



## Assigning sequences to species in the absence of large interspecific differences

Melanie Lou, G. Brian Golding\*

Department of Biology, McMaster University, Hamilton, Ontario, Canada L8S 4K1

### ARTICLE INFO

#### Article history:

Received 6 October 2009

Revised 30 December 2009

Accepted 2 January 2010

Available online 11 January 2010

#### Keywords:

Barcoding

Species identification

Coalescent

Segregating sites

*Drosophila* species

### ABSTRACT

Barcoding is an initiative to define a standard fragment of DNA to be used to assign unknown sequences to existing known species groups that have been pre-identified externally (by a taxonomist). Several methods have been described that attempt to place this assignment into a Bayesian statistical framework. Here we describe an algorithm that makes use of segregating sites and we examine how well these methods perform in the absence of an interspecific 'barcoding gap'. When a barcoding gap exists, that is when the data are clearly delimited, most methods perform well. Here we have used data from the *Drosophila* genus because this genus includes sibling species and the species relationships within this species while complex are, arguably, better understood than in any other group. The results show that the Bayesian methods perform well even in the absence of a barcoding gap. The sequences from *Drosophila* are correctly identified and only when the degree of incomplete lineage sorting is extreme in simulations or within the *Drosophila* species, do they fail in their identifications and even then, the "correct" species has a high posterior probability.

© 2010 Elsevier Inc. All rights reserved.

### 1. Introduction

DNA barcoding involves the use of a short DNA sequence as a means to taxonomically identify a specimen (Hebert et al., 2003a,b; Remigio and Hebert, 2003). The key to this concept is to standardize the segment of DNA used for barcoding and then to construct a database of this sequence from as many taxonomically identified species as possible. Storing these data in a searchable database permits new or unknown specimens to be identified via a comparison of their sequence with sequences from characterized species. The recognized utility of this methodology has resulted in a global, synchronized effort with more than 100 member organizations (including museums, zoos, botanical gardens and universities) involved in setting a global standard in taxonomy and in creating a database of DNA barcode sequences.

Although the usefulness of this approach is well established (see for example Hebert et al., 2003b, 2004; Hajibabaei et al., 2006), some taxonomic groups, such as cowries (Meyer and Paulay, 2005) and tiger moths (Schmidt and Sperling, 2008), have shown an unacceptably high error rate for identification by DNA barcodes. Part of the reason for this discrepancy is due to similar levels of intra- and interspecific divergence. Under these conditions there may be a small amount of divergence between species relative to the amount of divergence within species. The difference between intra- and interspecific divergences is known as the barcoding

gap. Cognato (2006) found substantial overlap between levels of intra- and interspecific variation within several orders of insects resulting in the failure to correctly diagnose insect species for 45% of the cases. Within Diptera, there are congeneric sequences whose distance is within 1% (Meier et al., 2008). Similarly, the Lepidopteran family Lycaenidae showed an 18% overlap between intra- and interspecific COI divergence (Wiemers and Fiedler, 2007). An overlap may occur for a number of reasons. It may occur when there is a wide variation in rates of molecular evolution among lineages (Sparks and Smith, 2006; Huang et al., 2008). The COI from some animals, such as coral (Huang et al., 2008), evolves too slowly to be useful for barcoding. Incomplete lineage sorting (paraphyly or polyphyly; Moritz and Cicero, 2004; Pollard et al., 2006; Wiemers and Fiedler, 2007; Aliabadian et al., 2009) and poor taxonomy may also explain the lack of a barcoding gap. An inference must be made as to which species (or other taxonomic group) the sequence belongs. It is often difficult to discern whether or not differences between the query sequences and sequences within the database are due to intraspecific differences or if they are an indication of interspecific differences. The effectiveness of barcoding is associated with a clear distinction between levels of divergence with the level of interspecific divergence greater than intraspecific divergence. Indeed it has been shown that the simplest of methods performs well under these circumstances (Ross et al., 2008; Austerlitz et al., in press). Although it is not impossible to identify a species in the absence of a barcoding gap, this deficiency makes it much more difficult.

However, these methods lack ways to measure the confidence with which an assignment is made. Hence, there is a need for

\* Corresponding author. Fax: +1 905 522 6066.

E-mail address: [Golding@McMaster.CA](mailto:Golding@McMaster.CA) (G. Brian Golding).

statistical methods to determine the most appropriate assignment and the degree of confidence with which this assignment can be made, particularly when a barcode gap might be small or non-existent. Frézal and Leblois (2008) note that population genetics theory is required to account for the level of uncertainty that is contributed by these processes. Here, only Bayesian methods will be examined because these provide the necessary statistical strength to distinguish between well supported assignments versus poor assignments and to provide a strong statistical framework.

There are two Bayesian methods that have been proposed to date. The first is a method that uses the coalescent (Abdo and Golding, 2007). This method calculates the likelihood of coalescents for sequences known to originate from a particular species and then calculates the change in the likelihood when the query sequence is considered a member of this species. The assignment of an unknown individual sequence is to the group,  $i$ , that minimizes the posterior risk,  $R_i$ . The posterior risk of group  $i$  reflects the posterior probability that the sequence belongs to a coalescent with sequences from species  $i$  and the 'loss' of making the decision that the query sequence originated from species  $i$ . Here, loss is defined as the difference between the sequence of the unknown individual and the consensus sequence of the assumed correct group  $k$ . The mathematical details for calculating the posterior risk, loss and posterior probability are given in Abdo and Golding (2007).

Coalescent methods can be time consuming for data sets with a large number of sequences since it must generate enough coalescent trees to adequately sample all possible coalescent events. Therefore, the coalescent method is amended in this paper by replacing the coalescent-based Markov Chain Monte Carlo (MCMC) algorithm with one that makes use of the number of segregating sites from the sequences of a single species. A segregating sites method uses only sites at which there is a nucleotide change. The theory behind segregating sites allows closed form solutions to be used in place of time consuming MCMCs. It is therefore very rapid. It does, however, involve a loss of information and compresses the entire collection of sequence data into a single number. For Barcoding sequences, which can generally be assumed to be closely related sequences, the loss of information is usually minor.

Another Bayesian method is the SAP (statistical assignment package) algorithm that incorporates taxonomic information from NCBI and uses this information to impose topology constraints on the trees sampled from a MCMC. The probability of assignment is the number of sampled trees showing the unknown sequence branching with a sequence from species  $i$  (Munch et al., 2008a,b). This approach assumes that the branching pattern, as delimited by the taxonomy, is realistic and accurate. It also does not take into account the variability that might be expected around this branching pattern due to unsampled intraspecific differences and it assumes that the species are monophyletic. However, several studies have shown that the expectation of monophyly for recently diverged species is not realistic (Knowles and Carstens, 2007; Hickerson et al., 2006; Hudson and Coyne, 2002). It is noted by Nielsen and Matz (2006) that false species assignments can be caused by incomplete lineage sorting and by random mutation processes that can mimic incomplete lineage sorting.

The comparison of population genetic methodologies to phylogenetic methods done here suggests that the posterior probability of species identification is, in general, much smaller for the former. This suggests that these methods are more conservative than phylogenetic methods. The underlying cause of these differences in posterior probabilities are shown to be because these methods estimate the probabilities of different quantities.

## 2. Materials and methods

### 2.1. Evaluating assignment with segregating sites

Following Abdo and Golding (2007), we evaluate the probability of assigning an unknown sequence to a taxonomic grouping in a Bayesian context. For some unknown DNA sequence,  $x$ , the goal is to assign the species from which this sequence was taken to the correct taxonomic group,  $k$ . Hence, we wish to find:

$$Pr(x \in k|x, D, \theta)$$

where  $D$  is a database of known sequences with  $n$  distinct taxonomic groups and  $\theta$  is a known collection of evolutionary parameters. The assignment of sequence  $x$  must be made to one of the taxonomic groups.

It is assumed that different groups that are potential targets of the assignment are fully pre-specified. Each group is assumed to form a panmictic population that follows a Wright–Fisher, neutral model of evolution that does not allow recombination, selection, or migration. Hence, the evolutionary process within each group is governed by one parameter, which is the expected number of mutational events between sequences. This quantity is dependant upon a population measure,  $\theta = 4N_e\mu$ , and is in turn, reflected in the number of segregating sites between sequences.

Using Bayes rule, assuming that the pre-sampled individuals are assigned correctly by external taxonomists, assuming independence of the evolutionary history between groups and assuming uniform priors, this can be calculated as:

$$Pr(x \in k|x, D, \theta) = \frac{Pr(x, D_k|x \in k, \theta_k)/Pr(D_k|\theta_k)}{\sum_j Pr(x, D_j|x \in j, \theta_j)/Pr(D_j|\theta_j)}$$

(see Abdo and Golding, 2007, for a derivation).

A risk function can be evaluated using this probability and traditionally, an assignment decision is based on the assignment with minimum risk. The risk function can be defined as:

$$R_i = \sum_k L(k, i)Pr(x \in k|x, D, \theta)$$

where  $R_i$  is the risk of making the assignment to species  $i$  and  $L(k, i)$  is the loss associated with an assignment to species  $i$  when the correct assignment should be to species  $k$  and  $Pr(x \in k|x, D, \theta)$  is the posterior probability of membership of the unknown sequence  $x$  to taxonomic group  $k$ .

In Abdo and Golding (2007) a method to evaluate  $Pr(x, D_k|x \in k, \theta_k)$  using the coalescent and an MCMC is implemented. However, it is also possible to evaluate  $Pr(x, D_k|x \in k, \theta_k)$  using the theory of segregating sites. If the sequence data  $\{x, D_k\}$  has  $s$  segregating sites, then the probability of the data given  $\theta_k$  can be approximated by the probability corresponding to the number of segregating sites,  $s$ . Hence,

$$Pr(x, D_k|x \in k, \theta_k) \sim Pr(S = s|n, \theta_k)$$

where  $s$  is the number of sequences in  $\{x, D_k\}$ . The basic recursive definition for the probability that a sample of  $n$  sequences will have  $s$  segregating sites is:

$$Pr(S = s|n, \theta_k) = \frac{n-1}{n-1+\theta_k} Pr(S = s|n-1, \theta_k) + \frac{\theta_k}{n-1+\theta_k} Pr(S = s-1|n, \theta_k)$$

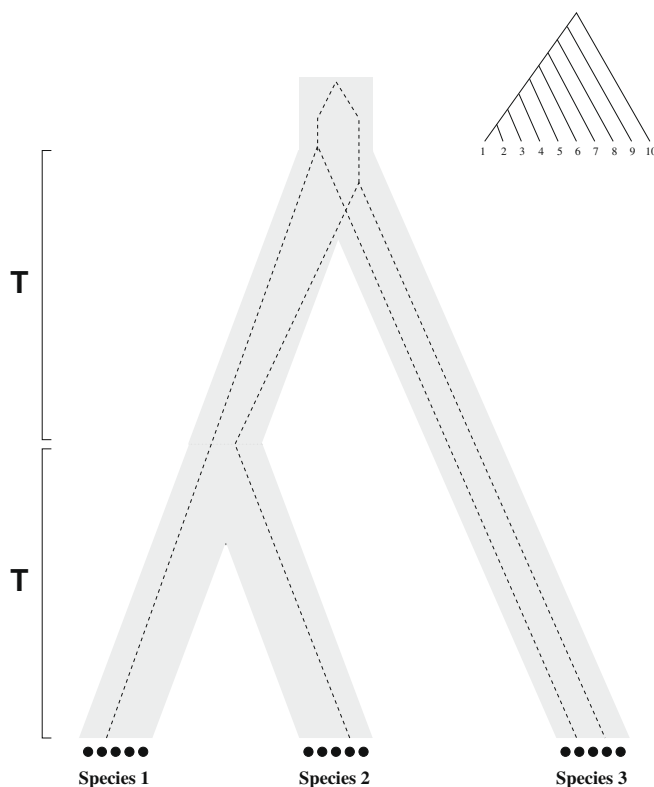
This recursion makes the assumption that an infinite sites model holds, that the populations are equilibrium single random mating populations of size  $N_e$  with mutation to new alleles at a rate  $\mu$ .

This recursion has been solved by Tavaré (1984) to yield a closed form solution of:

$$\Pr(S = s | n, \theta_k) = \frac{n-1}{\theta_k} \sum_{i=1}^{n-1} (-1)^{i-1} \binom{n-2}{i-1} \left( \frac{\theta_k}{i + \theta_k} \right)^{s+1}$$

Our implementation of this formula was found to occasionally be numerically unstable. Therefore, if the closed form solution did not satisfy the recursion with numerical accuracy, we then did an evaluation of the complete recursion.

Attention is focused here on the posterior probability rather than risk (multiple loss functions can be used to quantify risk as described in, [Abdo and Golding \(2007\)](#)) to make the results from the segregating sites algorithm comparable with those from the SAP algorithm. To test the assignment of unknown queries using the segregating sites algorithm, we conducted a simulation to test the performance of the algorithm in the absence of a ‘barcoding gap’. The simulations use a multi-species coalescent ([Degnan and Rosenberg, 2009](#)) to model 10 species with a pectinate species tree ([Fig. 1](#)). Each of the 10 species has five lineages. The ‘unknown’ query sequence is simulated as the sixth sequence from the first species. Sequences of length 600 bp were simulated using the coalescent tree. The entire length of the sequences were allowed to accumulate substitutions at a constant rate, defined by the parameter  $\theta$ . This value of  $\theta$  is the total mutation rate for all sites in the length of the sequence. At every time interval, defined by  $T$ , those sequences that had not yet coalesced to a common ancestor were added to the sequences from other “species”. The time intervals  $T$  were scaled according to  $2N_e$  generations and represent the time back to speciation events. However, the coalescents may extend beyond multiple speciation events depending on the size of  $T$ . In these simulations  $T$ , ranges from 0.5 to 3.0. When  $T = 3.0$ , the level of interspecific divergence is greater than the level intraspecific



**Fig. 1.** The simulation scheme used. Species are added to the tree sequentially up to a total of 10. Three species are expanded here. The length of time separating the divergence of each species can be short,  $T \ll N$ , allowing incomplete lineage sorting to occur (as illustrated here lineages within species #2 are more closely related to lineages within species #3 than they are to species #1 despite the implied species relationships).

divergence and this represents the ideal situation where a barcoding gap exists and each species is usually monophyletic and is distinct from every other species; in this scenario, we expect a high proportion of correct assignments. When  $T = 0.5$ , there is a lack of a barcoding gap which may lead to incomplete lineage sorting; we expect a lower proportion of correct assignments. The simulations were repeated 10,000 times and the results are given in [Table 2](#).

An advantage of the segregating sites algorithm is its speed. The method of segregating sites obviously involves a loss of information in moving from a full coalescent evaluation to an evaluation of a single number, the number of segregating sites. However, it gains a great deal of speed compared to a coalescent method. The analysis of 10,000 simulation runs took only seconds. In addition, for actual data collected from nature, the sequences are from highly conserved genes. Such sequences are anticipated to be very similar and the opportunity for multiple mutations to arise at a single site is small. The results described below document the efficacy of this method.

## 2.2. The SAP algorithm

SAP version 1.0.6 was downloaded and installed locally ([Munch et al., 2008a](#)). An in-house database constructed from sequences from the *Drosophila* genus were used for searches conducted with a local version of BLAST v. 2.2.17. The local database was annotated using the taxonomic information from NCBI. The set of sequence homologues were aligned using a local copy of CLUSTALW v. 2.0 ([Thompson et al., 1994](#)).

## 2.3. *Drosophila* sequences

The *Drosophila* species provide a good data set to test the ability of algorithms to assign sequences to species in the absence of a barcoding gap. Many species are sibling species with small interspecific differences and some have no barcoding gap at all with identical sequences shared among species.

A *Drosophila* data set consisting of 1542 CO1 sequences from 314 species was collected from NCBI and/or Flybase ([Tweedie et al., 2009](#)) February 2009. Alignment of sequences within a species was done using the corresponding amino acid sequence via MUSCLE ([Edgar, 2004](#)) and then translated back to DNA using TRANALIGN. Sequences with large indels (>10 amino acids) were removed. The sequences were trimmed to the barcode region (663 bp). Sequences were deleted entirely if they contained less than 650 bp. Species with two or fewer sequences were removed. Sequences were ensured to originate from distinct strains, from independent wild isolates or from different laboratories, as listed in the GenBank annotation. If there were multiple copies from the same source, the longest sequence from a single strain, isolate, or laboratory was used (refer to [Supplementary material](#) for a listing of strains and isolates of *Drosophila* species used in the study with, where available, references to literature containing information on where the strain or isolate originates). The remaining data set comprised of 616 sequences from 19 species. A summary of the sequences is shown in [Table 1](#) (the species and group designations were taken from NCBI; groups are listed only if there are multiple members present). Other commonly known *Drosophila* species have insufficient numbers of sequences or insufficient information that they represent distinct samples to be included by these criteria.

A diagram of the topological relationships between *Drosophila* species is shown in [Fig. 2](#). This diagram is patterned after a phylogeny constructed from Kimura 2-parameter distances ([Kimura, 1980](#)) using the Neighbor Joining method ([Saitou and Nei, 1987](#)) and with the phylogeny from Flybase (<http://flybase.org/>)

**Table 1**  
*Drosophila* COI sequences tested (Monophyly is taken from the diagram in Fig. 2).

Group	Species	Monophyletic	Sequences
Melanogaster	<i>D. mauritiana</i>	No	3
Melanogaster	<i>D. melanogaster</i>	–	10
Melanogaster	<i>D. simulans</i>	No	27
Quadrissetata	<i>D. barutani</i>	–	6
Quadrissetata	<i>D. beppui</i>	–	3
Quinaria	<i>D. falleni</i>	–	15
Quinaria	<i>D. innubila</i>	–	29
Quinaria	<i>D. recens</i>	No	136
Quinaria	<i>D. subquinaria</i>	No	136
Repleta	<i>D. arizonae</i>	–	17
Repleta	<i>D. mettleri</i>	–	24
Repleta	<i>D. mojavensis</i>	–	47
Repleta	<i>D. navojoa</i>	–	4
Repleta	<i>D. nigrospiracula</i>	–	10
Virilis	<i>D. montana</i>	–	42
Virilis	<i>D. virilis</i>	–	11
–	<i>D. angor</i>	No	13
–	<i>D. daruma</i>	–	4
–	<i>D. packea</i>	–	79
Total			616

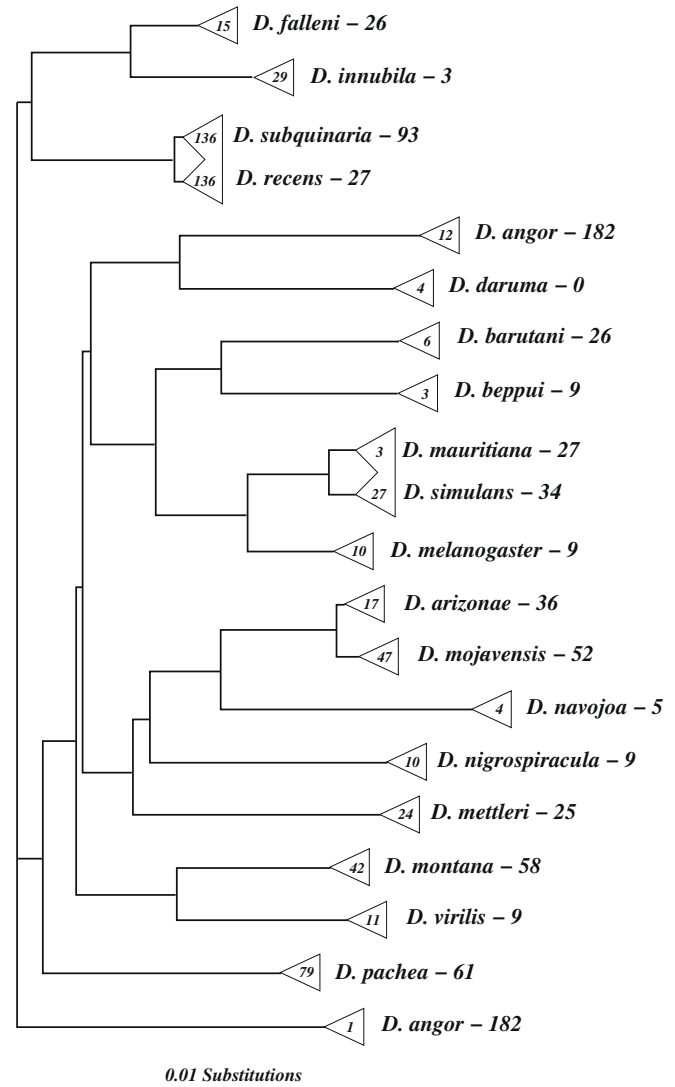
with the exceptions of species *D. angor*, *D. barutani*, *D. beppui*, and *D. daruma* (Wang et al., 2006) which are not listed in Flybase. Based on fossil, biogeographic, and molecular clock data, subgenera *Drosophila* (*D. melanogaster*, *D. simulans*, and *D. mauritiana*) and *Sophophora* are estimated to have diverged approximately  $62.9 \pm 12.4$  million years (MYA) (Powell, 1997; Tamura et al., 2004). Thus, there should be enough interspecific divergence to prevent the assignment of unknown sequences to the incorrect subgenus.

Some of these species are considered sibling species and are difficult to distinguish by anyone other than trained experts. Nevertheless, the species and their relationships are well known (Kelly and Noor, 1996; Powell, 1997). The ability of some taxa to create semi-sterile, usually uni-directional, hybrids has been well documented (Noor, 1995). The species pairs *D. arizonae* & *D. mojavensis*, *D. mauritiana* & *D. simulans*, and *D. recens* & *D. subquinaria* are considered sibling species. In the case of *D. mauritiana* and *D. simulans*, there is a haplotype identified as originating from *D. mauritiana* that is identical to that in *D. simulans* (Satta and Takahata, 1990; Ballard, 2000a,b). The divergence date of these species is estimated as  $0.93 \pm 0.49$  MYA (Tamura et al., 2004) and so this phenomenon may be due to incomplete lineage sorting or introgression. Similarly, 2 haplotypes (with 1 and 2 representative sequences respectively) out of 109 COI haplotypes from *D. subquinaria* are identical to 2 haplotypes (containing 16 and 66 sequences respectively) out of 36 COI haplotypes from *D. recens*. These are the result of *Wolbachia*-mediated introgression (Shoemaker et al., 2004; Jaenike et al., 2006). Although *D. arizonae* and *D. mojavensis* are sibling species with an estimated divergence time of 1.91–2.97 MYA, their sequences are similar but they do not share any haplotypes (Reed et al., 2007).

### 3. Results

#### 3.1. Simulation properties of a segregating sites algorithm

Simulations were conducted to test how well the segregating sites algorithm will assign queries when there is a known degree of similarity between the correct species and its closest relative(s). In this case, each species is progressively more and more distant from the first species (Fig. 1). The first species is the origin of the query sequence, and the branch length back to the common ancestor encompassing the next species ranges from  $T = 0.5$ –3.0. With



**Fig. 2.** A diagram of the relationships of the *Drosophila* COI sequences. The numbers in the triangle give the number of sequences used from each species and the number after the species name is the number of segregating sites within these sequences.

the simulation, the degree to which the histories of the individual species are distinct can be measured by examining the degree to which lineage sorting is complete. The results of this simulation are shown in Table 2.

The first row for each simulation run in Table 2 gives an indication of the extent of incomplete lineage sorting. When the interspecific distance between species is very short ( $T = 0.5$ ) lineage sorting is seldom complete within species 1. Only 15% of the 10,000 simulations have a distinct monophyletic lineage for the five sequences in species 1 while 28% have lineages that confuse species 1 and 2. Nevertheless, the segregating sites method correctly assigns 44% of the queries to species 1. Given the short divergence time and the comparatively small opportunity for distinct substitutions to occur, it is not surprising that the average posterior probabilities for these assignments are low. Because of the similarity between these species, the degree of confidence in these assignments is low.

In general, assignments to species further and further away from species 1 occur in rapidly declining numbers and with declining posterior probabilities. In addition, the estimated value of  $\theta$  increases. Thus, the assignments are made to more distantly related

**Table 2**

Simulation results based on the assignment of 10,000 queries. The query sequence always originates from Taxon #1. The first row indicates how many coalescents for Taxon #1 included sequences from other species (indicated by the column). The second row gives the number of times each species had the highest posterior probability.

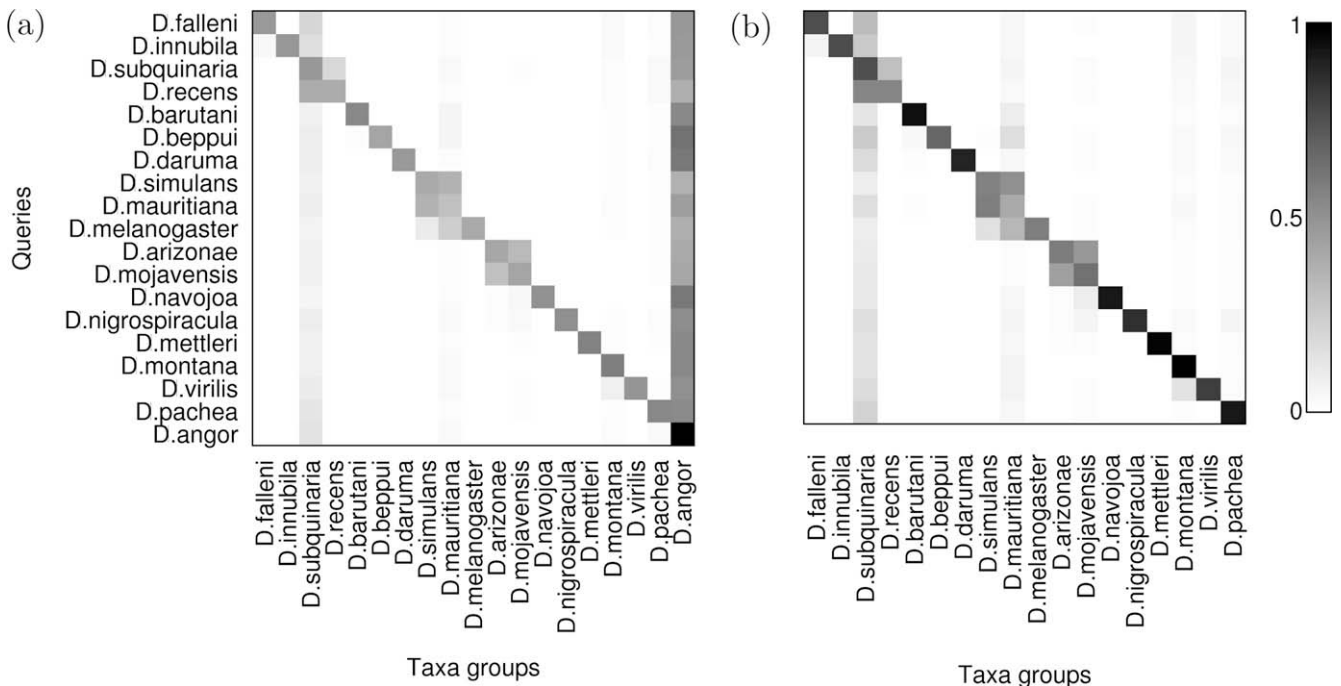
	Taxa									
	1	2	3	4	5	6	7	8	9	10
<i>T</i> = 3.0, $\theta$ = 2.0										
No. of taxon 1 coalescents including other taxa	9281	680	39	0	0	0	0	0	0	0
No. assigned to each taxa	9388	564	32	14	2	0	0	0	0	0
Avg. posterior	0.729	0.516	0.499	0.496	0.508	0	0	0	0	0
Avg. $\hat{\theta}$	1.645	2.425	3.425	4.460	4.560	0	0	0	0	0
<i>T</i> = 2.0, $\theta$ = 2.0										
No. of taxon 1 coalescents including other taxa	7961	1744	249	43	2	1	0	0	0	0
No. assigned to each taxa	8543	1245	173	27	7	3	2	0	0	0
Avg. posterior	0.601	0.444	0.404	0.477	0.439	0.347	0.551	0	0	0
Avg. $\hat{\theta}$	1.705	2.145	2.748	3.493	4.594	5.280	6.720	0	0	0
<i>T</i> = 1.0, $\theta$ = 2.0										
No. of taxon 1 coalescents including other taxa	4497	3497	1286	471	147	57	28	8	6	3
No. assigned to each taxa	6468	2238	836	311	95	33	9	7	2	1
Avg. posterior	0.394	0.322	0.286	0.265	0.254	0.244	0.277	0.306	0.303	0.203
Avg. $\hat{\theta}$	1.834	2.136	2.303	2.323	2.752	3.423	4.329	5.011	3.140	2.880
<i>T</i> = 0.5, $\theta$ = 2.0										
No. of taxon 1 coalescents including other taxa	1522	2846	2176	1350	823	502	304	201	109	167
No. assigned to each taxa	4431	2154	1367	849	520	308	162	114	61	34
Avg. posterior	0.263	0.229	0.209	0.195	0.186	0.179	0.170	0.176	0.161	0.161
Avg. $\hat{\theta}$	2.044	2.212	2.211	2.203	2.213	2.247	2.189	2.343	2.374	1.726

species when the number of mutations is, by chance, larger and further blurs the species level distinctions.

As *T* increases, the proportion of incomplete lineage sorting declines and the assignments become more accurate. In every circumstance, however, the proportion of correctly assigned query sequences is larger than the proportion of species #1 that have incomplete lineage sorting. Thus, the correct assignment of sequences can occur even without a barcoding gap but the confidence in that assignment can be variable.

3.2. The assignment of *Drosophila* sequences

Each *Drosophila* sequence was removed in turn and then assigned to a member species by the algorithms discussed here. The results for the segregating sites algorithm are shown in Fig. 3. The figure gives the average posterior probability that a query sequence (on the y-axis) is assigned to any one of the taxa (on the x-axis). The assignments of *Drosophila* sequences via the segregating sites algorithm (Fig. 3a) consistently suggest that



**Fig. 3.** Average posterior probability of assigning a query to each species group in the local database using the segregating sites algorithm; (a) with *D. angor* and (b) without *D. angor*. A grayscale ramp from white to black represents the average posterior probability assignment from 0.0 to 1.0 respectively. The origin of the query sequence is shown on the y-axis and the taxon for assignment is shown on the x-axis. Shadings off the main diagonal indicate posterior probabilities to incorrect taxon.

*D. angor* has a strong posterior probability for each and every one of the query sequences. Indeed, in many cases the posterior probability of an assignment to this group can be larger than that for the correct taxon. For example the average posterior probability for 11 *D. virilis* sequences is 0.4061 that they originated from a coalescent of the *D. angor* sequences and only 0.3908 that they originated from the coalescent formed by the remaining *D. virilis* sequences.

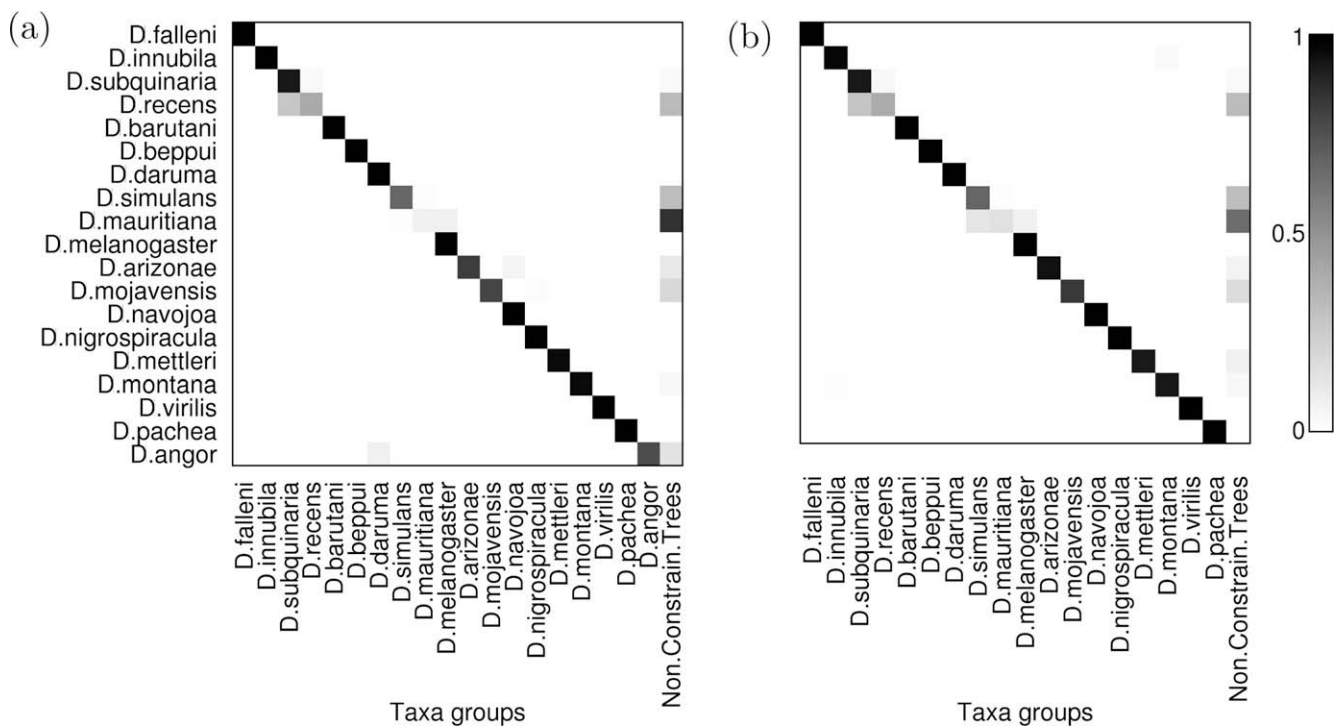
The *D. angor* sequences are an odd collection. The phylogeny shown in Fig. 2 suggests that these sequences branch polyphyletically throughout the tree. These 13 sequences form roughly five groups. The first group of four sequences are identical among themselves but differ from the others by 60–115 substitutions (within a length of 663 bp; a rather large level of intraspecific divergence). The second group of six sequences differ within the group by 2–46 substitutions. The third, fourth and fifth groups are each a single sequence that differs from every other *D. angor* sequence by 75–121, 81–119, 110–121 substitutions. That two sampled sequences from a single species should differ by fully a sixth of their nucleotides in a highly conserved sequence is unusual.

The effect of this on the assignments is to suggest that *D. angor* has a huge (and unrealistic) value of  $\theta$  and that the coalescent formed by the *D. angor* sequences can encompass any query. The potential addition of a query sequence to the *D. angor* group does not significantly alter the likelihood of the observed number of segregating sites. This is because only a comparatively few number of additional segregating sites are added with an already very large  $\theta$ . But since another entire sequence is added, the sample size has increased and, since the query is in the middle of this coalescent, the addition of another sequence with less variation actually improves the likelihood of the observation. This appears to be the cause of the high posterior probabilities of assignment to *D. angor* independent of the query sequence. To a lesser extent, this phenomenon also occurs with *D. subquinaria* since this taxon also has a large amount of sequence variation. To eliminate this effect the *D. angor* sequences were removed and the analysis redone as shown in Fig. 3b.

With the elimination of *D. angor*, most of the query sequences show the highest posterior probability to the taxon from which they originated. Missassignments occur most noticeably in three locations. The missassignment of *D. recens* to *D. subquinaria* (and to a lesser extent, the reverse), a symmetrical confusion between *D. arizonae* and *D. mojavensis*, and missassignments among *D. simulans* and *D. mauritiana*. The missassignments of *D. recens* sequences to the *D. subquinaria* species is because many of these sequences (82 from 2 distinct haplotypes) are identical to sequences labelled as originating from *D. subquinaria* (Shoemaker et al., 2004; Jaenike et al., 2006). The lack of resolution between the *D. arizonae* and *D. mojavensis* species is due to their sibling species status and recent divergence time (Reed et al., 2007). The distinction between *D. simulans* and *D. mauritiana* is even less clear due both to their shared haplotypes and recent divergence (Tamura et al., 2004).

The assignments by the SAP algorithm of query sequences to *Drosophila* species are shown in Fig. 4. This algorithm also had difficulty with the same group of taxa that the segregating sites algorithm had difficulty with. For the most part, these difficulties are not as apparent in the figure since a portion of the sampled trees from the MCMC do not match the given taxonomy from NCBI, termed here non-constrained trees. These trees, that do not match the NCBI annotated taxonomy, are classified separately. These trees represent an ambiguous component of the assignment.

The segregating sites algorithm spent roughly 3 s per assignment for the whole 616 sequence data set on a computer with a 1.6 GHz processor, running Linux. SAP spent roughly 8 min per assignment on the same system. A single assignment of a single query to the 42 sequences of *D. montana* using a coalescent assigner takes many hours to run and even then it is doubtful that it has reached stationarity. A single assignment to the 136 sequences of *D. recens* would take orders of magnitude longer. To complete the data set would require this to be repeated for each of the 616 queries. Hence it is not possible to provide comparable results for the coalescent assigner.



**Fig. 4.** Average posterior probability of assigning a query to each species using the SAP algorithm; (a) with *D. angor* and (b) without *D. angor*. A grayscale ramp from white to black represents a posterior probability assignment from 0.0 to 1.0 respectively. The origin of the query sequence is shown on the y-axis and the taxon for assignment is shown on the x-axis. Shadings off the main diagonal indicate posterior probabilities to incorrect taxon.

#### 4. Discussion

Barcoding involves the assignment of a sequence to a pre-existing taxonomic group. This is done using information drawn from a short DNA sequence, COI in many cases (*rbcl* and *matK* in the case of plants; Hollingsworth et al., 2009). The relationships of the sequences among the taxa contains information regarding their likelihood of being samples from a particular species. Unfortunately, when a collection of samples is first made, it is often difficult to determine their taxonomic species of origin. This is particularly the case if the group is little studied and has many sibling species. The *Drosophila* species have many sibling groups but have the advantage that the true species relationships are generally well known.

With the advent of better sequencing technologies, it is expected that the number of alternative species to which an assignment must be made will increase, consequently making the task of assignment more difficult. Thus, the performance of barcoding assignment methods, both in speed and accuracy, given increasing amounts of information, is important.

In general, a method to calculate the probability that an unknown sequence originated from a particular species, *x*, is desired. The segregating sites algorithm does not calculate this probability; rather it estimates the probability that the query sequence could originate from a coalescent implied from the knowledge of the current database. The segregating sites algorithm considers all of the sequences from each species. The data from the *D. angor* sequences illustrates this subtle difference. Similarly, the SAP program also does not determine the desired probability. Rather it estimates the probability that a sequence consistently branches next to a single member of species *x* given the current database. The sequences from *D. angor* do not generally alter these assignments.

The segregating sites algorithm, however, consistently suggest that for each query sequence, there is a significant probability that this sequence might have arisen from *D. angor*. The reason for this is that the given database and the given species identifications are assumed to be correct and, as such, given the huge amount of sequence divergence within the 'hypervariable' species *D. angor*, there is a very real possibility that any of these sequences might have originated from *D. angor*. Assuming that the given data is indeed accurate, this seems to be the correct answer. The taxonomic assignment of sequences to the species within the database (*D. angor*, for example) are assumed to be correct. This assumption is made at the species level for the segregating sites algorithm. It is similarly made for SAP at deeper taxonomic levels.

If the classification of the sequences of *D. angor* into a single species is correct then the segregating sites algorithm provides correct posterior probabilities. The further consideration of a risk measurement based on distances (which can be incorporated into a segregating sites algorithm) will however warn against over interpretation of the posterior probabilities. The presence of such a hypervariable species is also highlighted by the algorithm's results and suggests a possible alternate interpretation; that the species might be a candidate for further taxonomic scrutiny.

Even if an unknown query sequence is a perfect match to a sequence in a knowledge database, it does not imply that a perfect species identification has been achieved. Other species identifications might have a high or even a higher posterior probability. Therefore, given that a perfect match has been found in the database, this alone does not justify the conclusion that the species of origin has been identified.

The model based methods analyzed here capitalize on understanding the process governing the system under study and result in more informative and powerful tools to analyze sequence data generated from such systems. In applying any statistical method

it is important to understand the boundaries and limitations of its application. The application of the segregating sites algorithm and the SAP algorithm to *Drosophila* data illustrates well that they calculate posterior probabilities of somewhat different quantities. Which method is preferred and should be applied depends on which quantity is desired. The SAP algorithm measures where a sequence branches while the segregating sites algorithm measures if a sequence can 'fit' into an existing species.

The results presented indicate that both Bayesian methods work well to correctly identify species even in the absence of a 'barcode gap'. When uncertainty exists in the assignment, the methods correctly reflect and report this uncertainty. The degree of uncertainty in these methods is directly reflected in the accuracy of the taxonomic reconstructions.

The segregating sites algorithm is available at <http://info.mcmaster.ca/TheAssigner/>.

#### Acknowledgments

This research was funded by an NSERC discovery grant, NSERC/Genome Canada Barcode grants and an CRC award to GBG. ML is funded by NSERC/Genome Canada Barcode grants, grants from McMaster University and scholarships.

#### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.ympev.2010.01.002](https://doi.org/10.1016/j.ympev.2010.01.002).

#### References

- Abdo, Z., Golding, G.B., 2007. A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. *Syst. Biol.* 56, 44–56.
- Aliabadian, M., Kaboli, M., Nijman, V., Vences, M., 2009. Molecular identification of birds: performance of distance-based DNA barcoding in three genes to delimit parapatric species. *PLoS One* 4, e4119.
- Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., Veuille, M., Laredo, C., in press. DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics*.
- Ballard, J.W., 2000a. Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup. *J. Mol. Evol.* 51, 48–63.
- Ballard, J.W., 2000b. When one is not enough: introgression of mitochondrial DNA in *Drosophila*. *Mol. Biol. Evol.* 17, 1126–1130.
- Cognato, A.I., 2006. Standard percent DNA sequence difference for insects does not predict species boundaries. *J. Econ. Entomol.* 99, 1037–1045.
- Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340.
- Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.
- Frézal, L., Leblois, R., 2008. Four years of DNA barcoding: current advances and prospects. *Infect. Genet. Evol.* 8, 727–736.
- Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W., Hebert, P.D., 2006. DNA barcodes distinguish species of tropical *Lepidoptera*. *PNAS* 103, 968–971.
- Hebert, P.D., Cywinska, A., Ball, S.L., deWaard, J.R., 2003a. Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270, 313–321.
- Hebert, P.D., Ratnasingham, S., deWaard, J.R., 2003b. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. Biol. Sci.* 270, S96–S99.
- Hebert, P.D., Stoeckle, M.Y., Zemlak, T.S., Francis, C.M., 2004. Identification of birds through DNA barcodes. *PLoS Biol.* 2, e312.
- Hickerson, M.J., Meyer, C.P., Moritz, C.P., 2006. DNA barcoding will often fail to discover new animal species over broad parameter space. *Syst. Biol.* 55, 729–739.
- Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M.W., Cowan, R.S., Erickson, D.L., Fazekas, A.J., Graham, S.W., James, K.E., Kim, K.-J., Kress, W.J., Schneider, H., van AlphenStahl, J., Barrett, S.C.H., van den Berg, C., Bogarin, D., Burgess, K.S., Cameron, K.M., Carine, M., Chacn, J., Clark, A., Clarkson, J.J., Conrad, F., Devey, D.S., Ford, C.S., Hedderon, T.A.J., Hollingsworth, M.L., Husband, B.C., Kelly, L.J., Kesanaakurti, P.R., Kim, J.S., Kim, Y.-D., Lahaye, R., Lee, H.-L., Long, D.G., Madrin, S., Maurin, O., Meusnier, I., Newmaster, S.G., Park, C.-W., Percy, D.M., Petersen, G., Richardson, J.E., Salazar, G.A., Savolainen, V., Seberg, O., Wilkinson, M.J., Yi, D.-K., Little, D.P., 2009. A DNA barcode for land plants. *Proc. Natl. Acad. Sci. USA* 106, 12794–12797.

- Huang, D., Meier, R., Todd, P.A., Chou, L.M., 2008. Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. *J. Mol. Evol.* 66, 167–174.
- Hudson, R.R., Coyne, J.A., 2002. Mathematical consequences of the genealogical species concept. *Evolution* 56, 1557–1565.
- Jaenike, J., Dyer, K.A., Cornish, C., Minhas, M.S., 2006. Asymmetrical reinforcement and Wolbachia infection in *Drosophila*. *PLoS Biol.* 4 (10), e325.
- Kelly, J.K., Noor, M.A., 1996. Speciation by reinforcement: a model derived from studies of *Drosophila*. *Genetics* 143, 1485–1497.
- 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.
- Knowles, L.L., Carstens, B.C., 2007. Delimiting species without monophyletic gene trees. *Syst. Biol.* 56, 887–895.
- Meier, R., Zhang, G., Ali, F., 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the barcoding gap and leads to misidentification. *Syst. Biol.* 57, 809–813.
- Meyer, C.P., Paulay, G., 2005. DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol.* 3, e422.
- Moritz, C., Cicero, C., 2004. DNA barcoding: promise and pitfalls. *PLoS Biol.* 2 (10), 1529–1531.
- Munch, K., Boomsma, W., Huelsenbeck, J.P., Willerslev, E., Nielsen, R., 2008a. Statistical assignment of DNA sequences using Bayesian phylogenetics. *Syst. Biol.* 57, 750–757.
- Munch, K., Boomsma, W., Willerslev, E., Nielsen, R., 2008b. Fast phylogenetic DNA barcoding. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 3997–4002.
- Nielsen, R., Matz, M., 2006. Statistical approaches for DNA barcoding. *Syst. Biol.* 55, 162–169.
- Noor, M.A., 1995. Speciation driven by natural selection in *Drosophila*. *Nature* 375, 674–675.
- Pollard, D.A., Iyer, V.N., Moses, A.M., Eisen, M.B., 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* 2 (10), 1634–1647.
- Powell, J., 1997. *Progress and Prospects in Evolutionary Biology*. Oxford University Press, Inc., New York.
- Reed, L.K., Nyboer, M., Markow, T.A., 2007. Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. *Mol. Ecol.* 16, 1007–1022.
- Remigio, E.A., Hebert, P.D., 2003. Testing the utility of partial COI sequences for phylogenetic estimates of gastropod relationships. *Mol. Phylogenet. Evol.* 29, 641–647.
- Ross, H.A., Murugan, S., Li, W.L., 2008. Testing the reliability of genetic methods of species identification via simulation. *Syst. Biol.* 57, 216–230.
- Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.
- Satta, Y., Takahata, N., 1990. Evolution of *Drosophila* mitochondrial DNA and the history of the melanogaster subgroup. *Proc. Natl. Acad. Sci. USA* 87, 9558–9562.
- Schmidt, B.C., Sperling, F.A.H., 2008. Widespread decoupling of mtDNA variation and species integrity in *Grammia* tiger moths (Lepidoptera: Noctuidae). *Syst. Entomol.* 33, 613–634.
- Shoemaker, D.D., Dyer, K.A., Ahrens, M., McAbee, K., Jaenike, J., 2004. Decreased diversity but increased substitution rate in host mtDNA as a consequence of Wolbachia endosymbiont infection. *Genetics* 168, 2049–2058.
- Sparks, J.S., Smith, W.L., 2006. *Sicyopterus lagocephalus*: widespread species, species complex, or neither? A critique on the use of molecular data for species identification. *Mol. Phylogenet. Evol.* 40, 900–902.
- Tamura, K., Subramanian, S., Kumar, S., 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* 21, 36–44.
- Tavare, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* 26, 119–164.
- Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.
- Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R., Zhang, H., 2009. FlyBase: enhancing *Drosophila* gene ontology annotations. *Nucleic Acids Res.* 37, D555–D559.
- Wang, B.C., Park, J., Watabe, H.A., Gao, J.J., Xiangyu, J.G., Aotsuka, T., Chen, H.W., Zhang, Y.P., 2006. Molecular phylogeny of the *Drosophila virilis* section (Diptera: Drosophilidae) based on mitochondrial and nuclear sequences. *Mol. Phylogenet. Evol.* 40, 484–500.
- Wiemers, M., Fiedler, K., 2007. Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). *Front. Zool.* 4, 8.