MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 1864–1877 doi: 10.1111/j.1365-294X.2011.05239.x

ABGD, Automatic Barcode Gap Discovery for primary species delimitation

N. PUILLANDRE,* A. LAMBERT,† S. BROUILLET‡§ and G. ACHAZ‡§

**UMR 7138, Muséum National d'Histoire Naturelle, Departement Systématique et Evolution, 43, Rue Cuvier, 75231 Paris, France, †Laboratoire de Probabilités et Modèles Aléatoires (UMR 7599), UPMC Univ Paris 06, Univ Paris Diderot, CNRS, Paris, France, ‡Systématique, Adaptation et Evolution (UMR 7138), UPMC Univ Paris 06, CNRS, MNHN, IRD, Paris, France, §Atelier de Bioinformatique, UPMC Univ Paris 06, Paris, France*

Abstract

Within uncharacterized groups, DNA barcodes, short DNA sequences that are present in a wide range of species, can be used to assign organisms into species. We propose an automatic procedure that sorts the sequences into hypothetical species based on the barcode gap, which can be observed whenever the divergence among organisms belonging to the same species is smaller than divergence among organisms from different species. We use a range of prior intraspecific divergence to infer from the data a model-based one-sided confidence limit for intraspecific divergence. The method, called Automatic Barcode Gap Discovery (ABGD), then detects the barcode gap as the first significant gap beyond this limit and uses it to partition the data. Inference of the limit and gap detection are then recursively applied to previously obtained groups to get finer partitions until there is no further partitioning. Using six published data sets of metazoans, we show that ABGD is computationally efficient and performs well for standard prior maximum intraspecific divergences (a few per cent of divergence for the five data sets), except for one data set where less than three sequences per species were sampled. We further explore the theoretical limitations of ABGD through simulation of explicit speciation and population genetics scenarios. Our results emphasize in particular the sensitivity of the method to the presence of recent speciation events, via (unrealistically) high rates of speciation or large numbers of species. In conclusion, ABGD is fast, simple method to split a sequence alignment data set into candidate species that should be complemented with other evidence in an integrative taxonomic approach.

Keywords: DNA barcoding, integrative taxonomy, pairwise differences, speciation

Received 23 February 2011; revision received 4 June 2011; accepted 15 June 2011

Introduction

DNA barcodes, short DNA sequences that are present in a wide range of species and primarily used for taxonomic expertise, are now routinely used to identify, classify and analyse species diversity in many groups (http://www.barcodeoflife.org/content/resources/publications). The vast majority of barcoding studies, since Hebert *et al.* (2003), aim at testing the barcoding methodology, by first sequencing the Cytochrome Oxidase I (COI) gene (or other genes like *MatK* and *rbcl* for plants; CBOL Plant Working Group 2009) for a large number of individuals, and then by comparing the results obtained with previous knowledge of species boundaries (e.g. Armstrong & Ball 2005; Clare *et al.* 2006; Costa *et al.* 2007; Kerr *et al.* 2007; van Velzen *et al.* 2007; Rach *et al.* 2008). From previous studies, we know that DNA barcoding is efficient when intraspecific diversity for the COI gene is lower than the interspecific diversity, i.e. when COI sequences sampled within the same species are always more similar than sequences sampled from different species. Accordingly, DNA barcodes can be used as an identification tool, shortcutting the difficulties of a morphologically based identification (Stoeckle 2003; Blaxter 2004; Janzen *et al.* 2005), especially for environmental studies (Valentini

Correspondence: Guillaume Achaz, Fax: +33 1 44 27 63 12; E-mail: guillaume.achaz@upmc.fr

et al. 2009). In the presence of a reference data set with previously characterized species, the species of an organism can be automatically identified using its barcode sequence. The accuracy and power of several methods of assignation were recently assessed through simulations (Austerlitz *et al.* 2009).

DNA barcodes have also been put forward as an interesting tool to discover new species (Smith *et al.* 2005; De Salle 2006). Several new taxa that could potentially be new species were discovered. Specifically, several species, originally described by morphological or ecological characters, were including two or more groups of individuals that harbour very divergent COI sequences (Hebert *et al.* 2004; Campbell *et al.* 2008; Ståhls & Savolainen 2008; Locke *et al.* 2010). It is important to emphasize that when DNA barcoding suggests the existence of new species, it is not definitive proof and must be used along with other characters that make the species delimitation more reliable (De Salle 2006; Wiemers & Fiedler 2007). Indeed, COI gene can be affected by several biases and must be combined with, at least, the analysis of other independent genes, but also with morphological, geographical or ecological data to clearly delimit species in what is called an integrative framework (Dayrat 2005; Will *et al.* 2005; Ahrens *et al.* 2007; Miller 2007; Vogler & Monaghan 2007; Wiens 2007; Bond & Stockman 2008; Giraud *et al.* 2008; Dépraz *et al.* 2009; Damm *et al.* 2010; Goetze 2010; O'Meara 2010; Padial *et al.* 2010; Ross *et al.* 2010; Yeates *et al.* 2010).

DNA barcodes can also be used as an exploratory tool for unexplored groups. In this case, results obtained with DNA barcodes cannot be directly compared with other independent data (such as described species in existing literature). Instead, groups predicted from barcodes will be used as a first set of species hypotheses. The method we propose here is precisely designed for this purpose. We named it ABGD, an acronym for Automatic Barcode Gap Discovery.

A General Mixed Yule Coalescent (GMYC) model has been proposed to delimit species from single locus genetic data (Pons *et al.* 2006; Monaghan *et al.* 2009). Although grounded in a solid phylogenetic framework, this model heavily relies on the correctness of the Yule speciation model. Furthermore, it requires the phylogenetic reconstruction of all sequences in the data set, which is a very slow process for very large data sets and the model fit itself can be computationally intensive. Other methods based on Markov Chain Clustering (e.g. Zinger *et al.* 2009) are also used to build groups (named OTU, for Operational Taxonomic Units) but not specifically designed to delimit species, although Molecular Operational Taxonomic Units (MOTUs; Floyd *et al.* 2002) may overlap with the species, depending on the species definition. Numerous other methods also exist to delimit species based on DNA data, but they generally rely on a prior definition (e.g. based on morphological characters and geographical/ecological data; for review see, Sites & Marshall 2003).

Finally, several others methods based on multi-locus data sets have been proposed to delimit species without a priori knowledge (Knowles 2009; O'Meara 2010; Ross *et al.* 2010). However, sequencing multiple genes can be a laborious task, especially in largely unknown groups for which only few loci are generally characterized.

In the distribution of pairwise differences between all sequences of a typical barcode data set, one can observe a gap between intraspecific diversity and interspecific diversity; this gap has been named 'barcode gap'. Although several attempts have been made to establish a standard limit between intra- and inter-species divergence [e.g. 3% of divergence (Smith *et al.* 2005) or the 10× rule (Hebert *et al.* 2004)], none can be generalized to many groups of organisms (Fergusson 2002; Holland *et al.* 2004; Bichain *et al.* 2007; Gómez *et al.* 2007; Meier *et al.* 2008). Furthermore, as highlighted in several studies, intra- and interspecific distances frequently overlap, and visually defining a threshold becomes difficult (Meyer & Paulay 2005; Elias *et al.* 2007; Wiemers & Fiedler 2007; Smith *et al.* 2008). We propose here a method to automatically find the distance where the barcode gap is located, called Automatic Barcode Gap Discovery (ABGD). This method proposes a standard definition of the barcode gap and can be used even when the two distributions overlap to partition the data set into candidate species.

The data set is partitioned into the maximum number of groups (i.e. species) such that the distance between two sequences taken from distinct groups will always be larger than a given threshold distance (i.e. barcode gap). In the graph terminology, the sequences are nodes connected by edges if their distance is smaller than the threshold and the groups are the connected components of the graph. Naively applying this method requires (i) the knowledge of the threshold and (ii) the assumption that sequences belonging to the most closely related species have a greater divergence than the largest intraspecific divergence. Thus, the aim of the ABGD method is to (i) statistically infer the barcode gap from the data and to partition the data set accordingly, and (ii) recursively apply this procedure to the newly obtained groups of sequences, thereby allowing to work with data sets with multiple thresholds throughout taxa.

In this article, we first provide a complete description of the ABGD method. We then applied our method to both real and simulation data. Real data are necessary to accurately validate our method, but simulation data allow us to test the theoretical limitations of the method in a controlled scenario. For this reason, we believe that simulated data provide an interesting opportunity for testing methods based on the barcode gap strategy. For the sake of clarity and brevity, we leave for a future study the systematic comparison of all methods of species delineation based on genetics data and therefore only discuss qualitatively the benefits and drawbacks of ABGD.

Methods

Automatic Barcode Gap Discovery

Here, a data set is a set of n sequences, divided into an unknown number n_s of unknown species. The distribution of pairwise differences between the n sequences of a data set typically shows a gap when the mode of the distribution of intraspecific divergence is lower than the mode(s) of interspecific divergence. The ABGD method aims at (i) finding automatically the gap that divides the distribution between the left most significant mode and the other one(s); (ii) applying recursively this operation to get the finest partition of the data set into candidate species. Unlike other methods that split distributions (e.g. k-means; MacQueen 1967), this method does not rely on any specific properties of the distribution (e.g. variances).

The outline of ABGD is the following: (i) It finds the first barcode gap that occurs at a distance larger than some value $\text{dist}_{\text{limit}}$, a limit under which distances are statistically more likely to be intraspecific. $\text{dist}_{\text{limit}}$ is a simple function of the population mutation rate, estimated from the data set. It is estimated on a preliminary partition of the data set with a threshold P given by the user (P is the prior maximum divergence of intraspecific diversity). (ii) Taking a threshold equal to the barcode gap computed in step (i), it computes a so-called primary partition, where groups are the first candidate species. (iii) To account for mutation rate variability across taxa and overlap of intra and interspecific diversities, ABGD is only completed after recursive application of these first two steps to each cluster of the primary partition. This recursion splits the primary partition into secondary partitions, and so on until no further splittings occur.

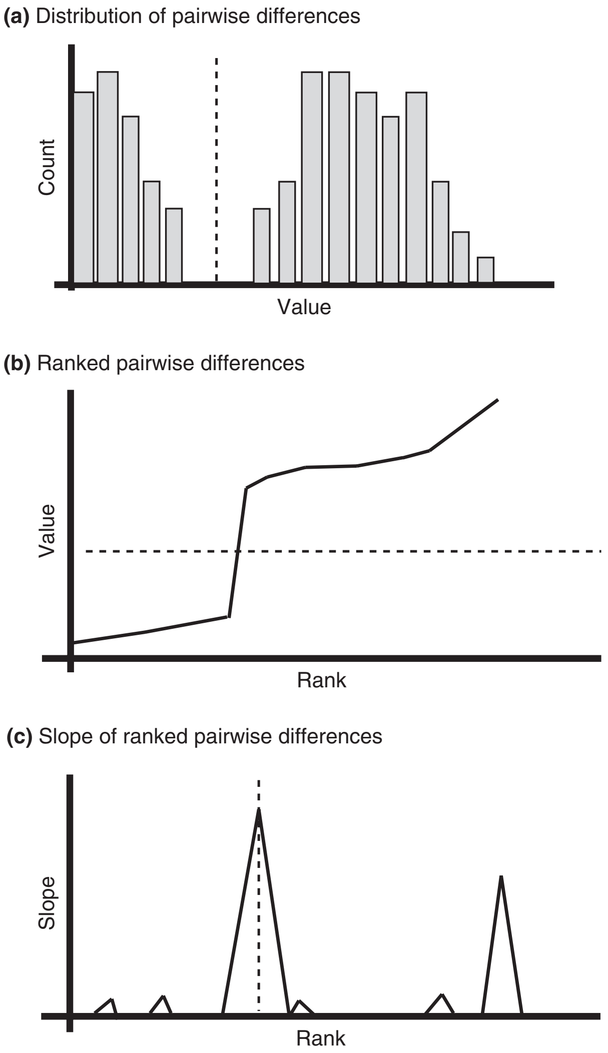
(a) Distribution of pairwise differences
Count
Value
(b) Ranked pairwise differences
Value
Rank
(c) Slope of ranked pairwise differences
Slope
Rank

Fig. 1 Schematic illustration of Automatic Barcode Gap Discovery (ABGD). (a) A hypothetical distribution of pairwise differences. This distribution exhibits two modes. Low divergence being presumable intraspecific divergence, whereas higher divergence represents interspecific divergence. (b) The same data can be represented as ranked ordered values. (c) Slope of the ranked ordered values. There is a sudden increase in slopes in the vicinity of the barcode gap. The ABGD method automatically finds the first statistically significant peak in the slopes.

Detecting gaps. Our method for detecting gaps is depicted in Fig. 1 : (i) All pairwise distances (i.e. number of differences, potentially corrected for multiple substitutions using distance models) are ranked by increasing values ($d_1 \leq d_2 \leq \cdots \leq d_p$, where $p = n(n-1)/2$ is the number of pairwise distances in the data set of n sequences); (ii) A local slope function is computed for a given window size w as: $s_{r,w} = (d_{r+w} - d_r)/w$ with $r \in [1,p-w]$; (iii) The method detects peaks of slope values (corresponding to gaps in the initial distribution) and reports the distance performing the local maximum as the exact value of the gap.

Obviously, $s_{r,w}$ depends on w and there is no easy choice for a single w value. Therefore, we computed increasing values of w, that increase by one tenth of its starting value (eg. 100, 110, 120, …) and considered that when the reported gap is identical for three successive

w, it is stable and therefore pertinent. It is noteworthy to mention that modifications of this arbitrary rule do typically not change the results. We set the default starting w to $p/10$, with 1 as a minimum and 1000 as a maximum.

Gaps within a single species. Here, we consider the reference case of one single species to compute the threshold value $\text{dist}_{\text{limit}}$, under which distances are statistically more likely to be intraspecific.

If we assume that a species is a single non-structured population, the distribution of pairwise differences from a sample of sequences can be retrieved using coalescent models (Tajima 1983; Slatkin & Hudson 1991). It is well documented that this distribution is sometimes multimodal (Slatkin & Hudson 1991). This is actually one feature of the distribution that has been put forward to estimate population growth (Rogers & Harpending 1992) or to build up a test of neutrality (Harpending *et al.* 1993), although because of the large variance of the gap widths, it is of limited use because it has strong bias for estimators and low power for statistics (Ramos-Onsins & Rozas 2002; Rosenberg & Hirsh 2003). Here, we characterized the distribution of the distance corresponding to the largest gap one could observe in a single panmictic population.

To do so, we ran standard coalescent simulations with a single species and, in each run, computed the distance that corresponds to the largest gap reported by the previously described method for detecting gaps. From 10^4 simulations, we estimated the upper boundary of a one-tail 95% confidence interval of this distance ($\text{dist}_{\text{limit}}$). In a given data set, if the distance corresponding to a gap exceeds $\text{dist}_{\text{limit}}$, we can exclude (with a risk of 5%) that this gap is within intraspecific divergence.

Standard coalescent simulations here take two parameters: n, the number of sequences in the sample and θ the population mutation rate. θ is the individual mutation rate μ per generation multiplied by twice the effective population size (i.e. $\theta = 2N_e\mu$). Further, the previously described gap detecting method requires a window size w. Therefore, the simulation process requires overall three parameters: n, θ and w.

Simulations (not shown) demonstrate that $\text{dist}_{\text{limit}}$ is independent of w and of n (as soon as $n > 10$). However, it increases with θ and shows a perfect fit to a linear relationship $\text{dist}_{\text{limit}} = a\theta$ where $a = 2.581$ (data not shown). As a consequence, the only mandatory value to assign the significance of a gap is θ.

From a single species to the whole data set. Although the only required value to compute $\text{dist}_{\text{limit}}$ for a given data set is θ, it is unfortunately often unknown and has to be estimated from the data. In the case of a single, non-structured species, we can use a well-known unbiased estimator of θ, noted $\hat{\theta_\pi}$, that is equal to the average pairwise differences (Tajima 1983). On the contrary, estimation of θ from a sample of sequences with an unknown number of species is a very difficult problem. We chose to use a prior limit to intraspecific diversity (P) to compute an estimator of θ in the general case. Based on this prior limit, we computed $\hat{\theta}_{\text{prior}}$, which is the average pairwise of all distances that are below P. Because this estimator is an average, it is only moderately sensitive to outliers ; therefore, even approximately good prior limit leads to estimators that are similar to the estimator one would get using the true limit.

Because of the large variance of our estimator (Tajima 1983), we assume that the true θ could be as large as twice the estimator and define $\text{dist}_{\text{limit}} = 2a\hat{\theta}_{\text{prior}}$. Then, to only capture 'large' gaps, the barcode gap is chosen as the first local maximum slope (of ranked distances) occurring after $\text{dist}_{\text{limit}}$ and X times larger than any gap in the prior intraspecific divergence.

Therefore, our method uses two user input values: P, a prior limit to intraspecific diversity and X, a proxy for the minimum gap width. P give approximate indications on the area where the barcod gap should be detected and X relates to the sensitivity of the method to gap width. By default, the method set $X = 1.5$ and assess the impact of P by reporting results from $P = 0.001$ to $P = 0.1$ (see Results).

Building the partitions. Once a barcode gap is computed, we partition the data set into groups of sequences, i.e. candidate species. Groups are chosen so that the distance between sequences from different groups is always larger than the gap distance, and for each sequence of each group, there is at least one other sequence in the group at a distance smaller than the gap distance. This primary partition of the data set assumes that a single gap can be defined for the entire data set. However, it is very likely that the gap distance differs for groups within the data set. Therefore, we chose to re-apply the same partition method to each group of the primary partition to build a secondary partition, which itself could potentially be further divided. The method is then applied recursively to the newly formed groups until no more split is made. Importantly, the same prior intraspecific divergence P is used, but the starting window size w may be resized as the groups get smaller and smaller.

Simulations

We generated artificial distributions of pairwise differences using simulations generating gene genealogies evolving inside phylogenetic trees. There are three main steps in our simulation process : (i) generation of a species tree, (ii) generation of a sequence tree within the species tree and (iii) addition of mutations on the sequence tree. All populations are assumed to currently have size N and time is always expressed in N generations. When only one species is considered, our simulation reduces to a standard coalescent with superimposed mutations as described in Hudson (1990).

Speciation tree. We implemented four different models of speciation: a radiation model, a Moran model, a Yule model and a critical model. They are all characterized by a single parameter b. In the radiation model, a single event of speciation occurred in the past and all species radiated at the same time; the radiation event is exponentially distributed with parameter b. In the Moran model, the number of species we follow is n_s and is held constant through time by replenishing instantaneously every species becoming extinct (Hubbell 2001; Durrett 2008). More specifically, each species, independently of the others, gives rise at constant rate b/n_s to a new daughter species, and to keep the species number constant, kills simultaneously a randomly chosen species. The resulting species tree is a Kingman coalescent tree (Kingman 1982) with rate $2b/n_s$. In other words, tracing the history of a group of k species backwards in time, its number can only decrease by 1, and it does so at rate $bk(k-1)/n_s$. Last, in both the Yule and the critical models, the species trees result from a branching process, in which species become extinct at rate d and speciate at rate b, independently. In particular, a group of species counting k species is incremented by 1 (speciation) at rate bk and decremented by 1 (extinction) at rate dk. Several properties of genealogical trees generated by such processes can be found in Lambert (2008). In the Yule model, $d = 0$, so that species never become extinct. In the critical model, $d = b$, so the size of a group of species remains constant in mean. The genealogy of a group of extant species can be described in such a manner that all coalescence times are independently and identically distributed (Rannala 1997; Lambert 2009). In the Yule model, the probability density of coalescence times is exponential with parameter b: $f_{\text{coal}}(t) = be^{-bt}$. In the critical model, the probability density of coalescence times is a Cauchy distribution: $f_{\text{coal}}(t) = b(1 + bt)^{-2}$. This density function decreases much more slowly than an exponential function (heavy tail), and has infinite mean.

Sequence tree. Sequence trees were generated using the Kingman coalescent. In an isolated population, the Kingman coalescent describes the dynamics of lineages of sequences as time goes backwards. The number of lineages decreases by 1 each time a common ancestor to two lineages is found; at each such time, called coalescence time, the pair which coalesces is chosen uniformly among all possible pairs; the waiting time between k and $k-1$ lineages is an exponential random variable with parameter $k(k-1)/2$. Equivalently, an exponential 'clock' with rate 1 is attached to each possible pair of lineages, and when the first clock rings, the concerned pair collapses into one lineage. The rule is modified for our purpose by attaching a clock only to those pairs of lineages lying in the same species, that is, by merely forbidding coalescence of lineages lying in different species. Importantly, two lineages starting within the same species do not necessarily coalesce before coalescence of their species with a sister species; their coalescence within the mother species can even be longer than the coalescence of one of them with a third lineage originating from a sister species (incomplete lineage sorting). An example of such algorithm is given in Simonsen *et al.* (1995).

Mutations. Once the sequence tree is constructed, the mutations are distributed assuming a Poisson molecular clock, with rate $\theta/2$, where $\theta = 2N\mu$ (all populations have the same size, N; μ is per capita, per generation mutation probability).

Results

The ABGD web-interface as well as a command-line program are available at: http://wwwabi.snv.jussieu.fr/public/abgd/ (sources are also provided).

Performance on real data

We apply the ABGD method on six chosen barcode data sets selected from previous analyses that cover diverse groups of the metazoans (Table 1). The size n of the data set varies from 334 to 2574 sequences (and therefore from $n(n-1)/2 =$ 55 611 to 3 308 877 pairwise distances). Each data set is associated with a published article to a barcode analysis (Hajibabaei *et al.* 2006; Pons *et al.* 2006; Kerr *et al.* 2007; Wiemers & Fiedler 2007; Elias-Gutierrez *et al.* 2008; Smith *et al.* 2008), where groups defined from DNA sequences (usually the COI gene) are compared with species hypotheses based on independent data, with good congruency overall. Conclusions are typically based on the overall congruency between COI genetic variation and species hypotheses as traditionally defined in the literature for these taxa. However, these data sets differ in several aspects. First, they correspond to different taxa among the metazoans (i.e. insects, crustaceans and vertebrates).

Table 1 Barcode data sets used in this study

Taxon	Number of sequences	Number of groups in References	Publication
Amphibian	339	39	Smith *et al.* (2008)
Bird	2574	643	Kerr *et al.* (2007)
Cladocera	335	58	Elias-Gutierrez *et al.* (2008)
Agrodiaetus	334	114	Wiemers & Fiedler (2007)
Rivacindela	386*	46	Pons *et al.* (2006)
Sphingidae	989	107	Hajibabaei *et al.* (2006)

*All three loci were concatenated, trimmed to a 1621 bp conserved block; 86 sequences with more than 300 missing characters (i.e. gaps) were then removed to avoid biases in distance estimations.

Second, they include a variable number of replicates within each species and a different number of species. Finally, the efficiency of DNA barcodes based on the COI gene has been debated in some cases (see Vences *et al.* 2005).

Automatic Barcode Gap Discovery takes as input either a sequence alignment or a pairwise distance matrix. For our purposes, sequences were aligned and then used to compute a matrix of pairwise distances using the Kimura two parameters model (Kimura 1980). The shape of the pairwise distance distribution greatly varies between the different data sets, some having a clearcut barcode gap, others not (Fig. 2).

As a first result, we would like to emphasize that ABGD method is extremely efficient in computation time. Starting from a distance matrix (that takes few minutes of computation time to build), it takes only two seconds to find the recursive partition (for a given prior limit) in the largest data set on a laptop (Intel 2.8 GHz, MacOSX). This is much faster, by several orders of magnitude, than any other method proposed so far, especially methods that require the reconstruction of a phylogenetic tree.

One critical parameter of the ABGD method is the prior maximum divergence of intraspecific diversity (P). Intuitively, if this parameter is set too high, the whole data set will be considered as a single species and on the contrary if it set too low, only identical sequences will be considered as part of the same species. On the six data sets, we ran the ABGD method with a prior P that ranges from 0.001 to 0.12 (Fig. 3). Results show that as expected the number of groups ranges from 1 (generally when $P = 0.1$) to a large number of groups that correspond to groups of identical sequences (generally when $P = 0.001$). Results also show that typically recursive partitions have more groups than primary ones (which is expected because the former derive from the latter). This illustrates the benefit of the recursive partitioning. Most importantly, results show that, for four data sets, the number of groups is very close (sometimes equal) to the number of species defined by the authors in their original study when $P = 0.01$. Not only the number of groups matches but also their composition (data not shown). When their numbers differ, some closely related species are merged into a single group or some species with a large diversity are split into a few groups. For the *Rivacindella* group, species were predicted using the GMYC method (Pons *et al.* 2006) that predicts more groups than ABGD. A biological discussion about the number of distinct species in this data set deserves further discussions that are beyond the scope of this article. Finally, for the *Agrodiaetus* data set [the 'restricted' data set from Wiemers & Fiedler (2007)], we were not able to identify a barcode gap. This is a consequence of the small number of sequences per species (332 sequences for 114 species: ~2.9 sequences per species). Simulations show that ABGD works when there are more than 3–5 sequences per species (data not shown).

Interestingly, although recursive partitions are expected to handle better heterogeneities in the data set, primary partitions are typically stable on a wider range of prior values and are usually close to the number of groups described by taxonomists. We therefore decided to report both primary and recursive partitions in the output of ABGD.

Theoretical limitations

We have investigated the theoretical limitation of the ABGD method and more generally of the methods that predict species based on the barcode gap. As one could expect, the number n_s of species in the sample has a large impact on the results. In our simulations, we built a species tree using one of the four models of speciation, then randomly assign the sequences to the species and reconstruct their genealogy using coalescent models and finally add mutations on the sequence tree assuming a neutral molecular clock. Then, we compute the distribution of the pairwise differences among all sequences and apply the ABGD method including recursive partition of sequences into groups.

As we are interested here in observing the theoretical limitation of the method, we used an excellent prior on the limit between intra- and interspecific divergence (P). This would mimic a situation where we already know approximatively the diversity among the species. Indeed, intraspecific divergence is bounded by the Kingman coalescent model, in which the time to the most recent common ancestor has a 95% probability of

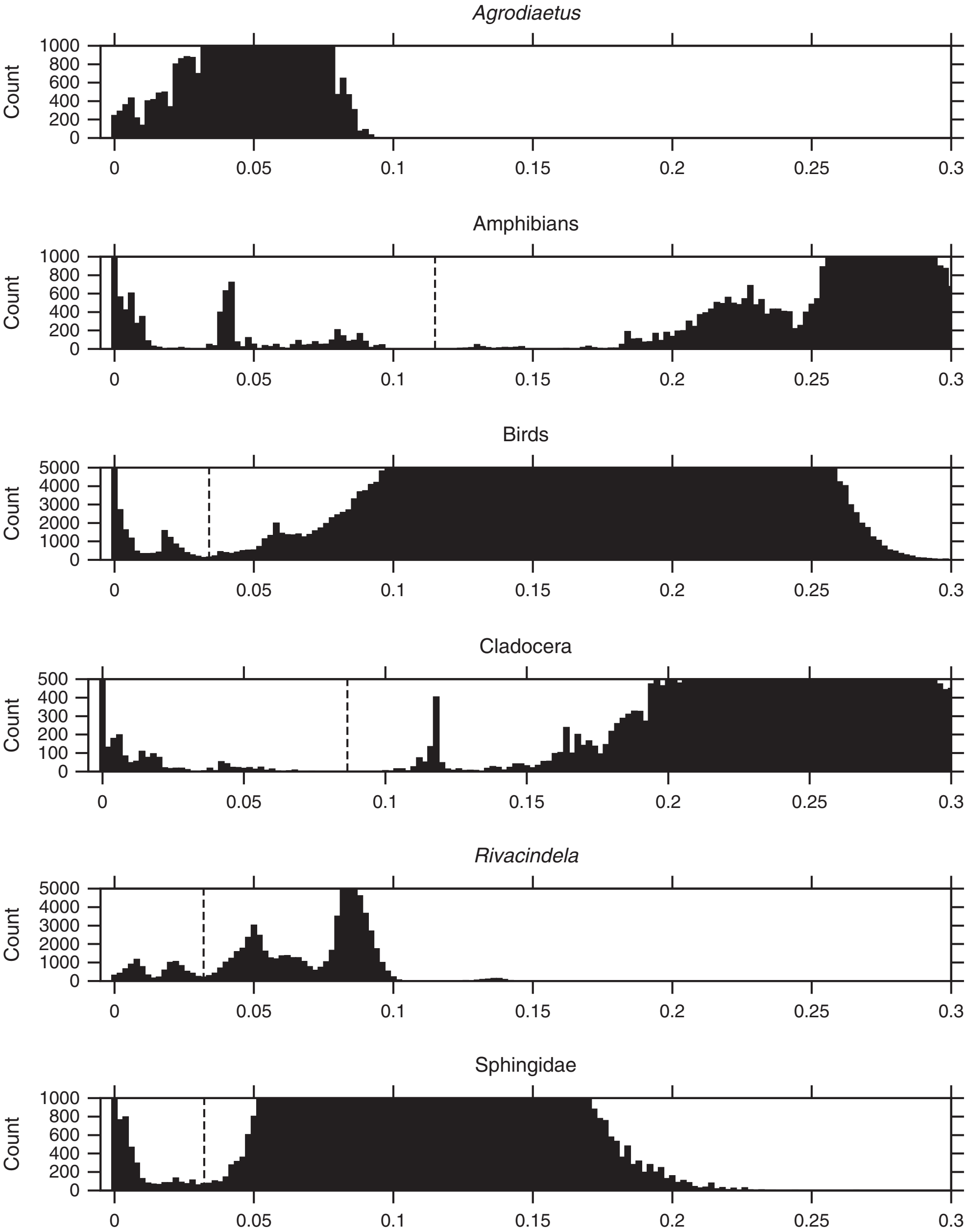

Fig. 2 Pairwise distance distribution of the six data sets. Distribution of the pairwise distances between sequences in five data sets. Saturation of the divergence is corrected using a Kimura two parameters model (transition and transversion rates). We also report using dashed lines the distance that is estimated using a prior of $P = 0.005$ of divergence. This figure illustrates, on the one hand, how variable can barcode data sets be and, on the other hand, that the distance corresponding to the primary partition differs from the prior distance.

being lower than approximatively $3.95 \times N$ generations, where N is the species population size. Therefore, for a given mutation rate, there is a bounded number of mutations between two sequences of the same species. In our simulations, we used a (rescaled) mutation rate of $\theta = 10$ and chose to set $P = 50$. The probability to observe more than 50 mutations is approximatively 0.02 (for a sample size larger than 10–15).

We evaluated the performance of the method, by reporting the number of species successfully delimited. For a given run, it ranges from 0 (none) to n_s (all). A species is considered as successfully delimited when all its members belongs to the same predicted group and no other sequences were added to it. This criterion is very conservative and therefore makes the interpretation of the results straightforward.

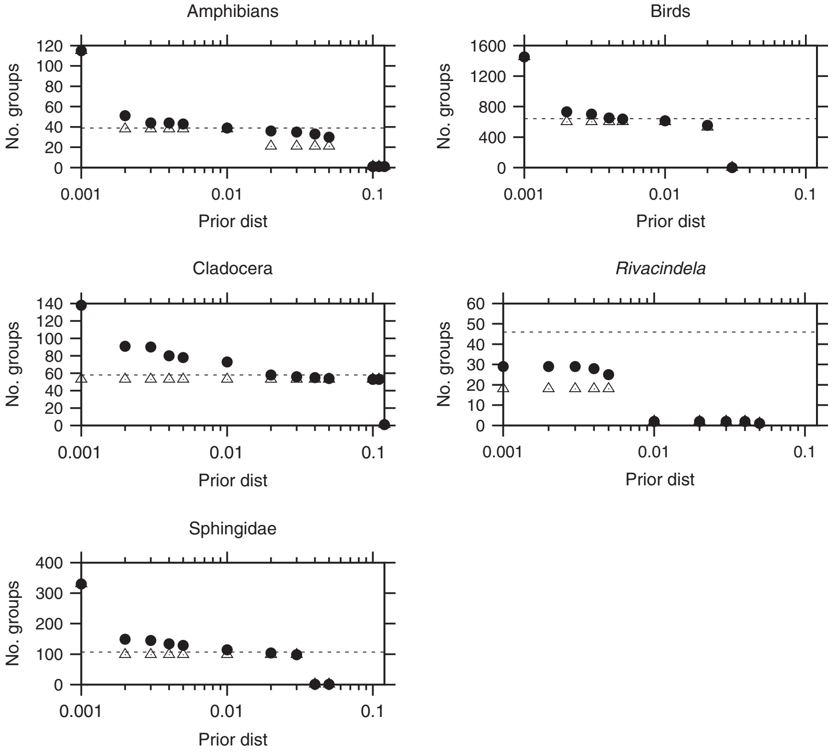

Fig. 3 Automatic partition of five data sets. This is a typical output from Automatic Barcode Gap Discovery (ABGD). We report the number of groups inside the partitions (primary and recursive) as a function of the prior limit between intra- and interspecies divergence. We also report the number of groups from the original studies. This figure illustrates that for a prior of 1–3%, the groups automatically defined by ABGD match the groups defined by the authors of the original studies.

Two species. We first assess the performance of ABGD with 50 sequences randomly assigned to two species. In this particular case, three models of speciation (the Moran, the Yule and the radiation models) are equal; only the critical model differs.

We ran the simulations using a variable speciation rate b that ranges from 0.001 to 10. When $b = 1$, the speciation and intraspecies coalescence events occur on the same time scale. As b gets smaller, the speciation times get longer (the speciation tempo is slower). Results (Fig. 4) show that the lower the speciation rate, the better the performance of the method. When b is very low ($b \ll 1$), the most recent speciation event is much older than the times to most recent common ancestor within species. When $b = 1$, the barcode gap vanishes: there are no more differences between intraspecific diver-

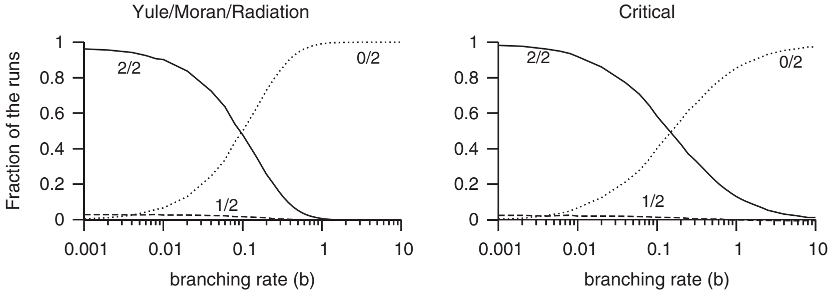

Fig. 4 Automatic Barcode Gap Discovery (ABGD) performance for two species. The fraction of the runs where 0, 1 or 2 species out of 2 are reported as a function of the branching rate (b) in the species tree. The following parameters were used: $\theta = 10$, $n_s = 2$, $n = 50$ and 10^4 replicates. When b is in the vicinity of 1, both the species and the intra-population tree have similar rates of coalescence. When $b \ll 1$, the speciation time is typically much longer than the total population coalescence time ; in this case, the genetic difference among and within species are well sorted apart. For this case, the Moran, the Yule and the Radiation Model are identical. The ABGD method almost never delimits only one species out of two.

gence and interspecific divergence. Interestingly, in the critical model, because the Cauchy distribution has a heavier tail than an exponential, speciation events tends to be older for the same *b*. This translates into a better performance of the ABGD for the same *b* value in the critical model than in the other models.

Importantly, we would like to mention that, in each run, the sequences were assigned to one of the two species randomly with equal probability. However, if the sampling is forced to consist in 49 sequences in one species and 1 in the other, results are identical (data not shown). This shows that the method is insensitive to the allocation of the sequences into species.

One of the most striking features of Fig. 4 is the almost absence of cases where only one of the two species is correctly assigned. Either 2/2 species or 0/2 are correctly delimited. When the detected gap distance corresponds to the true barcode gap, we expect 2/2 successfully delimited species. If the use of the detected gap distance were over-splitting one species, we would observe 1/2 successfully delimited species. However, because we filter out potential gaps that are within species, cases of this sort do not exceed a few per cent of the runs. Finally, when a single group is found (no split), we observe no (0/2) successfully delimited species. This last case systematically happens when $b > 1$, where divergence among and within species cannot be sorted apart.

Five species. We further tested the ABGD method using 50 sequences randomly assigned to 5 species. Results (Fig. 5) show that, although results under the four models of speciation are quantitatively different, results under the Moran, the Yule and the critical models share qualitative similarities. The ABGD method under radiation model shows however very different results. Under a radiation model, we observe either all (5/5) or no (0/5) species correctly delimited. Because there is a single event of speciation, when the speciation event is old enough, all species are found. For the Moran, Yule and critical models, we observe all cases but the 4/5 ones (that do not exceed a few per cent). Intermediate success (i.e. 1/5, 2/5, 3/5) that corresponds to the merging of different species into a single group are observed only for intermediate values of *b*. Indeed, for intermediate *b*, some speciation events are younger than times to most recent common ancestors within species while others are older. As $b \rightarrow 0$, all speciation events are older than those times and all species are found (5/5).

Impact of n_s. To measure more precisely the impact of n_s, we ran simulations with a fixed species branching rate ($b = 0.03$) but a variable number of species. We first studied the probability distribution of the number of correctly delimited species, with 250 sequences randomly assigned to 25 species (Fig. 6). Since the Moran, the Yule and the critical show qualitative similar results, we chose to report results for the Yule and the radiation models only. For the radiation model, 21.9% of the runs yield to 0/25 and 57.5% of them are 25/25. What remains is mainly 24/25 (17.8%) and 23/25 (2.5%). Cases like 24/25 correspond to the scenario

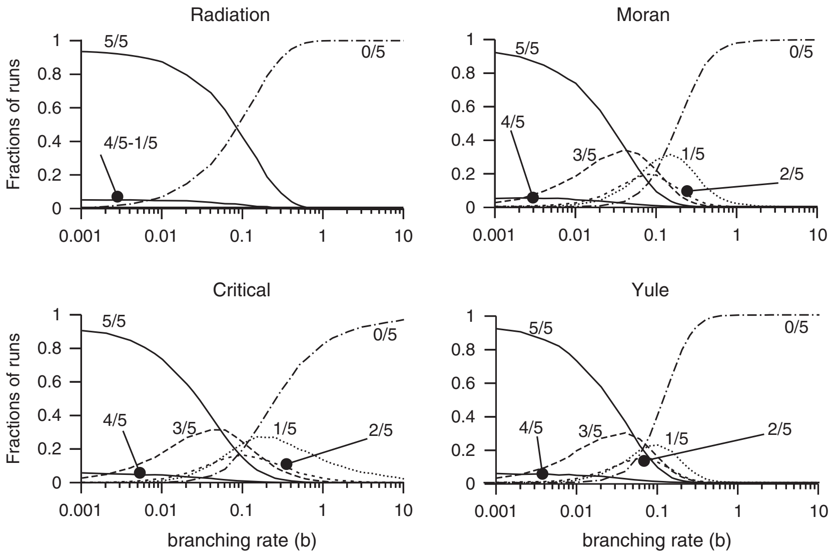

Fig. 5 Automatic Barcode Gap Discovery performance for five species. The fraction of the runs where 0, 1, 2, 3, 4 or 5 species out of 5 are reported as a function of the branching rate (*b*) in the species tree. The following parameters were used: $\theta = 10$, $n_s = 5$, $n = 50$ and 10^4 replicates.

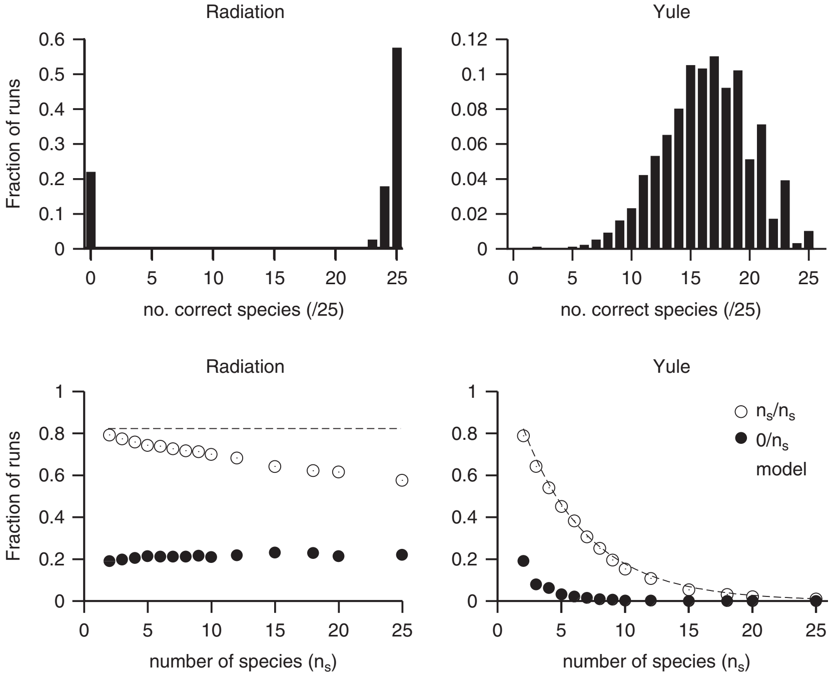

Fig. 6 Automatic Barcode Gap Discovery performance for n_s species. On the top panel, distribution of the number of correctly delimited species for 250 sequences divided into 25 species. Results are given for the Radiation and the Yule model ($\theta = 10$, $b = 0.03$). On the bottom panel, we report the fraction of runs that result in $0/n_s$ (complete failure) or n_s/n_s (complete success) for both the radiation and the Yule model. We also report the theoretical expectation for the n_s/n_s as dashed lines (setting the largest intraspecific divergence to $k = 65$ mutations). This figure overall illustrates the impact of the number of species in both types of models.

where one of the 25 species is over-split. Although this probability is very low for each species, the chances that one of many is oversplit becomes higher. The probability distribution of species successfully delimited is very different for the Yule model. Indeed, the number of successfully delimited species truly ranges from 0/25 to 25/25. The mode of the distribution is approximatively 15/25, which can be compared with 3/5 of Fig. 5.

Intuitively, for a given speciation rate, when the number of species increases, there is a higher chance that at least one of the speciation events is younger than common ancestors within species. This implies that the chances of correctly delimiting n_s/n_s species will decrease with the number of species. The same argument holds for the chances of finding no species ($0/n_s$). To assess the exact influence of the number of species on the chance of assigning correctly all species (n_s/n_s) or none ($0/n_s$), we ran simulations with an increasing number of species (from 2 to 25) with a total number of sequences of $n = 10 \times n_s$. As expected, results show that both curves (n_s/n_s and $0/n_s$) decrease strongly as the number of species increases. This suggests that in a large data set, the chances that all species are genetically differentiated becomes smaller and smaller as the number of species increases. Therefore, the predicted number of species is likely to be underestimated.

Modelling the impact of n_s. The ABGD method detects a difference between the intraspecific divergence and the interspecific divergence. To observe n_s/n_s correctly delimited species, sequences belonging to the two most closely related species should be separated by a number K of mutations greater than the largest intraspecific distance. In the Yule, Moran and radiation models, simple analytical expressions exist for the distribution of K. In these three models, the most recent speciation event is exponentially distributed; with rate $r = b(n_s - 1)$ for the Yule and the Moran models and with rate $r = b$ for the radiation model. Therefore, the number of mutations K separating two sequences belonging to the most closely related species has $P(K = k) = \int \text{Poisson}(k;\theta t)\text{Exp}(t;r)\text{d}t$, a geometric probability distribution with parameter $\theta/(r + \theta)$ (e.g. Tajima 1983). Therefore, we have:

$$P_{\text{radiation}}(K > k) = \left(\frac{\theta}{b+\theta}\right)^k \quad (1)$$

$$P_{\text{Moran/Yule}}(K > k) = \left(\frac{\theta}{b(n_s-1)+\theta}\right)^k \quad (2)$$

If k is close enough to the maximum intraspecific divergence (depending only on θ), then $P(K > k)$ can be taken as an approximation of the probability of delimiting correctly n_s/n_s species, and the preceding equations can be used to predict how well the method will perform for given b and θ.

Empirically, using $k = 65$ for $\theta = 10$ gives a good fit (Fig. 6), at least for the case of the Moran and Yule models. Importantly, eqn (1) predicts that, under the radiation model, the probability of delimiting all species will not be affected by the number of species. However, results show that this probability slightly decreases with n_s. This is mainly due to the fact that the probability of oversplitting one species increases in this particular case with n_s. Indeed, the preceding equations ignore the impact of oversplitting that would also make the n_s/n_s correctly delimited species less likely (although it is expected to be relatively rare in typical cases).

Discussion

We have introduced a new method, ABGD, to automatically formulate species hypotheses. Our method is meant to be used as a tool to detects a gap in the distribution of pairwise differences. Given the promising results and efficient run time, we suggest it to be used instead of any visual barcode gap definition that is less reliable because of its dependence on the bin size of the distribution and on some arbitrary, and not reproducible, decision that has to be made. The only input parameter that has to be set is the approximate maximum prior intraspecific distance, P. Importantly enough, this value needs not be defined precisely as the partitions are stable over a wide range of prior value and as several values are tested. We have repeatedly tried to avoid the use of this prior knowledge but were not able to do so. We would like to mention that the necessity of this prior knowledge could be dropped in theory, if a reliable estimator of θ could be computed from genetic data with an unknown number of species. Until this is possible, we suspect that it will be difficult to avoid the oversplitting of species into several groups.

Automatic Barcode Gap Discovery proposes the grouping of the input sequences into several hypothetical species by the sole use of pairwise differences (i.e. a distance matrix). As any method based on pairwise distances, it does not rely on an underlying genealogical tree. Although no ancestral states are inferred, this is very likely to be a benefit instead of a drawback as the method can a priori be used with nuclear sequences that have experienced recombination. Indeed, the tree representation is inadequate in the presence of recombination as ancestry is no longer represented by a binary tree but instead by an acyclic oriented graph, named the 'ancestral recombination graph' (Griffths & Marjoram 1997), that is, in practice, very difficult to reconstruct. Our method does not rely on tree shapes but on divergence, which, in case of recombination, are averages of neighbouring sequences with different genealogies. Furthermore, because we ignore the underlying genealogy, this method does not rely on properties of internal nodes of the species tree. Indeed, the method works when speciations are radiations, bifurcating events or even a mixture of both.

We have demonstrated that ABGD performs well on several large barcode data sets with previously hypothesized species. This however requires an appropriate prior of maximal intraspecific divergence. For the data set we have studied, this prior lies between 1% and 3% of divergence. In a few cases, ABGD founds multiple species hypotheses (e.g. one species split into two, or several species merged into a single one), but most of these cases were identified by the authors as problematic cases, where the COI gene was not totally congruent with the previously defined species boundaries, generally corresponding to species complexes. We would like to mention that although a 1–3% is potentially a reasonable default value for metazoans, it may well be inappropriate for other taxa such as bacterial or viral species which can harbour larger genetic diversity (and therefore larger estimated θ). Furthermore, intraspecific genetic diversity (and therefore θ) could vary from species to species. To minimize the number of false positive (oversplit species), one should use an estimate of θ for the most divergent species. The use of the recursion steps of partition should allow for further split within groups of smaller diversity, and thus limit the number of false negative (merged species). To explore the impact of P on the partitions, results computed on real data sets are by default reported for a range of $P \in [0.001,0.1]$. Typically, only few gaps have the potential to be considered as barcod gaps and the number of the partitions increases by discrete jumps as P gets smaller (see Fig. 3, where this is particularly striking for primary partitions).

However, when no barcode gap exists in a data set, ABGD cannot propose a primary partition and therefore is not suited for species delimitation. On the examples we have been working so far (e.g. in the *Agrodiaetus* example), this happens when the number of sequences per species is too small (i.e. <3–5).

Using controlled scenarios, we were able to show that both the speciation rate (b) and the number of species (n_s) within a data set are crucial parameters for the method to work. More generally, any method that detects a gap between intra- and interspecific variability will be strongly impacted by both b and n_s. The first caveat is the relative age of the speciation events to ancestry within populations. When the speciation events are too recent relative to population anctestry, there is no possibility of using genetic data to infer them. In terms of the model, when the speciation rate rises to values comparable to the speed of genetic drift, there is no more genetic differentiation between species. In our model, all times are expressed in *N* generations, which is the relevant time scale for population studies. Therefore, the larger the species, the older has to be the speciation event to observe differences between intra and interspecific diversity. This has to be interpreted as a secondary consequence of the growth of genetic diversity with the effective population size. We would like to mention that in most biological situations, the rate of speciation is very likely to be much smaller than the population drift rate. This illustrates well the benefit of the simulations that can be used to properly assess the theoretical limitation of the method. The second caveat comes from the number of species in the data set: the larger the number of species, the smaller the chance to find them all (to a lesser extent for the radiation model). Indeed, as the number of species grows, it becomes more likely that at least one speciation event is very recent. This again suggests that the only use of genetic data by itself may not be appropriate to delimit species.

We would like to emphasize that our method is excellent in terms of computation time. The GMYC method (Pons *et al.* 2006; Monaghan *et al.* 2009) is extremely slow and computation of a single partition could take up to several months for data sets of moderate size. However, as a next step, it would be very interesting to compare all methods available so far that automatically build partitions and test their benefits and drawbacks using simulations. Indeed, we believe simulations offer an interesting opportunity to test the theoretical limitations of all those methods in controlled scenarios.

Finally, one should always keep in mind that the partition output by ABGD is not meant to be interpreted as a final species delimitation. It is intended to be a first species partition hypothesis on which further work should be carried out. As emphasized in the Introduction, genetic analysis of a single locus is not robust enough to propose reliable species hypotheses. At least one other locus should be added as extra information. More appropriately, we strongly believe that the genetic strategy is an excellent shorthand to build a first species partition hypothesis. However, as exemplified by the model above, there will be cases where genetic data will not be informative (e.g. very recent species) and adding more loci will not systematically improve the results. Therefore, we think that the addition of non-genetic data, such as ecological or morphological ones, is essential to propose robust species hypotheses (De Queiroz 2007; Bond & Stockman 2008; Padial *et al.* 2010). Similarly, ABGD users have to rely on independent data (previously defined species, other barcoding studies that identified a prior divergence value for the same taxa or a closely related group, or any other available data for the studied species) to choose among the different partitions proposed by the ABGD method; ignore the unrealistic hypotheses and confront the plausible alternative ones in an integrative framework.

Acknowledgements

We thank Y. Kobayashi and F. Austerlitz as well as the anonymous reviewers for their constructive comments on the manuscript and on the website interface. We thank T. Treangen for English improvement and constructive scientific comments. G.A. was supported by grant 07-GMGE-004-04 from the Agence Nationale pour la Recherche (French national research agency) and by the JSPS for a long-term fellowship for research in Japan FY2009. A.L. was partially funded by projects BLAN06-3_146282 MAEV and NT09_545511 MANEGE of Agence Nationale pour la Recherche. N.P. was supported by a postdoctoral fellowship from the Centre National de la Recherche Scienfitique.

References

Ahrens D, Monaghan MT, Vogler AP (2007) DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). *Molecular Phylogenetics and Evolution*, **44**, 436–449.

Armstrong KF, Ball SL (2005) DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **10**, S10.

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

Bichain JM, Boisselier MC, Bouchet P, Samadi S (2007) Delimiting species in the genus Bythinella (Mollusca: Caenogastropoda: Rissooidea): molecular and morphometric approachs. *Malacologia*, **49**, 291–311.

Blaxter ML (2004) The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **359**, 669–679.

Bond JE, Stockman AK (2008) An integrative method for delimiting cohesion species: finding the population-species interface in a group of Californian trapdoor spiders with extreme genetic divergence and geographic structuring. *Systematic Biology*, **57**, 628–646.

Campbell DC, Johnson PD, Williams JD, *et al.* (2008) Identification of 'extinct' freshwater mussel species using DNA barcoding. *Molecular Ecology Resources*, **8**, 711–724.

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 12794–12797.

Clare EL, Lim BK, Engstrom MD, Eger JL, Hebert PDN (2006) DNA barcoding of Neotropical bats: species identification and discovery within Guyana. *Molecular Ecology Notes*, **7**, 184–190.

Costa FO, de Waard JR, Boutillier J, *et al.* (2007) Biological identifications through DNA barcodes: the case of the

Crustacea. *Canadian Journal of Fisheries and Aquatic Science*, **64**, 272–295.

Damm S, Schierwater B, Hadrys H (2010) An integrative approach to species discovery in odonates: from character-based DNA barcoding to ecology. *Molecular Ecology*, **19**, 3881–3893.

Dayrat B (2005) Towards integrative taxonomy. *Biological Journal of the Linnean Society*, **85**, 407–415.

De Queiroz K (2007) Species concepts and species delimitation. *Systematic Biology*, **56**, 879–886.

De Salle R (2006) Species discovery versus species identification in DNA barcoding efforts: response to Rubinoff. *Conservation Biology*, **20**, 1545–1547.

Dépraz A, Hausser J, Pfenninger M (2009) A species delimitation approach in the Trochulus sericeus/hispidus complex reveals two cryptic species within a sharp contact zone. *BMC Evolutionary Biology*, **9**, 171.

Durrett R (2008) *Probability Models for DNA Sequence Evolution*. Springer-Verlag, Ithaca, New York 4853–4201.

Elias M, Hill RI, Willmott KR, *et al.* (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proceedings of the Royal Society of London, B*, **274**, 2881–2889.

Elias-Gutierrez M, Jeronimo FM, Ivanova NV, Valdez-Moreno M, Hebert PDN (2008) DNA barcodes for Cladocera and Copepoda from Mexico and Guatemala, highlights and new discoveries. *Zootaxa*, **1839**, 1–42.

Fergusson JWH (2002) On the use of genetic divergence for identifying species. *Biological Journal of the Linnean Society*, **75**, 509–516.

Floyd R, Abebe E, Papert A, Blaxter M (2002) Molecular barcodes for soil nematode identification. *Molecular Ecology*, **11**, 839–850.

Giraud T, Refrégier G, Le Gac M, de Vienne DM, Hood ME (2008) Speciation in fungi. *Fungal Genetics and Biology*, **45**, 791–802.

Goetze E (2010) Species discovery in marine planktonic invertebrates through global molecular screening. *Molecular Ecology*, **19**, 952–967.

Gómez A, Wright PJ, Lunt DH, Cancino JM, Carvalho GR, Hughes RN (2007) Mating trials validate the use of DNA barcoding to reveal cryptic speciation of a marine bryozoan taxon. *Proceedings of the Royal Society of London, B*, **274**, 199–207.

Griffths RC, Marjoram P (1997) An ancestral recombination graph. *Progress in Population Genetics and Human Evolution, MA Volumes in Mathematics and its Applications*, **87**, 257–270.

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006) DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 968–971.

Harpending HC, Sherry ST, Rogers AR, Stoneking M (1993) The genetic structure of ancient human populations. *Current Anthropology*, **34**, 483–496.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, B*, **270**, 313–321.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, e312.

Holland BS, Dawson MN, Crow GL, Hofmann DK (2004) Global phylogeography of Cassiopea (Scyphozoa: Rhizostomeae): molecular evidence for cryptic species and multiple invasions of the Hawaiian Islands. *Marine Biology*, **145**, 1119–1128.

Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princetow, New Jersey 08540

Hudson RR (1990) Gene genealogy and the coalescent process. *Oxford Survey in Evolutionary Biology*, **7**, 1–44.

Janzen DH, Hajibabaei M, Burns JM, Hallwachs W, Remigio E, Hebert PDN (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **360**, 1835–1845.

Kerr KC, Stoeckle MY, Dove CJ, Weigt LA, Francis CM, Hebert PD (2007) Comprehensive DNA barcode coverage of North American birds. *Molecular Ecology Notes*, **7**, 535–543.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Kingman JFC (1982) The coalescent. *Stochastic Processes and Their Applications*, **13**, 235–248.

Knowles LL (2009) Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. *Systematic Biology*, **58**, 463–467.

Lambert A (2008) Population dynamics and random genealogies. *Stochastic Models*, **24**, 45–163.

Lambert A (2009) The allelic partition for coalescent point processes. *Markov Process and Related Fields*, **15**, 359–386.

Locke SA, Daniel McLaughlin J, Marcogliese DJ (2010) DNA barcodes show cryptic diversity and a potential physiological basis for host specificity among Diplostomoidea (Platyhelminthes: Digenea) parasitizing freshwater fishes in the St. Lawrence River, Canada. *Molecular Ecology*, **19**, 2813–2827.

MacQueen J (1967) Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (eds Le Cam LM, Neyman J), pp. 281–297. University of California Press, Los Angeles, California.

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. *Systematic Biology*, **57**, 809–813.

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, **3**, e422.

Miller SE (2007) DNA barcoding and the renaissance of taxonomy. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 4775–4776.

Monaghan MT, Wild R, Elliot M, *et al.* (2009) Accelerated species inventory on madagascar using coalescent-based models of species delineation. *Systematic Biology*, **58**, 298–311.

O'Meara BC (2010) New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology*, **59**, 59–73.

Padial JM, Miralles A, De la Riva I, Vences M (2010) The integrative future of taxonomy. *Frontiers in Zoology*, **7**, 16.

Pons J, Barraclough TG, Gomez-Zurita J, *et al.* (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**, 595–609.

Rach J, De Salle R, Sarkar IN, Schierwater B, Hadrys H (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society of London, B*, **275**, 237–247.

Ramos-Onsins SE, Rozas J (2002) Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution*, **19**, 2092–2100.

Rannala B (1997) Gene genealogy in a population of variable size. *Heredity*, **78**, 417–423.

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, **9**, 552–569.

Rosenberg NA, Hirsh AE (2003) On the use of star-shaped genealogies in inference of coalescence times. *Genetics*, **164**, 1677–1682.

Ross KG, Gotzek D, Ascunce MS, Shoemaker DD (2010) Species delimitation: a case study in a problematic ant taxon. *Systematic Biology*, **59**, 162–184.

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, **141**, 413–429.

Sites JW, Marshall JC (2003) Delimiting species: a renaissance issue in systematic biology. *Trends in Ecology and Evolution*, **19**, 462–470.

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, **129**, 555–562.

Smith MA, Fisher BL, Hebert PDN (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **360**, 1825–1834.

Smith MA, Poyarkov NA Jr, Hebert PDN (2008) COI DNA barcoding amphibians: take the chance, meet the challenge. *Molecular Ecology Resources*, **8**, 235–246.

Ståhls G, Savolainen E (2008) MtDNA COI barcodes reveal cryptic diversity in the Baetis vernus group (Ephemeroptera, Baetidae). *Molecular Phylogenetics and Evolution*, **46**, 82–87.

Stoeckle M (2003) Taxonomy, DNA, and the barcode of life. *BioScience*, **53**, 2–3.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

Valentini A, Miquel C, AliNawaz M, *et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular Ecology Resources*, **9**, 51–60.

van Velzen R, Bakker FT, van Loon JJA (2007) DNA barcoding reveals hidden species diversity in Cymothoe (Nymphalidae). *Proceedings of the Netherlands Entomological Society Meeting*, **18**, 95–103.

Vences M, Thomas M, Bonett RM, Vieites DR (2005) Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **360**, 1859–1868.

Vogler AP, Monaghan MT (2007) Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research*, **45**, 1–10.

Wiemers M, Fiedler K (2007) Does the DNA barcoding gap exist?—a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology*, **4**, 8.

Wiens JJ (2007) Species delimitation: new approaches for discovering diversity. *Systematic Biology*, **56**, 875–878.

Will KW, Mishler BD, Wheeler QD (2005) The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology*, **54**, 844–851.

Yeates D, Seago A, Nelson L, Cameron SL, Joseph L, Trueman JWH (2010) Integrative taxonomy, or iterative taxonomy? *Systematic Entomology*, **36,** 209–217.

Zinger L, Coissac E, Choler P, Geremia RA (2009) Assessment of microbial communities by graph partitioning in a study of soil fungi in two Alpine meadows. *Applied and Environmental Microbiology*, **75**, 5863–5870.

P.N. is a Post-Doc student at the Museum National d'Histoire Naturelle, Paris. He is interested in the taxonomy, diversification and evolution of marine gastropods, and in particular the Conoidea. A.L. is a Professor of mathematics at the University Pierre & Marie Curie (Paris VI). He is particularly interested in the stochastic modelisation of biological processes. S.B. is a Research Ingenior at the University Pierre & Marie Curie (Paris VI). She is specialized in the design of bioinformatic tools. G.A. is an Assistant Professor at the University Pierre & Marie Curie (Paris VI). He is developping researches on theorical biology, molecular evolution and bioinformatic.