies.

Barcode gap detection analyses were conducted as explained above. For the haploweb analyses, we formed diploid genotypes by picking up randomly pairs of conspecific sequences from the corresponding alignments (without replacement) 10 times for each simulation, yielding a total of 100 replicate populations per set of parameters. This procedure and the subsequent delineation of fields for recombination were performed using a custom perl script (countffrs.pl available in the Dryad data repository, http://dx.doi.org/10.5061/dryad.t7m5v). GMYC analyses were conducted by running the R package *splits* (Ezard et al. 2013) and bGMYC analyses were performed by running the eponymous R package on 1000 sampled trees. For each simulated genealogy, bGMYC was run over 110,000 generations, discarding the first 10,000 as burn-in and sampling every 100 generations afterward. The convergence of the MCMC was checked as described above. Because bGYMC returns a probability of conspecificity for each pair of sequences in the data set, its output is not directly comparable to ST-GMYC, MT-GMYC and haplowebs, which return groupings of individuals into species: to generate such groupings from the bGYMC output, we added a discretization step (retaining only pairs of individuals that had a probability of conspecificity higher than 0.95) followed by a transitivization step (during which individuals were aggregated into species: individuals X and Y were considered conspecific if X was conspecific with Z and Y conspecific with Z, even if the probability of conspecificity of X and Y was lower than the aforementioned 0.95 threshold). Both the raw (bGMYC^1) and discreticized–transitivized (bGMYC^2) results of bGMYC were used for calculating the error rates of this method (see below).

Taking advantage of the fact that the actual species delimitations were known for all the simulated data sets, we calculated the percentage of conspecific pairs of sequences returned as heterospecific (hereafter referred to as %oversplitting) and the percentage of heterospecific pairs of sequences returned as conspecific (%overlumping). In the case of bGMYC, we also computed the percentage of pairs of sequences that were neither confidently split nor lumped as the corresponding probability of conspecificity was between 0.05 and 0.95 (%indecision). Finally, we computed the percentage of pairs of sequences correctly determined (%success = 100–%oversplitting–%overlumping–%indecision). These different percentages were averaged over 30 replicate genealogies for SimCoal, 10 replicate genealogies for DendroPy or 100 replicate populations for the haploweb analyses.

Finally, in order to investigate the impact of mutation rate on the performance of GMYC methods, we performed GMYC analyses on phylogenetic trees inferred from simulated DNA sequence alignments. Because such analyses are computationally very expensive we limited ourselves to analyzing one set of conditions under which GMYC approaches yielded good results when applied directly on the gene genealogies (see the "Results" section): 100 gene copies per species, an effective population size of 10^4 haploid individuals (or 5×10^5 diploid ones), with 10^{-5}, 10^{-4} and 10^{-3} mutations per locus per generation and a birth rate of 10^{-6} for multispecies simulations. Phylogenetic analyses were performed in BEAST 1.7.4 (Drummond and Rambaut 2007) after removal of duplicate sequences (Fujisawa and Barraclough 2013) using a constant-size population model, a HKY substitution model (Hasegawa et al. 1985) and a strict clock model (length of the MCMC: 1,000,000,000 generations following a 100,000,000-generations burn-in, sampling every 1,000,000 generations). For each sampled tree population, a consensus tree was built with TreeAnnotator 1.7.4 (from the BEAST package; Drummond and Rambaut 2007) using the maximum clade credibility method and setting the posterior probability limit to 0. GMYC analyses were conducted on the consensus trees, and bGMYC analyses were performed on the 100 trees sampled during the MCMC. As above, bGMYC was run over 110,000 generations, discarding the first 10,000 as burn-in and sampling every 100 generations afterward.

Results

Empirical Test using Bonobos and Chimpanzees

Chimpanzees (*Pan troglodytes*; Blumenbach 1776) and bonobos (*Pan paniscus*; Schwartz 1929) have been considered as distinct species since the discovery of the latter (Schwartz 1929; Coolidge 1933) and are assumed to have diverged nearly 1 million years ago (Fischer et al. 2004; Won and Hey 2005). As the distinction between them is well supported both morphologically and genetically (Hey 2010), it is a good empirical model to test the performance of single-locus species delimitation approaches on a data set comprised of two closely related species.

The results are summarized in Figure 1. The mismatch distributions of five markers were unimodal (regions 11, 13, 14, 15 and 16) whereas the others were bimodal; however, among the seven markers that gave bimodal mismatch distributions, three (regions 22, 29 and 30) had bimodal distributions for intraspecific distances as well, with modes indistinguishable from the distributions of interspecific distances. The remaining four markers (Fischer regions 21, 25, 26 and 28) had unimodal distributions for intraspecific and interspecific distances; for these markers the mode of the distribution of intraspecific distances was smaller than the mode for interspecific distances, but the two distributions overlapped. Haploweb analyses properly delineated chimpanzees and bonobos for 6 markers out of 12 (Fisher regions 15, 21, 22, 25, 26 and 28), whereas the

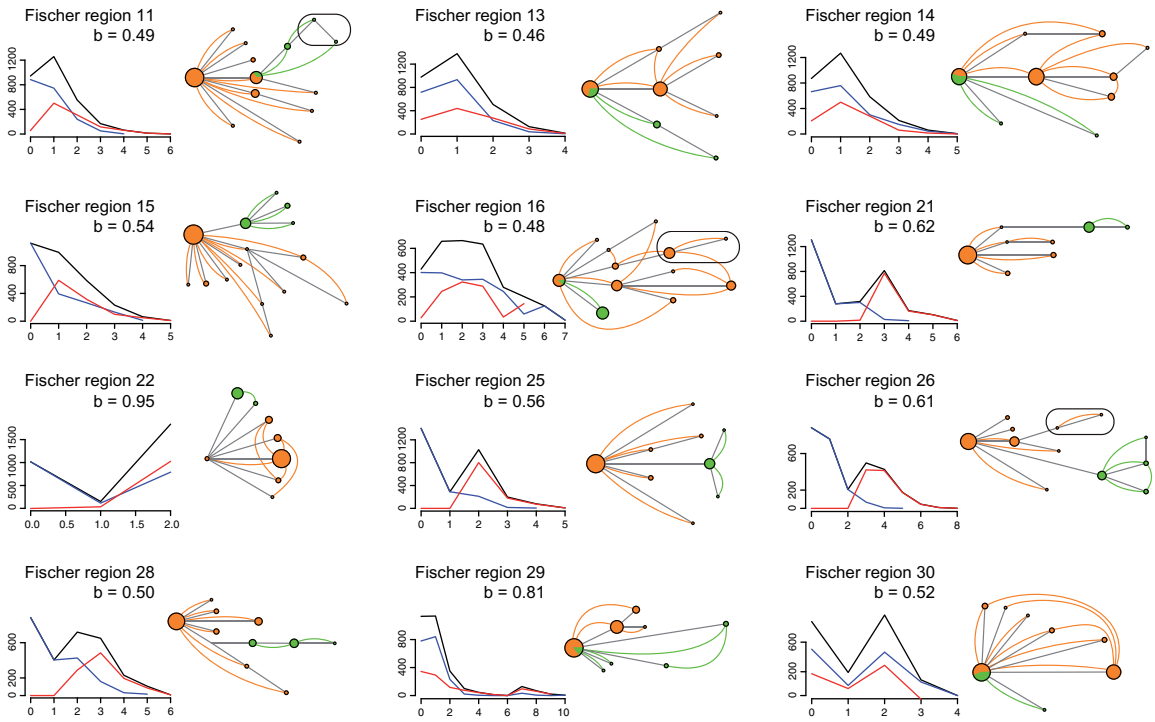

FIGURE 1. Comparison of the performance of barcode gap detection, haplowebs and GMYC-based approaches to delineate species of bonobos and chimpanzees using nuclear markers (Hey 2010). For each nuclear marker, the mismatch distribution and Sarle's bimodality coefficient are shown on the left side (in black), together with the distributions of intraspecific distances (in blue) and of interspecific distances (in red); the corresponding haploweb is shown on the right side (chimpanzee alleles are in orange whereas bonobo alleles are in green). ST-GMYC and MT-GMYC did not detect any significant split for any marker, whereas bGMYC considered some haplotypes as belonging to a distinct species (circled in black on the haplowebs of Fischer regions 11, 16 and 21).

6 other markers lumped them into a single putative species because of shared haplotypes. For each of the 12 markers, neither ST-GMYC nor MT-GMYC found the mixed Yule–coalescent model to be significantly more likely than the null hypothesis of pure coalescence, whereas bGMYC analyses of some markers did single out some individuals as probably belonging to a distinct species (see boxes on the haplowebs of regions 11, 16 and 26) but never recovered the correct boundary between bonobos and chimpanzees.

Barcode Gap Detection on Simulated Data Sets

We further analyzed the shape of the mismatch distribution over a large range of parameters using simulations. Boxplots of the resulting distributions of Sarle's bimodality coefficients are included in Figure 2 for the global distributions and in Figure 3 for intraspecific versus intraspecific distributions, whereas the global mismatch distributions are presented in Supplementary Figure S1 (available from http://dx.doi.org/10.5061/dryad.t7m5v) and the interspecific and intraspecific distributions of distances are plotted separately in Supplementary Figures S2 and S3 (respectively for three and six species).

Analyses of the distribution of Sarle's coefficient across replicate simulations revealed that intraspecific distributions obtained using a pure coalescent model (i.e., a single species) were at least as bimodal as the global distributions obtained when simulating three or six distinct species (Fig. 2). The lowest bimodality coefficients were obtained not for intraspecific simulations but for simulations of three or six species with large populations sizes (10^6 individuals) and/or low mutation rates (10^{-5} mutations per locus per generation). When plotted separately (Fig. 3), the distributions of Sarle's coefficients for intraspecific and interspecific mismatch distributions in simulations of three and six species were in nearly all cases above the 5/9 bimodality threshold; although the values were generally lower for intraspecific distributions than for interspecific ones, the lowest values were obtained for interspecific distributions with a low speciation rate (10^{-7}) combined with a high mutation rate (10^{-3} mutations per locus and per generation), in which case Sarle's coefficient was well under the 5/9 threshold.

Examination of the actual distributions of intraspecific and interspecific mismatches across the three-species simulation replicates (Supplementary Figure S2) revealed that a barcode gap was consistently present when speciation rate was low (10^{-7}) and population size was small (10^4 haploid individuals). All other cases displayed some overlap between intraspecific and interspecific distances; for combinations of high speciation rates (10^{-6} and 10^{-5}) and large population sizes (10^5 and 10^6) the distributions of intraspecific

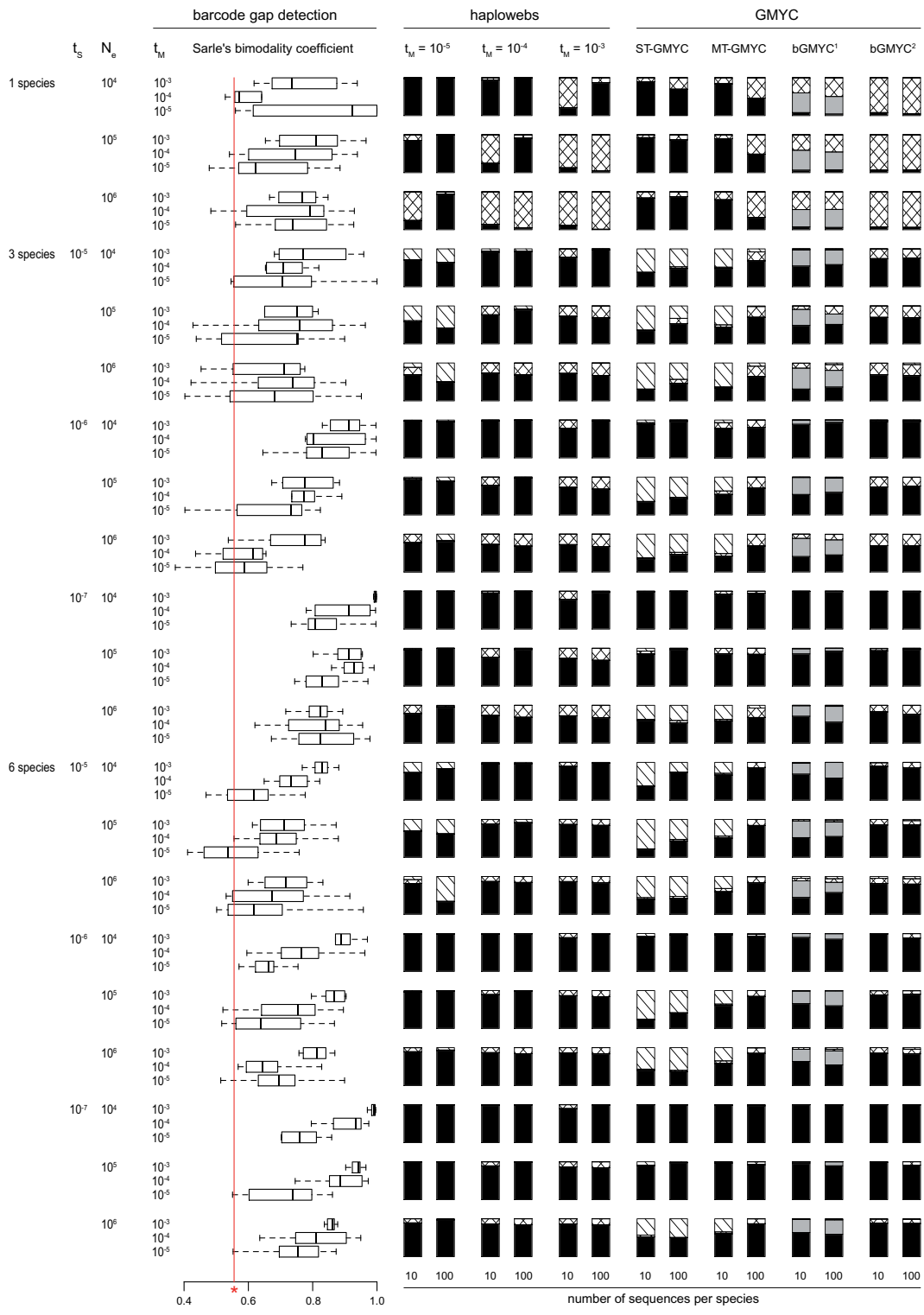

FIGURE 2. Species delimitation results based on simulated data sets. t_S refers to the speciation rate, N_e to the effective (haploid) population size and t_M to the mutation rate (number of mutations per locus per generation). The boxplots on the left summarize the distributions of Sarle's bimodality coefficient for the mismatch distributions: the limits of each box represent the 25th and 75th percentiles, the median is displayed as a line dividing the box, and whiskers extend from the lowest value within 1.5 interquartile range distance of the lower quartile to the highest value within 1.5 interquartile range distance of the upper quartile. The vertical line with a star indicates the threshold that separates unimodal distributions (Sarle's coefficient $< 5/9$) from bi- or multimodal distributions (Sarle's coefficient $\geq 5/9$). Haploweb and GMYC results are reported in barplots with, from top to bottom, overlumping (hatched), oversplitting (crosshatched), indecision (in gray), and success (in black). For bGMYC we report both the percentages based on the probability matrices (bGMYC^1) and those obtained after a discretization and transitivization step aimed at turning the probability matrices into species groupings (bGMYC^2, see the text for further details).

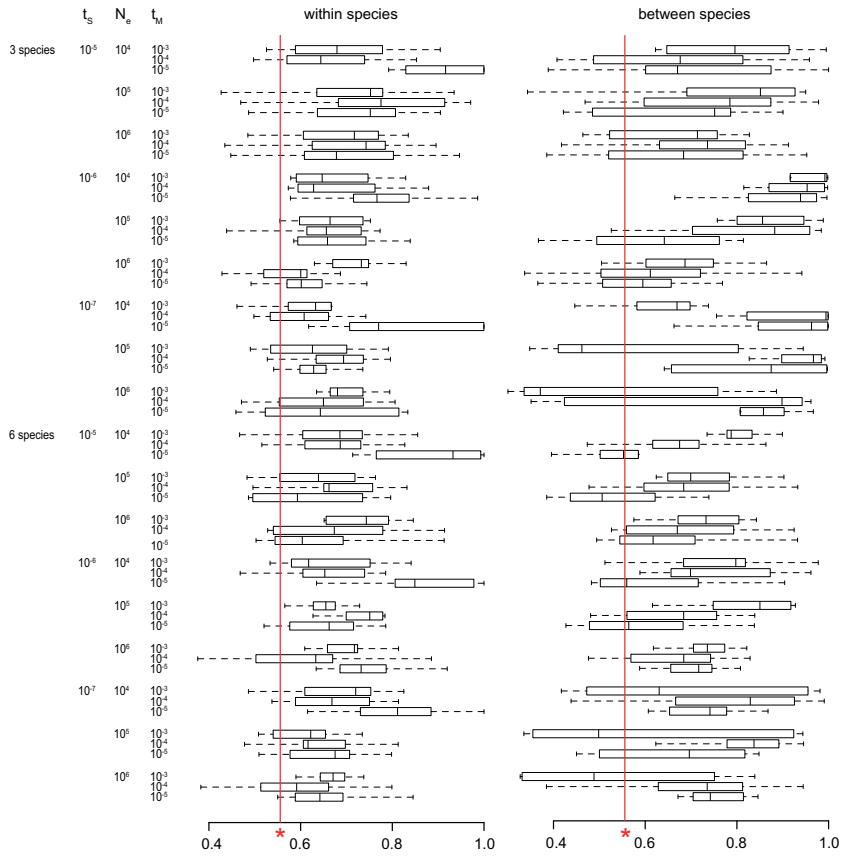

FIGURE 3. Boxplots summarizing the distributions of Sarle's bimodality coefficients computed for (left) the distributions of intraspecific mismatches (mismatches between sequences simulated for the same species) and (right) the distributions of interspecific mismatches (mismatches between sequences simulated for distinct species). t_S refers to the speciation rate, N_e to the effective (haploid) population size and t_M to the mutation rate (number of mutations per locus per generation). The limits of each box represent the 25th and 75th percentiles, the median is displayed as a line dividing the box, and whiskers extend from the lowest value within 1.5 interquartile range distance of the lower quartile to the highest value within 1.5 interquartile range distance of the upper quartile. The vertical line with a star indicates the threshold that separates unimodal distributions (Sarle's coefficient $< 5/9$) from bi- or multimodal distributions (Sarle's coefficient $\geq 5/9$).

and interspecific distances were indistinguishable. Similar results were obtained for six-species simulations (Supplementary Figure S3), except that a clear barcode gap was only observed for one particular condition of parameters: low speciation rate (10^{-7}), small population size (10^4 haploid individuals, i.e. 5×10^3 diploid individuals) and high mutation rate (10^{-3} mutations per locus and per generation).

Species Delimitation using Haplowebs on Simulated Data Sets

Haplowebs generally performed best with markers that were not very variable (10^{-5} or 10^{-4} mutations per locus and per generation), in species with small population sizes (10^4 or 10^5 gene copies, i.e. composed of 5×10^5 or 5×10^6 diploid individuals) and, in the case of three-species and six-species simulations, for speciation rates that were not too high (10^{-7} and 10^{-6}; Fig. 2). The most frequent type of error observed with haplowebs was oversplitting, particularly when dealing with highly variable markers (10^{-3} mutations per locus and per generation). When mutation rates were lower some overlumping was also observed, especially for data sets comprising a large number of gene copies.

Increasing sampling effort by sequencing 50 individuals per species instead of only 5 led to a decrease of oversplitting in most cases when oversplitting was a problem (as more heterozygote individuals were sampled), but also to an increase of overlumping under conditions when some alleles were shared between species (as the probability of picking up rare instances of allele sharing increased with sampling effort). This was particularly apparent when dealing with one-species data sets, in which case success rates ranged from 100% (for a low variability marker in a species with a small population size) to 0% (for a highly variable marker in a species with a large population size). In the latter case, only the two haplotypes of each heterozygote were connected in the resulting haplowebs, leading to the delimitation of each individual as a separate species.

Species Delimitation using GMYC on Simulated Data Sets

The error pattern obtained using GMYC was strikingly different depending on the method used (Fig. 2). For three-species and six-species simulations, the main error observed was overlumping (for ST-GMYC with both 10 and 100 genes copies sampled per species and for MT-GMYC with 10 gene copies sampled per species), whereas oversplitting was dominant for MT-GMYC when 100 gene copies were sampled per species and for bGMYC. For one-species simulations, oversplitting was observed using all methods in amounts ranging from moderate for ST-GMYC to very large for bGMYC. ST-GMYC was less error-prone than MT-GMYC and bGMYC for one-species data sets, but bGMYC outperformed the other two for three-species and six-species simulations (especially when the discretization and transitivization step was applied, which removed the otherwise very high uncertainty without increasing much the error rate). GMYC approaches performed best for low speciation rates (10^{-7}), small population sizes (10^4 or 10^5 haploid individuals) and markers with high mutation rates (Supplementary Figure S4).

Discussion

The empirical and simulated data analyzed here suggest that species-poor data sets pose specific challenges to species delimitation. Simulations of one-species data sets showed frequent oversplitting using GMYC as well as barcode gap detection and to a lesser extent haplowebs, resulting potentially in artifactual detection of non-existent species boundaries. In the case of GMYC this is due to a known limitation of this approach: ST-GMYC and MT-GMYC require several internode intervals to fit the "between-species" branching rates (Barraclough T., personal communication), whereas bGMYC does not test whether the mixed Yule-coalescent model fits the data significantly better than a pure coalescence (i.e., intraspecific) one (Reid and Carstens 2012; Reid N., personal communication). The observation that one-species data sets give rise to multimodal mismatch distributions that may be mistaken for evidence of distinct species was less expected. In any case, it appears that neither GMYC nor barcode gap detection approaches to species delimitation are suited when the hypothesis that all specimens sampled are conspecific cannot be excluded *a priori*; haplowebs, in contrast, may be used to test the hypothesis that the data set at hand comprises only one species, provided that the marker used is not too variable and that the sampling effort is sufficient (as was the case in Adjeroud et al. 2014).

Our analysis of empirical data from 12 independent nuclear markers in chimpanzees and bonobos showed that GMYC-based approaches were unable to detect the split between these two species for any of the markers analyzed. Barcode gap detection performed only slightly better on this data set: only 4 markers out of 12 exhibited a barcode gap separating intraspecific and interspecific distances; besides, in these four cases there was a large overlap between intraspecific and interspecific distributions. Haplowebs had the highest success rate of all approaches in this example, with 6 markers out of 12 delimiting properly bonobos from chimpanzees and the 6 other markers lumping them into a single species because of shared alleles.

When dealing with three-species and six-species simulated data sets, bGMYC and haplowebs outperformed ST-GMYC and MT-GMYC in nearly all cases. Our results for the GMYC-based approaches are consistent with those of previous theoretical and simulation analyses focusing on these methods. As Esselstyn et al. (2012), Reid and Carstens (2012) and Zhang et al. (2013), we found that the accuracy of GMYC-based approaches decreases sharply with increasing speciation rate, that is, when clades are undergoing rapid radiation. In their analysis of the ST-GMYC and MT-GMYC methods, Esselstyn et al. (2012) also pointed at the joint influence of speciation rate and population size on the accuracy of these methods. Similarly, Fujisawa and Barraclough (2013) argued that the main factor affecting GMYC accuracy was population size relative to the divergence time among species.

Not unexpectedly, our results show that barcode gap detection and haplowebs face challenges very similar to GMYC: they work best in the "sweet spot" where population sizes and speciation rates are low (and mutation rates are high), and perform poorly when population sizes and speciation rates are large (and mutation rates are low). Hence, one may extrapolate that no single-locus approach to species delimitation is likely to ever be capable of delineating properly species in cases when population sizes and speciation rates are both large. Choosing a marker characterized by a small effective population size and a high mutation rate may alleviate this problem to a certain extent: this is notably the case of mitochondrial markers such as COI (as they are haploid and maternally inherited; Moore 1995), but it applies also to nuclear ribosomal markers such as ITS1 and ITS2 thanks to their concerted mode of evolution (Navajas and Boursot 2003). Since COI is only present in one copy per individual, it can be analyzed using barcode gap detection and GMYC-based approaches but not using haplowebs; whereas ITS sequences of diploid individuals can be analyzed using all three types of methods thanks to their heterozygosity. For this reason, and because COI data sets are often plagued with nuclear pseudogenes (Bensasson et al. 2001; Song et al. 2008; Buhay 2009), ITS may be a better choice than COI for delineating species, especially now that methods are available to phase nuclear markers by directly sequencing PCR products without the having to go through the costly and time-consuming procedure of cloning them (Flot et al. 2006; Flot 2007, 2010; Dmitriev and Rakitov 2008; Harrigan et al. 2008).

In contrast to single-locus approaches, multilocus methods such as BPP (Yang and Rannala 2010; Rannala and Yang 2013) and SpedeSTEM (Ence and Carstens 2011) may be able to deal with high speciation rates

and large population sizes by leveraging information provided by several independent markers; however, future simulation studies will be required in order to find out whether it is really the case and to compare the performance of the various approaches available. In addition to the parameters explored in the present study, other challenges to species delimitation such as extinction, introgression and singletons/rare species (Lim et al. 2012) can also be simulated and we are planning to address them in the future.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.t7m5v.

Funding

Computational resources were provided by the Belgian High Performance Computing Centre co-funded by ULB and VUB (HPC cluster "Hydra"). S.D. is a research fellow of the Wiener-Anspach Foundation. J.-F.F. is supported by the European Research Council [ERC-2012-AdG 322790].

Acknowledgments

The authors are grateful to Jacob Esselstyn for sharing his Python script, to Noah Reid and Diego Fontaneto for their useful comments and advice, and to Jody Hey for providing us the chimpanzee and bonobo data set used in this study. They thank also Fabian Stiewe and Lukas Geyrhofer for useful discussions, and Andy Anderson, Tim Barraclough, Bryan Carstens and one anonymous reviewer for their comments on a previous version of this article.

References

Adjeroud M., Guérécheau A., Vidal-Dupiol J., Flot J.-F., Arnaud-Haond S., Bonhomme F. 2014. Genetic diversity, clonality and connectivity in the scleractinian coral *Pocillopora damicornis*: a multi-scale analysis in an insular, fragmented reef system. Mar. Biol. 161:531–541.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control 19:716–723

Bah T. 2011. Inkscape. Guide to a vector drawing program. 4th ed. Prentice Hall.

Baldwin B.G., Sanderson M.J. 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae). Proc. Natl. Acad. Sci. U. S. A. 95:9402–9406.

Bandelt H.J., Forster P., Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol Evol. 16:37–48.

Bensasson D., Zhang D.-X., Hartl D.L., Hewitt G.M. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol. Evol. 16:314–321.

Blumenbach J.F. 1776. De generis humani varietate nativa liber. Vandenhoeck, Goettingae.

Buhay J.E. 2009. "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. J. Crust. Biol. 29:96–110.

Burgess R., Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol. Biol. Evol. 25:1979–1994.

Carson H. 1957. The species as a field for recombination. In: Mayr E., editor. The species problem. Washington (DC): American Association for the Advancement of Science p. 23–38.

Carstens B.C., Pelletier T.A., Reid N.M., Satler J.D. 2013. How to fail at species delimitation. Mol. Ecol. 22:4369–4383.

Coolidge H.J. 1933. Pan paniscus. Pigmy chimpanzee from south of the Congo river. Am. J. Phys. Anthropol. 18:1–59.

Dayrat B. 2005. Towards integrative taxonomy. Biol. J. Linn. Soc. 85: 407–415.

Dmitriev D.A., Rakitov R.A. 2008. Decoding of superimposed traces produced by direct sequencing of heterozygous indels. PLoS Comput. Biol. 4:e1000113.

Doyle J.J. 1995. The irrelevance of allele tree topologies for species delimitation, and a non-topological alternative. Syst. Bot. 20: 574–588.

Drummond A., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Ellison A.M. 1987. Effect of seed dimorphism on the density-dependent dynamics of experimental populations of *Atriplex triangularis* (Chenopodiaceae). Am. J. Bot. 74:1280–1288.

Ence D.D., Carstens B.C. 2011. SpedeSTEM: a rapid and accurate method for species delimitation. Mol. Ecol. Resour. 11:473–480.

Esselstyn J.A., Evans B.J., Sedlock J.L., Anwarali Khan F.A., Heaney L.R. 2012. Single-locus species delimitation: a test of the mixed Yule–coalescent model, with an empirical application to Philippine round-leaf bats. Proc. R. Soc. B Biol. Sci. 279:3678–3686.

Excoffier L., Novembre J., Schneider S. 2000. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J. Hered. 91:506–509.

Ezard T., Fujisawa T., Barraclough T. 2013. R package *splits*: SPecies' LImits by Threshold Statistics, version 1.0-18/r45. Available from: URL http://r-forge.r-project.org/projects/splits/, last accessed January 2015.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Fischer A., Pollack J., Thalmann O., Nickel B., Pääbo S. 2006. Demographic history and genetic differentiation in apes. Curr. Biol. 16:1133–1138.

Fischer A., Wiebe V., Pääbo S., Przeworski M. 2004. Evidence for a complex demographic history of chimpanzees. Mol. Biol. Evol. 21: 799–808.

Flot J.-F. 2007. Champuru 1.0: a computer software for unraveling mixtures of two DNA sequences of unequal lengths. Mol. Ecol. Notes 7:974–977.

Flot J.-F. 2010. SeqPHASE: a web tool for interconverting PHASE input/output files and FASTA sequence alignments. Mol. Ecol. Resour. 10:162–166.

Flot J.-F., Blanchot J., Charpy L., Cruaud C., Licuanan W., Nakano Y., Payri C., Tillier S. 2011. Incongruence between morphotypes and genetically delimited species in the coral genus *Stylophora*: phenotypic plasticity, morphological convergence, morphological stasis or interspecific hybridization? BMC Ecol. 11:22.

Flot J.-F., Couloux A., Tillier S. 2010. Haplowebs as a graphical tool for delimiting species: a revival of Doyle's "field for recombination" approach and its application to the coral genus *Pocillopora* in Clipperton. BMC Evol. Biol. 10:372.

Flot J.-F., Dahl M., André C. 2013. *Lophelia pertusa* corals from the Ionian and Barents seas share identical nuclear ITS2 and near-identical mitochondrial genome sequences. BMC Res. Notes 6:144.

Flot J.-F., Tillier A., Samadi S., Tillier S. 2006. Phase determination from direct sequencing of length-variable DNA regions. Mol. Ecol. Notes 6:627–630.

Folmer O., Black M., Hoeh W., Lutz R., Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotechnol. 3:294–299.

Fujisawa T., Barraclough T.G. 2013. Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: a revised method and evaluation on simulated data sets. Syst. Biol. 62:707–724.

Harrigan R.J., Mazza M.E., Sorenson M.D. 2008. Computation vs. cloning: evaluation of two methods for haplotype determination. Mol. Ecol. Resour. 8:1239–1248.
Hasegawa M., Kishino H., Yano T.A. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.
Hebert P.D.N., Cywinska A., Ball S.L., deWaard J.R. 2003a. Biological identifications through DNA barcodes. Proc. R. Soc. B Biol. Sci. 270:313–321.
Hebert P.D.N., Penton E.H., Burns J.M., Janzen D.H., Hallwachs W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc. Natl. Acad. Sci. U. S. A. 101:14812–14817.
Hebert P.D.N., Ratnasingham S., deWaard J.R. 2003b. Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. Proc. R. Soc. B Biol. Sci. 270:S96–S99.
Hey J. 2010. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. Mol. Biol. Evol. 27:921–933.
Hurvich C.M., Tsai C.-L. 1989. Regression and time series model selection in small samples. Biometrika 76:297–307.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.
Komsta L., Novomestky F. 2012. *moments*: Moments, cumulants, skewness, kurtosis and related tests. R package version 0.13. Available from: URL http://cran.r-project.org/web/packages/moments/, last accessed January 2015.
Lanave C., Preparata G., Saccone C., Serio G. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86–93.
Lefébure T., Douady C.J., Gouy M., Gibert J. 2006. Relationship between morphological taxonomy and molecular divergence within Crustacea: proposal of a molecular threshold to help species delimitation. Mol. Phylogenet Evol. 40:435–447.
Li X. 2012. Molecular and evolutionary insights into sexual marine mammals and asexual bdelloid rotifers [thesis]. University of Namur. p. 181
Lim G.S., Balke M., Meier R. 2012. Determining species boundaries in a world full of rarity: singletons, species delimitation methods. Syst. Biol. 61:165–169.
Lohse K. 2009. Can mtDNA barcodes be used to delimit species? A response to Pons et al. (2006). Syst. Biol. 58:439–442.
Mendelson T.C., Shaw K.L. 2005. Rapid speciation in an arthropod. Nature 433:375–376.
Miralles A., Vences M. 2013. New metrics for comparison of taxonomies reveal striking discrepancies among species delimitation methods in *Madascincus* lizards. PLoS One 8:e68242.
Monaghan M.T., Wild R., Elliot M., Fujisawa T., Balke M., Inward D.J.G., Lees D.C., Ranaivosolo R., Eggleton P., Barraclough T.G., Vogler A.P. 2009. Accelerated species inventory on Madagascar using coalescent-based models of species delineation. Syst. Biol. 58:298–311.
Moore W.S. 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. Evolution 49:718–726.
Moyle R.G., Filardi C.E., Smith C.E., Diamond J. 2009. Explosive Pleistocene diversification and hemispheric expansion of a "great speciator". Proc. Natl. Acad. Sci. U. S. A. 106:1863–1868.
Navajas M., Boursot M. 2003. Nuclear ribosomal DNA monophyly versus mitochondrial DNA polyphyly in two closely related mite species: the influence of life history and molecular drive. Proc. R. Soc. B Biol. Sci. 270:S124–S127.
Papadopoulou A., Bergsten J., Fujisawa T., Monaghan M.T., Barraclough T.G., Vogler A.P. 2008. Speciation and DNA barcodes: testing the effects of dispersal on the formation of discrete sequence clusters. Phil. Trans. R. Soc. B Biol. Sci. 363:2987–2996.
Pfister R., Schwarz K.A., Janczyk M., Dale R., Freeman J. 2013. Good things peak in pairs: a note on the bimodality coefficient. Front Psychol. 4:700
Phillimore A.B., Price T.D. 2008. Density-dependent cladogenesis in birds. PLoS Biol. 6:e71.
Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.
Posada D. 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25:1253–1256.
Puillandre N., Lambert A., Brouillet S., Achaz G. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Mol. Ecol. 21:1864–1877.
Rannala B., Yang Z. 2013. Improved reversible jump algorithms for Bayesian species delimitation. Genetics 194:245–253.
Reid N., Carstens B. 2012. Phylogenetic estimation error can decrease the accuracy of species delimitation: a Bayesian implementation of the general mixed Yule-coalescent model. BMC Evol. Biol. 12:196.
Roach J.C., Glusman G., Smit A.F.A., Huff C.D., Hubley R., Shannon P.T., Rowen L., Pant K.P., Goodman N., Bamshad M., Shendure J., Drmanac R., Jorde L.B., Hood L., Galas D.J. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636–639.
Rowe K.C., Aplin K.P., Baverstock P.R., Moritz C. 2011. Recent and rapid speciation with limited morphological disparity in the genus *Rattus*. Syst. Biol. 60:188–203.
Russell A., Cox M., Brown V., McCracken G. 2011. Population growth of Mexican free-tailed bats (*Tadarida brasiliensis mexicana*) predates human agricultural activity. BMC Evol. Biol. 11:88.
Schwartz E. 1929. Das Vorkommen des Schimpansen auf den linken Kongo-Ufer. Revue de zoologie et de botanique africaines 16: 425–426.
Song H., Buhay J.E., Whiting M.F., Crandall K.A. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc. Natl. Acad. Sci. U. S. A. 105:13486–13491.
Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.
Tang C.Q., Humphreys A.M., Fontaneto D., Barraclough T.G. 2014. Effects of phylogenetic reconstruction method on the robustness of species delimitation using single-locus data. Methods Ecol. Evol. 5:1086–1094.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.
Won Y.-J., Hey J. 2005. Divergence population genetics of chimpanzees. Mol. Biol. Evol. 22:297–307.
Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U. S. A. 107: 9264–9269.
Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29:2869–2876.