2014

A preliminary framework for DNA barcoding, incorporating the multispecies coalescent

Mark Dowton
University of Wollongong, mdowton@uow.edu.au

Kelly A. Meiklejohn
University of Wollongong, km988@uow.edu.au

Stephen Cameron
Queensland University of Technology

James F. Wallman
University of Wollongong, jwallman@uow.edu.au

Publication Details

Dowton, M., Meiklejohn, K., Cameron, S. L. & Wallman, J. (2014). A preliminary framework for DNA barcoding, incorporating the multispecies coalescent. Systematic Biology, 63 (4), 639-644.

A preliminary framework for DNA barcoding, incorporating the multispecies coalescent

Abstract

The capacity to identify an unknown organism using the DNA sequence from a single gene has many applications. These include the development of biodiversity inventories (Janzen et al. 2005), forensics (Meiklejohn et al. 2011), biosecurity (Armstrong and Ball 2005), and the identification of cryptic species (Smith et al. 2006). The popularity and widespread use (Teletchea 2010) of the DNA barcoding approach (Hebert et al. 2003), despite broad misgivings (e.g., Smith 2005; Will et al. 2005; Rubinoff et al. 2006), attest to this. However, one major shortcoming to the standard barcoding approach is that it assumes that gene trees and species trees are synonymous, an assumption that is known not to hold in many cases (Pamilo and Nei 1988; Funk and Omland 2003). Biological processes that violate this assumption include incomplete lineage sorting and interspecific hybridization (Funk and Omland 2003). Indeed, simulation studies indicate that the concatenation approach (in which these two processes are ignored) can lead to statistically inconsistent estimation of the species tree (Kubatko and Degnan 2007). The purpose of this article is to initiate the development of a framework for "next-gen barcoding": one that incorporates the multispecies coalescent, but does so by comparing multiple gene sequences from an unknown taxon with a database of sequences.

Disciplines

Medicine and Health Sciences | Social and Behavioral Sciences

Publication Details

Dowton, M., Meiklejohn, K., Cameron, S. L. & Wallman, J. (2014). A preliminary framework for DNA barcoding, incorporating the multispecies coalescent. Systematic Biology, 63 (4), 639-644.

Running Head: DNA BARCODING WITH THE MULTISPECIES COALESCENT

A Preliminary Framework for DNA Barcoding, Incorporating the Multispecies Coalescent

Mark Dowton[1]*, Kelly Meiklejohn[2], Stephen Cameron[3], James Wallman[2]

[1]*Centre for Medical Bioscience, School of Biological Sciences, University of Wollongong, NSW 2522, Australia*

[2]*Institute for Conservation Biology and Environmental Management, School of Biological Sciences, University of Wollongong, NSW 2522, Australia*

[3]*School of Earth, Environmental & Biological Sciences, Queensland University of Technology, QLD 4001, Australia*

*Correspondence to be sent to: *School of Biological Sciences, University of Wollongong, NSW 2522, Australia; E-mail: mdowton@uow.edu.au*

DNA barcoding continues to enjoy widespread use, despite broad, mostly philosophical criticisms. Most notably, DNA barcoding assumes that gene trees and species trees are synonymous, but biological processes (such as incomplete lineage sorting and interspecific hybridization) are known to violate this assumption. Recently, an analytical solution was found for one of the major shortcomings of DNA barcoding– incomplete lineage sorting can now be incorporated into the model of analysis of gene sequences. Here, we propose a preliminary framework for DNA barcoding that incorporates the multispecies coalescent.

Keywords: barcode, coalescent, incomplete lineage sorting, Bayesian, species delimitation.

DNA Barcoding

The capacity to identify an unknown organism using the DNA sequence from a single gene has many applications. These include the development of biodiversity inventories (Janzen et al. 2005), forensics (Meiklejohn et al. 2011), biosecurity (Armstrong and Ball 2005), and the identification of cryptic species (Smith et al. 2006). The popularity and widespread use (Teletchea 2010) of the DNA barcoding approach (Hebert et al. 2003), despite broad misgivings [e.g. (Smith 2005, Will et al. 2005, Rubinoff et al. 2006)], attest to this. However, one major shortcoming to the standard barcoding approach is that it assumes that gene trees and species trees are synonymous, an assumption that is known not to hold in many cases (Pamilo and Nei 1988, Funk and Omland 2003). Biological processes that violate this assumption include incomplete lineage sorting and interspecific hybridization (Funk and Omland 2003). Indeed, simulation studies indicate that the

concatenation approach (in which these two processes are ignored) can lead to statistically inconsistent estimation of the species tree (Kubatko and Degnan 2007).

However, recent developments make a barcoding approach that utilizes a single locus outdated. The cost of sequencing multiple gene fragments is no longer inhibitory, but more importantly, a range of analytical approaches have been developed that account for incomplete lineage sorting (Degnan and Salter 2005, Edwards et al. 2007, Liu et al. 2008, Kubatko et al. 2009, Heled and Drummond 2010, Yang and Rannala 2010). These approaches incorporate coalescent theory into the analysis of species trees and species delimitation (Fujita et al. 2012), and are conveniently accessible as software programs (e.g. BEST, BPP, *BEAST, MrBayes v. 3.2, STEM, COAL). Although the GMYC (General Mixed Yule Coalescent) approach has also been developed for species delimitation (Pons et al. 2006), we do not consider it further here. It operates quite differently to the approaches outlined above [i.e. BEST, BPP, *BEAST, MrBayes v. 3.2, STEM, COAL]. The GMYC approach seeks to identify the shift in the rate of lineage branching that should be evident when interspecific evolutionary processes switch to population-level processes (Pons et al. 2006). Both empirical (Esselstyn et al. 2012) and simulation studies (Esselstyn et al. 2012, Fujisawa and Barraclough 2013) report that it performs poorly when effective population sizes and speciation rates are high, but within biologically relevant ranges.

Ideally, a ‘next-generation’ barcoding approach would (1) identify a minimal set of barcoding genes (perhaps specific to certain lineages), (2) generate a large and cladistically divergent database for comparisons, and (3) identify species using species delimitation approaches that incorporate the multispecies coalescent. The first two of these conditions are straightforward, and require only discussion (requirement 1) and resources

(requirement 2). However, the third requirement is much more problematic. Some of the recently developed approaches for species delimitation could not be used alone; for example, BPP requires a user-specified guide tree (Yang and Rannala 2010). All of the recently developed approaches are computationally intensive (Degnan and Rosenberg 2009), with many having practical limitations on the number of individuals that can be compared. By contrast, the current barcoding approach is able to compare enormous numbers of sequences in a very short time, primarily because the approach is analytically simple; a single sequence is compared with all sequences in the database by calculating all possible pairwise K2P distances. As long as exemplars exist within the database that have K2P distances below some predetermined threshold (usually 4%), the species is considered identified. The speed of analysis is due primarily to the use of distance-based measures.

The purpose of this paper is to initiate the development of a framework for 'next-gen barcoding': one that incorporates the multispecies coalescent, but does so by comparing multiple gene sequences from an unknown taxon with a database of sequences.

ANALYTICAL APPROACH

Our philosophy is that a second generation barcoding approach should incorporate the multispecies coalescent. There are currently three approaches that we are aware of that delimit species with a consideration of the multispecies colascent (Fujita et al. 2012): Brownie (O'Meara 2010), SpedeSTEM (Ence and Carstens 2011) and BPP (Yang and Rannala 2010). Brownie offers the most elegant approach, with estimation of the species tree and species delimitation (i.e. estimation of the delimited species tree) achieved in a single analysis. However, the approach is still being developed – it would therefore be premature to adopt this approach for barcoding. SpedeSTEM suffers from the assumption that the

underlying gene tree is known, whereas the BPP approach does not. In principle, any of these three approaches could be used for species delimitation, but BPP appears to be superior in the few studies that have made comparisons between them (Camargo et al. 2012). Nevertheless, this field is relatively new, and further comparisons would be useful (e.g. Monaghan et al. 2009).

For the above reasons, the BPP approach was used here for species delimitation. However, BPP is not designed to deduce the delimited species tree – it is designed only to delimit species. BPP must be provided with a guide tree of species lineages, with internal nodes sequentially collapsed in order to test alternative models of species delimitation. For these reasons, we paired BPP with *BEAST, a Bayesian approach for deducing the species tree in which the multispecies coalescent is modelled. The *BEAST analysis produced the guide tree for the subsequent BPP analysis.

TEST DATABASE

Our test database contained 428 specimens of sarcophagine flies (Diptera: Sarcophagidae), from 39 different species, collected for both traditional barcoding (Meiklejohn et al. 2012) and phylogenetic studies (Meiklejohn et al. 2013). Thus, all sequences have been reported in previous studies, and have been deposited in GenBank. Some species were represented by as many as 40 specimens, with populations sampled from across their Australian distribution. We consider that this resembles the depth of within-species sampling that is typical of the current (Barcode of Life) BOLD database. Two genes from every specimen were sequenced; mitochondrial cytochrome oxidase I (COI) and nuclear CAD (carbomoylphosphate synthase domain of *rudimentary*). This low number of

genes was chosen for both practical purposes, and because it is the minimum number needed for coalescent-based barcoding.

Generating Unknowns For Testing

We wanted to test whether an unknown specimen could be identified using our database and a coalescent-based barcoding approach. To generate unknown specimens, we randomly chose a single representative from each species in our dataset. The sequence of that represenative was then removed so that an identical match was not present. When choosing representatives, we focussed on species from the genus *Sarcophaga*, as the remaining taxa in our dataset served primarily as outgroups. In our dataset, there were 32 species of *Sarcophaga* that could be confidently identified by morphological examination; choosing a single representative from each *Sarcophaga* species produced 32 unknowns. One of these species was divided into two groups, however, because of high levels of molecular and morphological divergence discovered during a phylogenetic analysis of this group (Meiklejohn et al. 2013); these are represented as *bancroftorum* clade 1 and *bancroftorum* clade 2 in Table 1. Choosing a representative from each of these *bancroftorum* clades produced 33 unknowns in total. Some species in our database were represented by a single individual, so we were also interested to assess whether these singletons could be resolved as 'new species' (i.e. unrepresented in the database).

Alignment And Substitution Models

Nucleotide sequences were aligned as described elsewhere (Meiklejohn et al. 2013). There was only a single instance of an internal gap (in COI) in the amino acid sequence alignments. For this reason, when subsets of the database were used, aligned

sequences were extracted, rather than extracting the unaligned sequences and re-aligning them. The bestfit model of nucleotide analysis was then determined for each gene, using MrModelTest 2.3 (Nylander 2004) and the Aikaike Information Criterion. We initially considered dividing each gene into codon partitions (1st, 2nd and 3rd codon positions), and running each through MrModelTest. However, inspection of the codon position alignments indicated that (as expected) most of the variation was present at the 3rd codon position, with the 2nd codon position almost invariant. We saw little advantage in isolating the 2nd codon position and determining a model for an invariant or almost invariant partition. For this reason, we determined the bestfit model for each gene (i.e. all codon positions analysed as a single partition), but this assumption deserves further investigation.

*BEAST ANALYSES

Both genes were then analyzed in *BEAST (Heled and Drummond 2010). Each gene was analyzed as a separate partition, specifying the substitution model and site heterogeneity model as found by MrModelTest. Estimated base frequencies were used. The species tree prior was specified as the Yule Process, and the population size model was specified as the piecewise linear and constant root. The ploidy type was set as 'mitochondrial' for COI, and 'autosomal nuclear' for CAD. The clock rate prior was chosen from a gamma distribution. Tree search was initiated from a random tree. The MCMC analysis ran for 10 million generations with trees sampled every 1000 generations. Two independent runs were conducted for every analysis, and the results compared. Stationarity was assessed by importing the parameter file into Tracer v. 1.5 (Drummond et al. 2012), and assessing whether the ESS of all parameter values was >100. Generally, stationarity was reached after 1000 trees had been sampled (i.e. after 1 million

generations), but in some instances stationarity was not reached until 6000 trees had been sampled. Trees that were sampled prior to stationarity were discarded as burnin. An estimate of the phylogenetic tree was then made using TreeAnnotator (Drummond et al. 2012), using the Maximum Clade Credibility Tree setting, and the Medium Node height setting. The maximum clade credibility tree was then viewed in FigTree v. 1.3.1. This tree was then used as the guide tree in BPP analyses (Yang and Rannala 2010).

BPP ANALYSES

Datasets were then analyzed using BPP (Rannala and Yang 2003, Yang and Rannala 2010). A gamma prior G (2, 1000), with mean 2/2000 = 0.001, was used on the population size parameters (θs). The age of the root in the species tree (τ_0) was assigned the gamma prior G (2, 1000), while the other divergence time parameters were assigned the Dirichlet prior (Yang and Rannala, 2010: equation 2). Each dataset was analyzed four times: (1) with rjMCMC algorithm 0, ε = 5 (command speciesdelimitation = 1 0 5), no heredity multipliers; (2) with rjMCMC algorithm 1, α = 2, *m* = 1 (command speciesdelimitation = 1 1 2 1), no heredity multipliers; (3) with rjMCMC algorithm 0, ε = 5 (command speciesdelimitation = 1 0 5), with heredity multipliers (CAD heredity scalar = 1, COI heredity scalar = 0.25); and (4) with rjMCMC algorithm 1, α = 2, *m* = 1 (command speciesdelimitation = 1 1 2 1), with heredity multipliers (CAD heredity scalar = 1, COI heredity scalar = 0.25). We did not perform duplicates of each of these analyses and we concede that more comprehensive analyses are possible (e.g. systematically varying ε, α and *m*). Instead, we considered variation across the four analyses as likely to capture any variation that would be revealed by limitations in the analytical approach.

FILTERING THE NUMBER OF INDIVIDUALS IN THE DATABASE

Our analytical approach seems straightforward: analyse a dataset using *BEAST, then use the deduced phylogeny as a guide tree for the subsequent BPP analysis. Indeed this approach has been adopted by a number of authors to investigate the species-status of a number of specific groups (Leaché and Fujita 2010, Yang and Rannala 2010, Fuchs et al. 2011, Satler et al. 2013). However, the purpose of barcoding is the identification of an unknown specimens by comparison with a very large database, yet the analytical burden of both *BEAST and BPP is too great to work with large numbers of taxa (e.g. Degnan and Rosenberg 2009). This limitation demands that the database be initially screened to identify a subset of candidate taxa for further analysis.

In order to reduce the computational burden in *BEAST, we applied two strategies. In the first strategy, we reduced the total size of the database by reducing the number of replicates of each species. For each species that was represented by more than five individuals, we randomly chose five representatives. This reduced the size of the database from 416 representatives of *Sarcophaga* to 136. However, *BEAST analyses of unknowns together with this reduced database remained prohibitively slow. We further reasoned that, in order to achieve a barcoding identification, one need only identify conspecifics and closely related congenerics. In the second strategy to reduce the computational burden, we used the COI sequence of each unknown, and screened our database for individuals with K2P distances that were less than 10%. We used this fairly high level of sequence divergence (the standard barcode approach uses 4% to identify species) in order to capture both conspecifics and a broad range of candidate sister taxa. This approach did serve to reduce the number of individuals that were then included in subsequent analyses. Table 1 describes the various datasets that were produced with this approach. In some cases, large

numbers of taxa (up to 115) and species (up to 38 when unknowns were included) remained in the dataset. Nevertheless, analysis times were not prohibitive, with most runs complete within two days on a desktop computer (64-bit Windows with a CPU speed of 3.30 GHz). We anticipate that these run times could be markedly reduced, particularly if the software was written to take advantage of parallel architecture.

Although these two strategies enabled us to carry out *BEAST analyses for each of our 33 'unknowns' (Table 1), BPP can only perform analyses with a maximum of 19 taxa. As can be seen in Table 1, many of our analyses produced datasets with more than 19 taxa. In order to perform BPP analyses with these datasets, we inspected the *BEAST tree, and pruned the most remotely related taxa (compared with the unknown) from our dataset, until there were only 19 taxa for analysis. We then assessed the success rate of each of our 'unknown' identifications.

BPP Analyses After Filtering The Number Of Individuals In The Database

Analyses of the datasets represented in Table 1 by BPP were generally successful at identifying the unknowns. Of the 33 'unknowns' that could be reliably identified morphologically, 26 had conspecifics remaining in the database (after withdrawal of the unknown), while seven were singletons (indicated by an 'S' in Table 1). The inclusion of singletons was important, as it allowed us to assess whether the analyses would correctly identify them as new species. Of the 26 unknowns that had conspecifics in the database, each of the 26 was recovered with its conspecifics (Table 1). Further, for 22 of the unknowns, the node circumscribing the unknown and the conspecifics had low posterior probability in each of the four BPP analyses (range 0 – 0.44), indicating that the unknown should be identified as the conspecific; that is, there was low confidence in a speciation

event between the unknown and the conspecifics. In the remaining four unknowns that were less confidently identified, the posterior probabilities for the node circumscribing the unknown and the conspecifics was higher (>0.50, boxed values in Table 1) in some of the BPP analyses, but in no case were all four consistently high. For example, for *S. torvida*, posterior probability values ranged from 0.63 – 0.86, and for *S. crassipalpis*, values ranged from 0.47 – 0.79.

Each of the seven singletons was reliably identified as 'new species'. In all four of the BPP analyses, the node circumscribing the unknown and the closest relative had high posterior probability (range 0.99 – 1.00), indicating strong evidence for a speciation event between the unknown and its closest relatives. This result suggests that the analytical framework is able to identify new species relatively easily, and that if the level of evidence required to identify new species is set relatively high, the approach would have a high level of success. For example, if the posterior probability value has to be 0.98 or higher in each of the four BPP analyses for a new species to be recognized, then the success rate of the analyses described here would be 100%.

A Comparison with Barcoding

The coalescent-based *BEAST/BPP approach used here was very successful at identifying species boundaries. We also recently assessed how well barcoding delimits species in this same set of *Sarcophaga* species (Meiklejohn et al. 2012). We found that, of the 31 species of *Sarcophaga* examined (we exclude *S. bancroftorum* from our comparison here, as its taxonomic limits are not confidently known), 27 could be reliably distinguished by barcoding when a 4% sequence divergence threshold was applied. The four problematic taxa were *S. megafilosia*, *S. meiofilosia*, *S. crassipalpis* and *S ruficornis*. S. *megafilosia* and *S.*

meiofilosia had an interspecific divergence of 2.81%, while *S. crassipalpis* and *S ruficornis* had an interspecific divergence of 3.75%. The success rate of barcoding for this set of taxa is thus 87%, while the *BEAST/BPP approach had a success rate of 100%. However, ours is just one case study, and we encourage other investigators to compare the success rates of these two approaches in a broad range of taxonomic groups.

LIMITATIONS

The purpose of the present paper is to encourage the development of smarter approaches to barcoding – ones that incorporate the multispecies coalescent into the analysis of multilocus sequence data. We hope to convince the barcoding community that more biologically realistic approaches to barcoding are tractable, despite their computational overhead.

There are a number of limitations to our approach, but ones that we consider are easy to overcome with broader analyses. Perhaps the major limitation is reliance on a guide tree in the BPP analysis. Initial studies suggest that misspecification of the guide tree does not drastically influence the BPP analyses (Yang and Rannala 2010), but this deserves much more comprehensive investigation. There may be instances where a well-supported phylogeny is available for a particular group. In these instances, it may be wise to use this phylogeny as the guide tree in the BPP analyses, rather than the *BEAST generated phylogeny.

In order to identify a species, any database needs to be reduced in size so that not all conspecifics are included. We chose to include only five randomly selected specimens, but did not systematically optimize this choice. The relatively high success rate of species

identification in this study (100% if the criteria for new species recognition are chosen carefully) suggests that five is sufficient, and indeed, of the four unknowns that were less reliably identified, two were represented by just one or two conspecifics (although the other two were represented by four conspecifics). Although one could argue that increasing the number of conspecifics may improve the success rate, our initial analyses suggest new species are identified with high confidence, making it possible to simply set the cutoff for species identification high.

Another limitation of our approach is the filtering of candidates using K2P distances. Although we set this level high (10%) to limit our reliance on this non-phylogenetic measure, a phylogenetic filter would be much more philosophically consistent. An alternative approach could be to perform a *BEAST analysis with single representatives of each species, in order to identify some subset of the database for more detailed analysis. However, any barcoding database will have thousands to millions of representatives, and we find it difficult to envisage the avoidance of a non-phylogenetic filter being used in the initial stages of analysis.

We have shown that a coalescent-based approach to barcoding is possible with our *Sarcophaga* fly database. Although the traditional barcoding approach works reasonably well with this dataset (Meiklejohn et al. 2012), there were instances where the *BEAST/BPP approach appeared superior to traditional barcoding. For example, *S. megafilosia* and *S. meiofilosia* are considered sister species; both are parasitoids of the marine snail *Littoraria filosa*. Snails with shell lengths greater than 10 mm are parasitised by *S. megafilosia*, while snails with shells between four and 10 mm are parasitised by *S. meiofilosia* (Pape et al. 2000). Yet traditional barcoding does not recover them as separate species, due to the

relatively low level of interspecific COI sequence variation of 2.81%. By contrast, BPP analyses firmly resolve them as distinct species. As a first step towards our towards our proposed 'next-gen' barcoding framework, we encourage others to also assess their taxon of interest using analytical approaches that incorporate the multispecies coalescent.

FUNDING

This work was supported by the Australian Biological Resources Study; the Australian Research Council (grant numbers LP0883711, FT120100746); the Australian Federal Police; the NSW Police Force; and the Taxonomic Research Information Network (Emerging Priorities Program).

REFERENCES

Armstrong K.F., Ball S.L. 2005. DNA barcodes for biosecurity: invasive species identification. Philos. Trans. R. Soc. Lond. B 360:1813-1823.

Camargo A., Morando M., Avila L.J., Sites J.W. 2012. Species delimitation with ABC and other coalescent-based methods: A test of accuracy with simulations and an empirical example with lizards of the *Liolaemus darwinii* complex (Squamata: Liolaemidae). Evolution 66:2834-2849.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332-340.

Degnan J.H., Salter L.A. 2005. Gene tree distributions under the coalescent process. Evolution 59:24-37.

Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969-1973.

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104:5936-5941.

Ence D.D., Carstens B.C. 2011. SpedeSTEM: a rapid and accurate method for species delimitation. Mol. Ecol. Res. 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. Lond. B 279:3678-3686.

Fuchs J., Fjeldsa J., Bowie R.C. 2011. Diversification across an altitudinal gradient in the Tiny Greenbul (*Phyllastrephus debilis*) from the Eastern Arc Mountains of Africa. BMC Evol. Biol. 11:117.

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.

Fujita M.K., Leaché A.D., Burbrink F.T., McGuire J.A., Moritz C. 2012. Coalescent-based species delimitation in an integrative taxonomy. Trends Ecol. Evol. 27:480-488.

Funk D.J., Omland K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397-423.

Hebert P.D.N., Ratnasingham S., de Waard J.R. 2003. Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. Proc. R. Soc. Lond. Ser. B-Biol. Sci. 270:S96-S99.
Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570-580.
Janzen D.H., Hajibabaei M., Burns J.M., Hallwachs W., Remigio E., Hebert P.D.N. 2005. Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. Philos. Trans. R. Soc. Lond. B 360:1835-1845.
Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinf. 25:971-973.
Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17-24.
Leaché A.D., Fujita M.K. 2010. Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). Proc. R. Soc. Lond. B 277:3071-3077.
Liu L., Pearl D.K., Brumfield R.T., Edwards S.V. 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution 62:2080-2091.
Meiklejohn K.A., Wallman J.F., Cameron S.L., Dowton M. 2012. Comprehensive evaluation of DNA barcoding for the molecular species identification of forensically important Australian Sarcophagidae (Diptera). Invert. Syst. 26:515-525.
Meiklejohn K.A., Wallman J.F., Dowton M. 2011. DNA-based identification of the forensically important Australian Sarcophagidae (Diptera). Int. J. Legal Med. 125:27-32.
Meiklejohn K.A., Wallman J.F., Pape T., Cameron S.L., Dowton M. 2013. Utility of COI, CAD and morphological data for resolving relationships within the genus *Sarcophaga* (*sensu lato*) (Diptera: Sarcophagidae): a preliminary study. Mol. Phylogenet. Evol. in press.
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*., et al.* 2009. Accelerated species inventory on Madagascar using coalescent-based models of species delineation. Syst. Biol. 58:298-311.
Nylander J.A.A. 2004. MrModeltest 2.0. Uppsala University, Program distributed by the author.
O'Meara B.C. 2010. New heuristic methods for joint species delimitation and species tree inference. Syst. Biol. 59:59-73.
Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568-583.
Pape T., McKillup S.C., McKillup R.V. 2000. Two new species of *Sarcophaga* (*Sarcorohdendorfia*) Baranov (Diptera: Sarcophagidae), parasitoids of *Littoraria filosa* (Sowerby) (Gastropoda: Littorinidae). Aust. J. Entomol. 39:236-240.
Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595-600,603-609.
Rannala B., Yang Z.H. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645-1656.
Rubinoff D., Cameron S., Will K. 2006. A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. J. Hered. 97:581-594.
Satler J.D., Carstens B.C., Hedin M. 2013. Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, *Aliatypus*). Syst. Biol. 62:805-823.
Smith M.A., Woodley N.E., Janzen D.H., Hallwachs W., Hebert P.D.N. 2006. DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). Proc. Natl. Acad. Sci. USA 103:3657-3662.
Smith V.S. 2005. DNA barcoding: perspectives from a “partnerships for enhancing expertise in taxonomy” (PEET) debate. Syst. Biol. 54:841-844.

Teletchea F. 2010. After 7 years and 1000 citations: comparative assessment of the DNA barcoding and the DNA taxonomy proposals for taxonomists and non-taxonomists. Mitochondrial DNA 21:206-226.
Will K.W., Mishler B.D., Wheeler Q.D. 2005. The perils of DNA barcoding and the need for integrative taxonomy. Syst. Biol. 54:844-851.
Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. USA 107:9264-9269.