ASAP: assemble species by automatic partitioning

Nicolas Puillandre, Sophie Brouillet, Guillaume Achaz

ASAP: Assemble Species by Automatic Partitioning

Running title: Assemble Species by Automatic Partitioning

Nicolas Puillandre[1,*], Sophie Brouillet[1,*], Guillaume Achaz[3,2,1]

[1] Institut Systématique Evolution Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles. 57 rue Cuvier, CP 26, 75005 Paris, France.

[2] SMILE group, CIRB, UMR 7241, Collège de France, CNRS, INSERM, Paris, France.

[3] Éco-anthropologie, Muséum National d'Histoire Naturelle, CNRS UMR 7206, Université de Paris, Paris, France

.

Corresponding author: N. Puillandre, puillandre@mnhn.fr.

*: equally contributed to the work.

ABSTRACT

We describe ASAP (Assemble Species by Automatic Partitioning), a new method to build species partitions from single locus sequence alignments (*i.e.* barcode datasets). ASAP is efficient enough to split datasets as large 10^4 sequences into putative species in several minutes. Although grounded in evolutionary theory, ASAP is the implementation of a hierarchical clustering algorithm that only uses pairwise genetic distances, avoiding the computational burden of phylogenetic reconstruction. Importantly, ASAP proposes species partitions ranked by a new scoring system that uses no biological prior insight of intra-specific diversity. ASAP is a stand-alone program that can be used either through a graphical web-interface or that can be downloaded and compiled for local usage. We have assessed its power along with three others programs (ABGD, PTP and GMYC) on 10 real COI barcode datasets representing various degrees of challenge (from small and easy cases to large and complicated datasets). We also used Monte-Carlo simulations of a multi-species coalescent framework to assess the strengths and weaknesses of ASAP and the other programs. Through these analyses, we demonstrate that ASAP has the potential to become a major tool for taxonomists as it proposes rapidly in a full graphical exploratory interface relevant species hypothesis as a first step of the integrative taxonomy process.

KEYWORDS

ASAP, species delimitation, integrative taxonomy, ABGD, DNA barcoding.

INTRODUCTION

During the last 15 years, following the success of the DNA-barcoding projects and the increase in sequencing capacities, many methods of species delimitation based on DNA sequences have been developed. They can be roughly classified into two main categories. A first one includes methods that compute the likelihood of competing partitions of species hypotheses ("models") in the so-called "multi-species coalescent" framework. In this category, the most popular methods are SpedeSTEM (Ence & Carstens, 2011), BPP (Yang & Rannala, 2014) and BFD (Leaché, Fujita, Minin, & Bouckaert, 2014), reviewed (with other methods) in several articles (Camargo & Sites, 2013; Carstens, Pelletier, Reid, & Satler, 2013; Fujita, Leaché, Burbrink, McGuire, & Moritz, 2012; Leavitt, Moreau, & Lumbsch, 2015; Rannala, 2015). They were designed for multilocus data and are computationally (extremely) demanding. As a consequence, they have been mainly applied to datasets with limited number of sequences and species, and to well-studied groups, for which competing partitions of species have been proposed in the literature; they generally correspond to species complexes, typically in the grey zone (De Queiroz, 2005).

A second category of methods corresponds to exploratory ones, *i.e.* methods that propose *de novo* species partitions, typically from a single-locus, DNA-barcoding-like, datasets. Although sometimes criticized because a single gene tree poorly represents the species tree (Degnan & Rosenberg, 2009; Nichols, 2001), these methods are widely used, as they are easy to apply on DNA-barcoding datasets, even large, and precisely because they do not necessitate pre-defined species hypotheses. The most popular ones are GMYC –General Mixed Yule-Coalescent model– (Pons et al., 2006), PTP –Poisson Tree Process– (Zhang, Kapli, Pavlidis, & Stamatakis, 2013), both first developed in a maximum likelihood

framework, and later extended to a Bayesian framework (Reid & Carstens, 2012), and ABGD –Automatic Barcode Gap Discovery– (Puillandre, Lambert, Brouillet, & Achaz, 2012). GMYC and PTP take as input a phylogenetic tree and estimate rates of branching events to infer which part of the tree more likely follows a speciation model (the deepest part) and which part follows a coalescent model (subtrees of the shallowest part). The species partition is found by maximizing the likelihood of the transition between these two branching rates, GMYC in absolute time (hence the need for an ultrametric tree), PTP in mutational time at different nodes of the tree. GMYC and PTP first inferred a single transition event between the two rates (speciation *vs* coalescent) and were later expanded to infer "multiple thresholds", allowing several transitions to occur in different subtrees (Kapli et al., 2017; Monaghan et al., 2009).

Contrary to the two previous methods, ABGD uses only pairwise genetic distances (no tree is inferred) and automatically identifies in their distribution the so-called "barcode gap". This gap marks the limit between the smaller intra-specific distances and the larger inter-specific distances. From the gap, a distance threshold is estimated and used to partition the samples into putative species. A coalescent model is used to identify the position of the most likely barcode gap, based on a maximal genetic intraspecific divergence *P* defined *a priori* by the user. Consequently, users must provide a range of *P* in which ABGD identifies one or several barcode gaps and the method outputs the corresponding species partitions. For a single dataset, ABGD thus eventually proposes several partitions that correspond to different prior values *P*. In its recursive version, ABGD is applied on each group of the initial partition, and eventually splits them when internal barcode gaps are detected.

The relative performances of these three exploratory methods, GMYC, PTP and ABGD, sometimes together with less used methods (Flot, Couloux, & Tillier, 2010; Ratnasingham &

Hebert, 2013) have been compared in various taxa: mammals (Derouiche, Vercammen, Bouhadad, & Fernandes, 2017), amphibians (Vacher et al., 2017), squamates (Blair & Bryson, 2017), fishes (Ramirez et al., 2017), echinoderms (Boissin, Hoareau, Paulay, & Bruggemann, 2017), insects (Lin, Stur, & Ekrem, 2015), spiders (Ortiz & Francke, 2016), crustaceans (Larson, Castelin, Williams, Olden, & Abbott, 2016), pycnogonids (Dömel, Melzer, Harder, Mahon, & Leese, 2017), rotifers (Papakostas et al., 2016), annelids (Decaëns et al., 2016), molluscs (Fourdrilis et al., 2016), flatworms (Scarpa et al., 2017), nemerts (Leasi & Norenburg, 2014), cnidarians (Arrigoni et al., 2016), plants (Lithanatudom et al., 2017), algae (Zou et al., 2016), lichens (Pino-Bodas, Burgaz, Teuvo, & Stenroos, 2018), fungi (Alors, Lumbsch, Divakar, Leavitt, & Crespo, 2016) and foraminifera (André et al., 2014).

Although the results obtained with the various methods often vary depending on dataset characteristics (e.g. Blair & Bryson, 2017), the main conclusions of these studies are:

1. all methods generally perform well (but see e.g. Dellicour & Flot, 2018) being mostly congruent (*i.e.* providing similar species partitions) with each other and with the species partitions inferred from independent data (e.g. other molecular markers, morphological data, ecological data);
2. all of them perform poorly when the number of sampled individuals per species is too low (Ahrens et al., 2016), or when the contrast of intra- *vs*. interspecific divergences is mild. This contrast varies with species ages, mutation rates, population sizes, strengths of the selection and degrees of within-species population structure (Pante et al., 2015; Pentinsaari, Vos, & Mutanen, 2016; Ritchie, Lo, & Ho, 2016); mPTP was in particular developed to overcome this issue (Kapli et al., 2017);
3. partitions proposed by the three methods sometimes differ, each of them being able to infer the “correct” species when the two others fail. This led some authors to propose

that all three methods (among with eventually others) should be applied jointly and compared (Ducasse, Ung, Lecointre, & Miralles, 2020);

4. Although there are several exceptions (e.g. Blair & Bryson, 2017), ABGD in particular, and PTP to a lesser extent, tend to lump species more than GMYC (Pentinsaari et al., 2016). Conversely, the multiple-threshold version of GMYC is particularly prone to oversplitting (Fujisawa & Barraclough, 2013; Kekkonen & Hebert, 2014).

In comparison with GMYC and PTP, ABGD has the advantage of being very fast, mainly because it bypasses the phylogenetic reconstruction. Furthermore, because ABGD identifies a species partition for each value of *P* defined a priori, several partitions may be proposed, reflecting the uncertainty stemming from the data and encouraging the user to evaluate the relevance of the ABGD partitions in the light of other data, as it is recommended in an "integrative taxonomy" approach. However, ABGD does not provide a score for each partition that would help the user to identify the "best" partition(s), and this probably constitutes the main drawback of ABGD (judging from the numerous comments and questions the authors of ABGD have received from the users).

In this article, we describe a new method of species delimitation, still based on pairwise genetic distances, but which implementation provides a score for each defined partition and overcomes the challenge of *a priori* defining *P*. Our new algorithm, ASAP (Assemble Species by Automatic Partitioning), still provides several partitions, more or less fine-grained, but ranked using a new scoring system. Importantly, we also develop a full graphical web-interface to ease its usage. However, ASAP, like any other method, must not replace the taxonomist work, as any partition of species must be subsequently tested against other

evidences in an integrative taxonomy framework. This is especially crucial as ASAP uses single-locus data that are known to bear weaknesses.

MATERIAL AND METHODS

Overview of the ASAP software

ASAP is a C self-contained program. Users can use ASAP either through a full graphical web-interface (https://bioinfo.mnhn.fr/abi/public/asap), or download and compile the sources for local usage (same url).

Our algorithm is an ascending hierarchical clustering, merging sequences into *groups* that are successively further merged until all sequences form a single group. At each merging step, the assignment of all sequences into groups is named a *partition*. The first partition contains as many groups as sequences (no grouping was yet done) whereas the last partition is a single group with all sequences inside. Larger groups are created by merging groups of the previous partition together. We characterize all newly created partition in two complementary ways. First, we assign to it a probability that quantifies the chances that each of its new groups is a single species. Second, we compute the width of the barcode gap between the previous and this new partition. Both metrics (probability and barcode gap width) are combined into a single *asap-score* that is used to rank the partitions.

ASAP in details

i) Ranked distances

We first start by computing, when not provided, all pairwise distances between the *n* sequences of the alignment. Distances are then ranked by increasing values. The efficiency of

the algorithm stems from the fact that each distance is only considered once in increasing order for clustering purposes.

ii) Hierarchical clustering

The clustering process starts with a first partition where each sequence belongs to a different group. ASAP then treats each of the ranked distances one by one in increasing order (equal distances are treated together) as a threshold value for delimiting groups: sequences separated by a distance equal to the current value d_C are clustered into the same group. Consequently, when sequences that were in different groups are clustered together, the previous groups are merged into a new larger group, and is associated to the current clustering distance, d_C. Importantly, a new partition can have a single new group or several new ones when several sequences from different groups are merged independently into different groups for the same distance d_C. When a new partition is built, the clustering process pauses. ASAP then scores all new groups with a probability of panmixia. It also scores the new partition using an ad-hoc score computed from both the barcode gap width and probabilities of panmixia. After the group(s) and partition scoring, ASAP then continues the clustering by looking after the next distances until another partition is built. The algorithm stops when all sequences are merged into a single final group.

iii) Computing p-values

a. For each group: we aim at computing a p-value for a newly created group that is a merge of two or more subgroups. We compute Π_{intra} the average pairwise distance between sequences within the subgroups and Π_{inter} the average pairwise distance among sequences of different subgroups (Figure 1). We then compare Π_{intra} to its theoretical distribution, computed by Monte-Carlo simulations of a neutral coalescent model assuming a single panmictic species with a sample size m and a coalescent mutation rate $\theta = \Pi_{inter} / [2\times(1-1/m)]$.

The value of θ is set so that in the simulations the distance between sequences connected by the Most Recent Common Ancestor (MRCA) of the group (π_{inter}) is equal, on average, to the observed one: $E[\pi_{inter}]=\Pi_{inter}$. This relates to the average time to the MRCA that is $2\times(1-1/m)$, expressed in coalescent time (Wakeley, 2009). We compute the p-value as the fraction of replicates where the simulated π_{intra} is equal or lower than the observed Π_{intra}. The number of replicates is updated on the fly to have correct estimations of low p-values. Put differently, it quantifies under H0 (one single species) the probability of observing a diversity Π_{intra} or less within the subgroups given that the divergence between the subgroups is on average Π_{inter}.

b. For partitions: we compute the probability to observe π_{intra} or less diversity within all subgroups of the *current* partition (that are groups of *previous* partition before the merge) assuming that all new groups of the *current* partition are independent coalescent realizations with θ estimated for each group independently.

iv) Recursive splits

Once a new partition is built, ASAP tests for each of the groups of the partition whether its p-value is lower than a given risk (by default 1%) and consequently should be split. When a group is split, ASAP recursively descends to all its subgroups and assesses whether they should be split as well.

v) Relative barcode gap width

ASAP also computes a relative barcode gap width associated to the current partition (Supplementary Material 1). The partition is associated to a threshold distance d_T that is the mid-point between the current distance, d_C (with rank r_C), that triggered the merging and the previous distance in the list d_{C-1} (with rank r_C-1). A barcode gap corresponds to a "jump" in the distance values in only few ranks. While increasing only few ranks in the list, the distance will "jump" from a value that is (much) less than d_T to a value that is (much) higher than d_T.

To quantify the barcode gap width, ASAP scans downward the distance list from d_{C-1} until it finds the first distance smaller than $0.9d_{C-1}$: this is d_L which rank is r_L in the list. It then scans from d_C the distance list upward until it finds the first distance above $1.1d_C$: this is d_H which rank is r_H. The relative gap width W is defined as:

$$W = [(d_H - d_L) / (d_H+d_L+1)] / (r_H - r_L).$$

We normalized the difference of distance ($d_H - d_L$) by (d_H+d_L+1) to compute the "relative" width of the gap; the "+1" only prevents the ratio to be very high when distance values are very small. The higher the W, the larger the barcode gap.

vi) Outputs

At the end of the clustering, ASAP scores and sorts all the different partitions using two criteria: their p-value sorted (see iii.b) by increasing order (the smallest p-value has rank 1) and their rank of relative barcode gap width (see v) sorted by decreasing order (the largest gap has rank 1). The *asap-score* is the average of both ranks: the smaller, the better. Furthermore, ASAP produces a graphical output where each node of the hierarchical clustering is color-coded depending on its probability of being a panmictic species (see iii.a). Thus, the color guides the user finding which nodes may be split into smaller groups. Several other graphical options are provided to help the user navigate among partitions and choose the "most relevant" partition, beyond a simple naive use of the *asap-score* (Supplementary Material 2).

Tests on empirical data

To compare the results obtained by four methods (ASAP, (m)PTP, (m)GMYC and ABGD), we selected 10 empirical COI datasets covering various taxa (birds, mammals, amphibians, insects, crustaceans and molluscs) and including 44 to 2,574 specimens that belong to 5 to 643 species (Table 1) (Borisenko, Lim, Ivanova, Hanner, & Hebert, 2008; Elias-Gutierrez,

Jeronimo, Ivanova, Valdez-Moreno, & Hebert, 2008; Hajibabaei, Janzen, Burns, Hallwachs, & Hebert, 2006; Kerr et al., 2007; Puillandre, Cruaud, & Kantor, 2010; Puillandre, Baylac, Boisselier-Dubayle, Cruaud, & Samadi, 2009; Puillandre, Fedosov, Zaharias, Aznar-Cormano, & Kantor, 2017; Puillandre et al., 2011, 2012; Smith, Poyarkov Jr., & Hebert, 2008). Among them, five correspond to datasets published by one of the authors to facilitate the interpretations of the results. An eleventh dataset, including 9,396 sequences of moths (publicly available from BOLD), was used to estimate and compare the computation times of ABGD and ASAP. A dataset of this size could not be analyzed by (m)GMYC or (m)PTP as the phylogenetic reconstruction is too costly.

For all empirical datasets, we used the web version of ABGD, with default parameters. Only the initial partitions were considered, and only the more stable partition(s) (i.e. the partition(s) found with several *P* in the vicinity of the barcode gap) was (were) reported. For ASAP, we used a recursive split probability of 0.01 (see iv), and report a) the partition with the best *asap-score* as well as b) the partition that is closest to the "correct" one among the two best partitions, according to their *asap-scores*. For GMYC and mGMYC, ultrametric trees were reconstructed using BEAST 2 (Bouckaert et al., 2014), with an independent GTR substitution model for each codon position. Relative divergence times were estimated using a relaxed log-normal clock with a coalescent prior and a constant population size, following the recommendations of Monaghan et al. (2009). The number of MCMC steps were 20M (*Gemmuloborsonia*, *Benthomangelia*, *Lophiotoma* and *Eumunida* datasets), 100M (Amphibians, Cladocera, Mammals, Sphingidae and Turridae datasets) and 200M (Birds dataset), sampled every 2,000, 10,000 and 20,000 steps respectively. Convergence of the runs was assessed using TRACER 1.6 (Rambaut & Drummond, 2014) to check that all effective

sample size values exceeded 200. Consensus trees were calculated after discarding the first 25% of the trees as burn-in, with the option "Common Ancestry" for node height.

For PTP and mPTP, the web server at https://mptp.h-its.org/#/tree was used, with default parameters. The input tree was obtained with RAxML (Stamatakis, 2006), with an independent GTR substitution model for each codon position. All phylogenetic analyses were performed on the Cipres Science Gateway (http://www.phylo.org/portal2), using the BEAST2 on XSEDE (2.1 - 2.4.8) and RAxML-HPC2 on XSEDE (8.2.10) tools.

Simulations

We measured the power of ABGD, GMYC, (m)PTP and ASAP to retrieve the correct species partition in various scenarios using Monte Carlo simulations. We used a "multispecies coalescent" framework (Rannala & Yang, 2003) with different options and parameters using Monte-Carlo simulations, as described previously (Puillandre et al., 2012). Note that contrarily to the standard multispecies coalescent, the species tree is here drawn from a probability distribution. The home-made C simulator is available upon request.

Briefly, for each replicate, we generate a species tree using either a Yule model (all lineages have the same birth rate) or a radiation model (all species arose at the same time). Radiation (hard polytomy) models cases where all speciation events follow each other quickly and where no mutations have occurred between the first (the root) and the last speciation event. We used a backward coalescent version of these models that we have previously used for ABGD evaluation (Puillandre et al., 2012). For the radiation model a unique speciation event, exponentially distributed with rate r, is drawn. For the Yule model (n_{sp}-1) speciation events are drawn with identical rate (Lambert & Stadler, 2013).

Once the species tree is obtained, we assign sequences to species uniformly, with at least 1 sequence per species. All species (current and ancestral) are assumed to be of equal effective size (*i.e.* N individuals). The genealogy of the sequences is then simulated in backward time using a standard Kingman coalescent process but forbidding coalescent events between lineages from different species. Once the genealogy is obtained, a Poisson random number of mutations – with mean $L\theta/2$, where L is the total tree length and θ the population mutation rate – are distributed uniformly on the tree and the resulting polymorphic sites are generated. The whole simulation process is tuned by 4 parameters:

- a total number of sequences $\boldsymbol{n}$,
- a number of species $\boldsymbol{n_{sp}}$ with one or more sequences,
- a speciation rate $\boldsymbol{r}$, expressed in coalescent time (*i.e.* in N generations),
- a mutation rate $\boldsymbol{\theta}$, expressed in coalescent scale ($\theta = 2 N \mu$), set to $\theta=10$ for 600bp of simulated sequence. Mutations are only substitutions following a Jukes-Cantor model.

ABGD and ASAP use the pairwise distance matrix as input. For ABGD, we used a prior value of 0.083 (5x10/600) that is an excellent prior representing a situation where the user has near perfect knowledge on maximal diversity within species. For GMYC and (m)PTP, we used as input the 'true' gene genealogy (the one simulated for the replicates) not only to fasten the simulation (*i.e.* skipping the phylogenetic reconstruction) but also to assess their power when the phylogeny is perfectly reconstructed. We would like to emphasize that only ASAP used unprocessed data (polymorphic sites) without any biological insights (no prior, no phylogeny reconstruction nor calibration).

RESULTS

Empirical datasets

We first assessed the ability of ASAP through a proxy that is its ability to retrieve the "correct" number of species in 10 empirical datasets (Table 1). The datasets were selected to represent test cases of different sizes (from 44 sequences/5 species to 2,574 sequences/643 species). We first report the number of species predicted in the partition with the best *asap-score* (ASAP 1st): we found that in 4/10 of the datasets, the partition with the best *asap-score* is *very close* to the reference one (less than 5% difference in terms of species numbers) and that 8/10 is *close* (less than 10% difference). If we also consider the partition with the second best *asap-score* (ASAP 1st and 2nd), the degree of accuracy increases to 6/10 for the *very close* ones and 9/10 for the *close* ones. This is a good indication that ASAP users should consider not only the partition with the best *asap-score* but also few subsequent ones. It is important to report that here no extra biological knowledge was considered for ASAP predictions. One could for example use threshold distances (e.g. d_T or d_C) to prefer one partition over another despite a poorer *asap-score* (e.g. in most clades intra-specific diversity is typically on the order of 1%, not on the order of 10%). Obviously, other criteria and characters should also be used to choose a final species partition, in an integrative taxonomy context.

One of the ASAP main qualities is that it is extremely fast compared to any method that relies on tree reconstruction. The online version takes 45 seconds for the largest dataset of Table 1 (2,574 aligned sequences; 643 species) for all steps of the complete method: mainly creating the distance matrix, performing the clustering and computing probabilities by Monte Carlo at each node. We observed that the CPU time increases linearly with the number of species in the datasets (Figure 2) and only to a lesser extent with the number of sequences (data not shown). We estimate the CPU cost at 0.07 sec per species for the current web version. This

suggests that most of the CPU time is taken by probability estimations of significant nodes (see method, section iii) (non-significant ones are not as costly in our implementation as we increase the number of replicates only for nodes with low probabilities). The number of significant nodes likely increases approximately linearly with the number of species. The time for distance matrix computation and clustering both increase quadratically with the number of sequences and are independent from the number of species.

On a curated unpublished moth dataset, it took 6 min 35 on the website to delimit 2,466 species (best *asap-score*) or 2,067 (second best *asap-score*) from 9,396 sequences. Subsequent partitions with lower *asap-scores* are close to one or the other of these two first partitions. Because of its rapidity, ASAP web server accepts up to 10^4 sequences (unlike the ABGD server).

We also took the opportunity of analyzing the 10 datasets to assess the performance of other methods: ABGD which is solely based on pairwise distances, PTP and mPTP that were run on an ML trees (*i.e.* RaxML) and GMYC and mGMYC on an ultrametric trees estimated by a Bayesian MCMC method (*i.e.* BEAST). Results (Table 1) show that ABGD performance is similar to ASAP 1^{st}-2^{nd}, that PTP and mPTP tend to not perform very well, that GMYC performs very well provided that the number of species is not too large and that, as previously reported in the literature, mGMYC generally oversplits (Fujisawa & Barraclough, 2013; Kekkonen & Hebert, 2014). Note that ABGD performances are somehow overestimated as we report the partition that is the closest to the reference one over the whole range of *P*. We could not use GMYC for the largest dataset as the Bayesian tree reconstruction did not converge after several weeks of computation.

Simulated datasets

We then assess the theoretical performance of ASAP using Monte-Carlo simulations of a multispecies coalescent framework. In brief, a random species tree is generated using either a Radiation model, where all species arose in single event, or a Yule model, where the speciation events occur at constant rate independently in all branches. In both model, we tune the separation of time scales (speciation versus intra-specific coalescent events) using a speciation rate that is expressed in coalescent time (*i.e.* N generations per unit of time). The lower the speciation rate, the better the separation of time scales. For example, when the speciation rate is 0.1, speciation events are 10 times slower than pairwise coalescent events within species.

The impact of speciation rate on ASAP

We first examine the ability of ASAP to correctly retrieve four species in both speciation models as a function of the speciation rate (from 0.001 to 1). We report in Figure 3 the fraction of runs where ASAP was able to correctly retrieve the four species (top panel) and the average number of predicted species, regardless of their composition (bottom panel). We assess the quality of the partition with the best *asap-score* (ASAP 1^{st}) as well as the quality of the partition that is the closest to the truth among the two best partitions (ASAP 1^{st}-2^{nd}).

We observe that for low rates of speciation, the best partition proposed by ASAP correspond exactly to the four species. This is an “easy” case where the two time scales are well separated. As the speciation rate increases, both time scales overlap and it becomes harder to delineate species using pairwise genetic differences at a single locus. When the speciation rate is larger than 1, speciation events are more recent than intra-specific divergence so that individuals within species are no more different than individuals between species.

ASAP performs usually better with the Radiation than with the Yule model. This is especially striking for moderate speciation rate (e.g. 0.03). For radiations, most of the errors correspond to oversplit, as illustrated by the average number of predicted species that is larger than four. Under the Yule model with four species, there are three independent speciation events and consequently there is a higher chance to generate at least one very recent speciation event that would be invisible in regard of sequence divergence. Indeed, the most recent event is exponentially distributed with rate 3*r*. As a consequence, contrarily to the radiation model, ASAP failures correspond for this rate to cases where it lumps the two closest species into a single one.

The impact of the number of species on ASAP

Second, we explore the impact of the number of species for a fixed sample size of 200 sequences, with r=0.01, a moderately challenging speciation rate. We report the average number of predicted species regardless of their composition for both the radiation and the Yule models. Results (Figure 4) show a) that ASAP very well predicts the species under a radiation model, regardless of the number of species and b) that it only finds a fraction of them for the Yule model. Under the Yule model, the problem of finding a threshold between intra- and inter-specific distance becomes harder as the most recent speciation event is exponentially distributed with rate $r.(n_{sp}-1)$; the more species, the more recent the last speciation event. Furthermore, the higher the number of species the higher the chance to have a very old coalescent MRCA (Most Recent Common Ancestor) within one of the species. This old MRCA translates into a high divergence among individuals of this species, which would also obscure the threshold between intra- and inter-specific genetic divergences.

The impact of the number of species on ABGD, PTP and GMYC

We apply the same analysis to ABGD, (m)PTP and GMYC. We would like to emphasize again that we assessed their power under optimal conditions: a single “excellent” prior for ABGD representing a perfect knowledge of intraspecific diversity and the “true” simulated tree for (m)PTP and GMYC, bypassing their main limitations, that is having a correctly reconstructed phylogenetic tree. As a consequence, we here overestimate their power for realistic biological situations where only a set of sequences is available (neither the true tree nor prior knowledge of intraspecific diversity is known). ASAP, on the contrary, directly uses the sequences and needs no prior biological insight or phylogenetic reconstruction.

The power assessments of the methods (Figure 4) show that ABGD retrieves well the correct partition when speciation occur as a single radiation but has a limited power when speciations follow a Yule model. On the contrary, we found that GMYC performs very well for the Yule model but is less efficient for a radiation model. Interestingly mPTP consistently split a constant small number of species. It thus performs poorly when the number of species is low but quite well when the number of species is 50 or more.

DISCUSSION

We introduced a new species delimitation program, ASAP, fully exploratory, in the sense that it does not require any *a priori* knowledge, neither on the number of species, the species composition, or any biological information, such as a phylogenetic tree or *a priori*-defined intraspecific genetic distances. Only pairwise genetic distances are used to build a list of partitions ranked by a score. This composite score is computed using the probabilities of groups to be panmictic species and the barcode gap widths. ASAP overcomes the two mains

limitations of ABGD, namely (i) the need for an *a priori* defined *P* and (ii) the lack of a scoring system.

However, and contrary to some other methods, ASAP still outputs several partitions, ranked by their *asap-scores*. A list of the "best" partitions (10 by default) is provided in the output together with their gap-width score, their p-value, their threshold distance d_T and the number of species they correspond to.

The graphical output of ASAP has four main components (Supplementary Material 2):

(1) a list of partitions ranked by their asap-score that putatively correspond to species hypothesis,

(2) a plot of the *asap-score* as a function of d_C. We report the *asap-score* of all partitions (not only the best ones) as a function of the clustering distance d_C to appreciate whether all good partitions have similar d_C or whether "potentially good" partitions can drastically differ in size.

(3) an ultrametric clustering tree of all sequences, where the distance to the leaves lengths correspond to the distance d_C at which these sequences were clustered in the same group. All nodes of this tree are color-coded depending on their p-value (the darker the more it differs from a panmictic species).

(4) a "boxed-species" graph, where species hypotheses in the different partitions are represented as vertical boxes in front of the ultrametric tree.

When a partition is selected by a click in any of the three panels, it is automatically highlighted in the two other components.

We also propose a complementary representation, where we display the hierarchical tree with, at its leaves, the 10 best ASAP partitions where their groups are depicted as boxes (that are similar to the boxes of Figure 1).

We have evaluated ASAP strengths and weaknesses using both real and simulated data. Our benchmark shows that ASAP performs well delivering partitions in a matter of minutes even for datasets as large as 10^4 sequences. ASAP is thus meant to be applied on large single-locus datasets when no species hypothesis is available, as typically produced in DNA-barcoding projects. Although the web version limits the input to 10^4 sequences, more sequences can be analyzed using a local command-line version of ASAP (sources are available on the webserver).

The comparison with the other programs shows that ASAP and ABGD both perform well for a Radiation model, because there are no "recent" invisible speciation events. Indeed, both methods use a phenetic approach were similar sequences are simply clustered in the same group/species. On the contrary, (m)GMYC and (m)PTP that are explicitly based on a phylogenetic approach behave differently, performing quite well under a Yule model. More generally, (m)GMYC and (m)PTP are both relying on a different property to propose species hypotheses, compared to ABGD and ASAP: specimens belonging to the same species, *i.e.* to the same diverging lineage, share a common evolutionary history, *i.e.* they form a clade. Indeed, phenetic differences are calculated by simply counting the differences among sequences, whereas the phylogenetic criterion requires the reconstruction of a proper phylogenetic tree. This additional step in the (m)GMYC and (m)PTP methods potentially introduces a bias, because a) phylogenetic trees reconstructed on a single locus may differ drastically from the species tree, and b) the limited number of sites in a single marker may lead to incorrectly reconstructed trees. Consequently, (m)GMYC and (m)PTP have been shown to be sensitive to the reconstruction method (Tang, Humphreys, Fontaneto, & Barraclough, 2014). On the contrary, it could be argued that relying only on genetic distances, *i.e.* without testing if these differences actually correspond to distinct evolutionary histories,

and not to homoplasy, must be used with caution. Indeed, the efficiency of each method in delimiting species probably depends on various characteristics of the species and datasets (number of samples, number of species, population sizes…), and applying several methods to a given dataset is a strategy commonly applied to maximize the probability to detect species complexes, identified as groups of species whose limits vary depending on the method.

Importantly, several other methods can also be used to delimit species, such as BINs (Ratnasingham & Hebert, 2013), Jmotu (Jones, Ghoorah, & Blaxter, 2011) or VSEARCH (Rognes, Flouri, Nichols, Quince, & Mahé, 2016), among others (e.g. Rannala & Yang, 2020). We are also aware that the number of predicted species is only a proxy to assess the performance of the different methods. Indeed, other metrics such as the F-measure (Larsen & Aone, 1999) or the number of splits or merges (Ratnasingham & Hebert, 2013) give also insightful information. Some of them are even implemented in meta-analysis software such as LIMES (Ducasse, Ung, Lecointre, & Miralles, 2020), which could be used to perform a more extensive benchmark of all existing methods using a wider spectrum of metrics.

More generally, and as advocated by the proponents of the integrative approach in taxonomy, the use of a single marker with a single method of species delimitation should be avoided, precisely because each method has its own limitations. Some methods are based on a phenetic criteria (e.g. ASAP and ABGD) while others on phylogenetic criteria (e.g. (m)GMYC and (m)PTP). Furthermore a single locus may not follow the species history, because of introgression and incomplete lineage sorting. This is particularly true for species in the grey zone, in which the gene tree may differ from the species tree, and the coalescent events may be older than the speciation events (De Queiroz, 2005). For this reason, we recommend that single-locus methods are to be used as a first step of the species delimitation process that is to propose primary species *hypotheses*. This is for example useful in groups for which there is

no pre-existing hypotheses to test, or for which unknown/incorrectly delimited species represent the majority of the diversity (e.g. microbial communities or hyperdiverse groups of eukaryotes, such as insects, spiders, nematodes, mollusks…). Furthermore, DNA barcodes are now routinely produced using NGS approaches, providing large numbers of sequences often not assignable to known and sequenced species (Kennedy et al., 2020), and for which methods such as ASAP are welcome to *e.g.* compare species diversity among sites.

In a second step it is then the responsibility of the taxonomist to evaluate with other methods (in particular, methods that will evaluate alternative partitions of species) and/or lines of evidence (such as other genetic markers, morphology or ecology) whether the proposed hypotheses are robust, or not. In this context, methods such as ASAP, ABGD, (m)PTP and (m)GMYC should thus be seen as a formalized and reproducible way to propose species hypotheses in groups where no such hypotheses exist, or, if they do exist, that are better to be ignored.

ACKNOWLEDGMENTS

This work was supported by the CONOTAX project funded by the French National Research Agency (grant number ANR-13-JSV7-0013-01). The authors would like to thank Amaury Lambert for his valuable advices, Malcolm Sanders, Paul Zaharias, Laure Corbari and Romain Sabroux for having tested preliminary versions of ASAP, and all the users of ABGD.

REFERENCES

Ahrens, D., Fujisawa, T., Krammer, H.-J., Eberle, J., Fabrizi, S., & Vogler, A. P. (2016). Rarity and incomplete sampling in DNA-based species delimitation. *Systematic Biology*, *65*(3), 478–494.

520 Alors, D., Lumbsch, H. T., Divakar, P. K., Leavitt, S. D., & Crespo, A. (2016). An integrative approach for
521 understanding diversity in the *Punctelia rudecta* species complex (Parmeliaceae,
522 Ascomycota). *PloS One*, *11*(2), e0146537.
523 André, A., Quillévéré, F., Morard, R., Ujiié, Y., Escarguel, G., De Vargas, C., … Douady, C. J. (2014). SSU
524 rDNA divergence in planktonic foraminifera: molecular taxonomy and biogeographic
525 implications. *PLoS One*, *9*(8), e104641.
526 Arrigoni, R., Berumen, M. L., Chen, C. A., Terraneo, T. I., Baird, A. H., Payri, C., & Benzoni, F. (2016).
527 Species delimitation in the reef coral genera *Echinophyllia* and *Oxypora* (Scleractinia,
528 Lobophylliidae) with a description of two new species. *Molecular Phylogenetics and*
529 *Evolution*, *105*, 146–159.
530 Blair, C., & Bryson, R. W. (2017). Cryptic diversity and discordance in single-locus species delimitation
531 methods within horned lizards (Phrynosomatidae: *Phrynosoma*). *Molecular Ecology*
532 *Resources*, *17*(6), 1168–1182.
533 Boissin, E., Hoareau, T. B., Paulay, G., & Bruggemann, J. H. (2017). DNA barcoding of reef brittle stars
534 (Ophiuroidea, Echinodermata) from the southwestern Indian Ocean evolutionary hot spot of
535 biodiversity. *Ecology and Evolution*, *7*(24), 11197–11203.
536 Borisenko, A. V., Lim, B. K., Ivanova, N. V., Hanner, R. H., & Hebert, P. D. N. (2008). DNA barcoding in
537 surveys of small mammal communities: a field study in Suriname. *Molecular Ecology*
538 *Resources*, *8*, 471–479.
539 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., … Drummond, A. J. (2014). BEAST
540 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *10*(4),
541 e1003537.
542 Camargo, A., & Sites, J. Jr. (2013). Species delimitation: a decade after the renaissance. In I. ISBN:
543 978-953-51-0957-0 DOI: 10. 5772/52664 (Ed.), *The Species Problem - Ongoing Issues, book*

edited by Igor Ya. Pavlinov, ISBN 978-953-51-0957-0, Published: February 6, 2013 under CC BY 3.0 license. Pavlinov, I. Y.

Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species delimitation. *Molecular Ecology*, *in press*.

De Queiroz, K. (2005). A unified concept of species and its consequences for the future of taxonomy. *Proceedings of the California Academy of Sciences*, *56*, 196–215.

Decaëns, T., Porco, D., James, S. W., Brown, G. G., Chassany, V., Dubs, F., … Rossi, J.-P. (2016). DNA barcoding reveals diversity patterns of earthworm communities in remote tropical forests of French Guiana. *Soil Biology and Biochemistry*, *92*, 171–183.

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, *24*(6), 332–340.

Dellicour, S., & Flot, J. (2018). The hitchhiker's guide to single-locus species delimitation. *Molecular Ecology Resources*.

Derouiche, L., Vercammen, P., Bouhadad, R., & Fernandes, C. (2017). Genetic evidence supporting the taxonomic separation of the Arabian and Northwest African subspecies of the desert hedgehog (*Paraechinus aethiopicus*). *Gene*, *620*, 54–65. doi: 10.1016/j.gene.2017.04.009

Dömel, J. S., Melzer, R. R., Harder, A. M., Mahon, A. R., & Leese, F. (2017). Nuclear and Mitochondrial Gene Data Support Recent Radiation within the Sea Spider Species Complex *Pallenopsis patagonica*. *Frontiers in Ecology and Evolution*, *4*, 139.

Ducasse, J., Ung, V., Lecointre, G., & Miralles, A. (2020). LIMES: a tool for comparing species partition. *Bioinformatics*, *36*(7), 2282–2283.

Elias-Gutierrez, M., Jeronimo, F. M., Ivanova, N. V., Valdez-Moreno, M., & Hebert, P. D. N. (2008). DNA barcodes for Cladocera and Copepoda from Mexico and Guatemala, highlights and new discoveries. *Zootaxa*, *1839*, 1–42.

568 Ence, D. D., & Carstens, B. C. (2011). SpedeSTEM: a rapid and accurate method for species
569 delimitation. *Molecular Ecology Resources*, *11*, 473–480.
570 Flot, J.-F., Couloux, A., & Tillier, S. (2010). Haplowebs as a graphical tool for delimiting species: a
571 revival of Doyle's" field for recombination" approach and its application to the coral genus
572 Pocillopora in Clipperton. *BMC Evolutionary Biology*, *10*(1), 1.
573 Fourdrilis, S., Mardulyn, P., Hardy, O. J., Jordaens, K., de Frias Martins, A. M., & Backeljau, T. (2016).
574 Mitochondrial DNA hyperdiversity and its potential causes in the marine periwinkle
575 *Melarhaphe neritoides* (Mollusca: Gastropoda). *PeerJ*, *4*, e2549.
576 Fujisawa, T., & Barraclough, T. G. (2013). Delimiting Species Using Single-locus Data and the
577 Generalized Mixed Yule Coalescent (GMYC) Approach: A Revised Method and Evaluation on
578 Simulated Datasets. *Systematic Biology, 62*, 707–724.
579 Fujita, M. K., Leaché, A. D., Burbrink, F. T., McGuire, J. A., & Moritz, C. (2012). Coalescent-based
580 species delimitation in an integrative taxonomy. *Trends in Ecology and Evolution*, *27*, 480–
581 488.
582 Hajibabaei, M., Janzen, D. H., Burns, J. M., Hallwachs, W., & Hebert, P. D. N. (2006). DNA barcodes
583 distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences*,
584 *103*(4), 968–971.
585 Jones, M., Ghoorah, A., & Blaxter, M. (2011). jMOTU and Taxonerator: turning DNA barcode
586 sequences into annotated operational taxonomic units. *PLoS ONE*, *6*, e19259.
587 Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A., & Flouri, T. (2017). Multi-rate
588 Poisson tree processes for single-locus species delimitation under maximum likelihood and
589 Markov chain Monte Carlo. *Bioinformatics*, *33*(11), 1630–1638.
590 Kekkonen, M., & Hebert, P. D. N. (2014). DNA barcode-based delineation of putative species: efficient
591 start for taxonomic workflows. *Molecular Ecology Resources*, *14*(4), 706–715. doi:
592 10.1111/1755-0998.12233

Kennedy, S. R., Prost, S., Overcast, I., Rominger, A. J., Gillespie, R. G., & Krehenwinkel, H. (2020). High-throughput sequencing for community analysis: the promise of DNA barcoding to uncover diversity, relatedness, abundances and interactions in spider communities. *Development Genes and Evolution*, 1–17.

Kerr, K. C. R., Stoeckle, M. Y., Dove, C. J., Weigt, L. A., Francis, C. M., & Hebert, P. D. N. (2007). Comprehensive DNA barcode coverage of North American birds. *Molecular Ecology Notes*, *7*, 535–543.

Lambert, A., & Stadler, T. (2013). Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theoretical Population Biology*, *90*, 113–128.

Larsen, B., & Aone, C. (1999). *Fast and effective text mining using linear-time document clustering*. 16–22.

Larson, E. R., Castelin, M., Williams, B. W., Olden, J. D., & Abbott, C. L. (2016). Phylogenetic species delimitation for crayfishes of the genus *Pacifastacus*. *PeerJ*, *4*, e1915.

Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP data. *Systematic Biology*, *63*(4), 534–542.

Leasi, F., & Norenburg, J. L. (2014). The necessity of DNA taxonomy to reveal cryptic diversity and spatial distribution of meiofauna, with a focus on Nemertea. *PLoS One*, *9*(8), e104385.

Leavitt, S. D., Moreau, C. S., & Lumbsch, H. T. (2015). The dynamic discipline of species delimitation: progress toward effectively recognizing species boundaries in natural populations. In *Recent Advances in Lichenology* (pp. 11–44). Springer.

Lin, X., Stur, E., & Ekrem, T. (2015). Exploring genetic divergence in a species-rich insect genus using 2790 DNA barcodes. *PloS One*, *10*(9), e0138993.

Lithanatudom, S. K., Chaowasku, T., Nantarat, N., Jaroenkit, T., Smith, D. R., & Lithanatudom, P. (2017). A First Phylogeny of the Genus *Dimocarpus* and Suggestions for Revision of Some Taxa Based on Molecular and Morphological Evidence. *Scientific Reports*, *7*(1), 6716.

618 Monaghan, M. T., Wild, R., Elliot, ., Fujisawa, T., Balke, M., Inward, D. J. G., … Vogler, A. P. (2009).
619 Accelerated species inventory on Madagascar using coalescent-based models of species
620 delineation. *Systematic Biology*, *58*, 298–311.
621 Nichols, R. (2001). Gene trees and species trees are not the same. *Trends in Ecology and Evolution*,
622 *16*(7), 358–364.
623 Ortiz, D., & Francke, O. F. (2016). Two DNA barcodes and morphology for multi-method species
624 delimitation in Bonnetina tarantulas (Araneae: Theraphosidae). *Molecular Phylogenetics and*
625 *Evolution*, *101*, 176–193.
626 Pante, E., Puillandre, N., Viricel, A., Arnaud-Haond, S., Aurelle, D., Castelin, M., … Samadi, S. (2015).
627 Species are hypotheses: avoid connectivity assessments based on pillars of sand. *Molecular*
628 *Ecology*, *24*(3), 525–544. doi: 10.1111/mec.13048
629 Papakostas, S., Michaloudi, E., Proios, K., Brehm, M., Verhage, L., Rota, J., … Fontaneto, D. (2016).
630 Integrative taxonomy recognizes evolutionary units despite widespread mitonuclear
631 discordance: evidence from a rotifer cryptic species complex. *Systematic Biology*, *65*(3), 508–
632 524.
633 Pentinsaari, M., Vos, R., & Mutanen, M. (2016). Algorithmic single-locus species delimitation: effects
634 of sampling effort, variation and nonmonophyly in four methods and 1870 species of beetles.
635 *Molecular Ecology Resources*.
636 Pino-Bodas, R., Burgaz, A. R., Teuvo, A., & Stenroos, S. (2018). Taxonomy of *Cladonia angustiloba* and
637 related species. *The Lichenologist*, *50*(3), 267–282.
638 Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., … Vogler, A. P.
639 (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects.
640 *Systematic Biology*, *55*, 595–609.
641 Puillandre, N., Cruaud, C., & Kantor, Y. I. (2010). Cryptic species in *Gemmuloborsonia* (Gastropoda:
642 Conoidea). *Journal of Molluscan Studies*, *73*, 11–23.

643 Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery
644 for primary species delimitation. *Molecular Ecology*, *21*, 1864–1877.
645 Puillandre, N., Baylac, M., Boisselier-Dubayle, M.-C., Cruaud, C., & Samadi, S. (2009). An integrative
646 approach to species delimitation in *Benthomangelia* (Mollusca: Conoidea). *Biological Journal*
647 *of the Linnean Society*, *96*(3), 696–708.
648 Puillandre, N., Fedosov, A. E., Zaharias, P., Aznar-Cormano, L., & Kantor, Y. I. (2017). A quest for the
649 lost types of *Lophiotoma* (Gastropoda: Conoidea: Turridae): integrative taxonomy in a
650 nomenclatural mess. *Zoological Journal of the Linnean Society*, *181*(2), 243–271.
651 Puillandre, N., Macpherson, E., Lambourdière, J., Cruaud, C., Boisselier-Dubayle, M.-C., & Samadi, S.
652 (2011). Barcoding type specimens helps to identify synonyms and an unnamed new species
653 in *Eumunida* Smith, 1883 (*Decapoda: Eumunididae*). *Invertebrate Systematics*, *25*(4), 322–
654 333. doi: 10.1071/IS11022
655 Puillandre, N., Modica, M.-V., Zhan, Y., Sirovich, L., Boisselier, M.-C., Cruaud, C., … Samadi, S. (2012).
656 Large-scale species delimitation method for hyperdiverse groups. *Molecular Ecology*, *21*(11),
657 2671–2691. doi: 10.1111/j.1365-294X.2012.05559.x
658 Rambaut, A., & Drummond, A. J. (2014). *Tracer v1.6*. Available from http://beast.bio.ed.ac.uk/Tracer.
659 Ramirez, J. L., Birindelli, J. L., Carvalho, D. C., Affonso, P. R., Venere, P. C., Ortega, H., … Galetti Jr, P.
660 M. (2017). Revealing Hidden Diversity of the Underestimated Neotropical Ichthyofauna: DNA
661 Barcoding in the Recently Described Genus *Megaleporinus* (Characiformes: Anostomidae).
662 *Frontiers in Genetics*, *8*, 149.
663 Rannala, B., & Yang, Z. (2020). Species delimitation. In *Phylogenetics in the Genomic Era*
664 (Scornavacca, C., Delsuc, F., Galtier, N.).
665 Rannala, B. (2015). The art and science of species delimitation. *Current Zoology*, *61*(5), 846–853.
666 Rannala, Bruce, & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral
667 population sizes using DNA sequences from multiple loci. *Genetics*, *164*(4), 1645–1656.

668 Ratnasingham, S., & Hebert, P. D. (2013). A DNA-based registry for all animal species: the Barcode
669 Index Number (BIN) system. *PloS One*, *8*(7), e66213.
670 Reid, N. M., & Carstens, B. C. (2012). Phylogenetic estimation error can decrease the accuracy of
671 species delimitation: a Bayesian implementation of the general mixed Yule-coalescent
672 model. *BMC Evolutionary Biology*, *12*(1), 196.
673 Ritchie, A. M., Lo, N., & Ho, S. Y. (2016). Examining the sensitivity of molecular species delimitations
674 to the choice of mitochondrial marker. *Organisms Diversity & Evolution*, 1–14.
675 Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source
676 tool for metagenomics. *PeerJ*, *4*, e2584.
677 Scarpa, F., Sanna, D., Cossu, P., Lai, T., Curini-Galletti, M., & Casu, M. (2017). A molecular approach to
678 the reconstruction of the pre-Lessepsian fauna of the Isthmus of Suez: the case of the
679 interstitial flatworm *Monocelis lineata* sensu lato (Platyhelminthes: Proseriata). *Journal of*
680 *Experimental Marine Biology and Ecology*.
681 Smith, M. A., Poyarkov Jr., N. A., & Hebert, P. D. N. (2008). CO1 DNA barcoding amphibians: take the
682 chance, meet the challenge. *Molecular Ecology Resources*, *8*, 235–246.
683 Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
684 thousands of taxa and mixed models. *Bioinformatics*, *22*, 2688–2690.
685 Tang, C. Q., Humphreys, A. M., Fontaneto, D., & Barraclough, T. G. (2014). Effects of phylogenetic
686 reconstruction method on the robustness of species delimitation using single-locus data.
687 *Methods in Ecology and Evolution*, *5*(10), 1086–1094.
688 Vacher, J.-P., Kok, P. J., Rodrigues, M. T., Lima, J. D., Lorenzini, A., Martinez, Q., … Gaucher, P. (2017).
689 Cryptic diversity in Amazonian frogs: Integrative taxonomy of the genus *Anomaloglossus*
690 (Amphibia: Anura: Aromobatidae) reveals a unique case of diversification within the Guiana
691 Shield. *Molecular Phylogenetics and Evolution*, *112*, 158–173.
692 Wakeley, J. (2009). *Coalescent theory: an introduction*.

693 Yang, Z., & Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple
694 loci. *Molecular Biology and Evolution*, msu279.
695 Zhang, J., Kapli, R., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with
696 applications to phylogenetic placments. *Bioinformatics*, *Advance Access*.
697 Zou, S., Fei, C., Song, J., Bao, Y., He, M., & Wang, C. (2016). Combining and comparing coalescent,
698 distance and character-based approaches for barcoding microalgaes: A Test with *Chlorella*-
699 like species (Chlorophyta). *PloS One*, *11*(4), e0153833.
700

DATA ACCESSIBILITY STATEMENT

ASAP is available at https://bioinfo.mnhn.fr/abi/public/asap. Data sharing is not applicable to this article as no new data were created or analyzed in this study.
The software used to simulate multispecies coalescent with random speciation time was written in C and is available upon request, as well as the simulated datasets. All real datasets are directly accessible from the ASAP website.

AUTHOR CONTRIBUTIONS

SB, GA and NP designed the method; GA developed the algorithm and tested it on simulated datasets; SB wrote the program and created the web-interface; NP performed the tests on real datasets; GA and NP wrote the manuscript.

TABLES

Table 1. Results of the analyses of the empirical datasets.

Dataset	Reference	#seq	#spec	ASAP 1st	ASAP 1st-2nd	ABGD	PTP	mPTP	GMYC	mGMYC
Benthomangelia	Puillandre et al. 2009	44	5	2/4/5	5	5	6	5	5	11
Gemmuloborsonia	Puillandre et al. 2010	80	5	5	5	5	5	5	5	8
Lophiotoma	Puillandre et al. 2017	276	10	9	10	9	17	13	10	12
Eumunida	Puillandre et al. 2011	127	16	16	16	16	18	16	16	24
Amphibians	Smith et al. 2008	339	39	20	37	38	44	33	38	49
Cladocera	Elias-Gutierrez et al. 2008	355	58	54	54	53	60	54	67	89
Mammals	Borisenko et al. 2008	521	73	66	66	76	73	55	80	95
Turridae	Puillandre et al. 2012	1,000	87	81	88	87	103	69	95	115
Sphingidae	Hajibabaei et al. 2006	989	107	107	107	98	135	105	140	159
Birds	Kerr et al. 2007	2,574	643	527	529	601	634	475	n.a.	n.a.

Each line represents a dataset which numbers of sequences (#seq) and species (#spec) are reported in the provided reference. We compare the "true" number of species to the predictions made by the partition ranked first by ASAP (ASAP 1st), by the "best" partition among the two first predicted by ASAP (ASAP 1st-2nd), the "best" partition by ABGD and the unique partition predicted by PTP, mPTP, GMYC and mGMYC. There is no partition for Birds by GMYC and mGMYC as we were not able to obtain a Bayesian tree given the large number of sequences. Cells were colored in dark grey when predictions were very accurate (at most 5% different from the referenced number of species) and with light grey when accurate (between 5% and 10%).

FIGURE CAPTION

Figure 1. An illustration of the clustering algorithm on a small dataset of nine sequences. On the lower part, we report how ASAP proceeds (downward in the figure) through the list of ranked distances (on the left), merging successively sequences into groups (highlighted in colored blocks). For each new group, ASAP computes a p-value that this new group is a panmictic species (values reported on the right part) based on pairwise differences within (intra) and between (inter) subgroups. Furthermore, each time a new group is created, a new partition is built (a sequence of blocks in the central part) that is associated to the current distance d_C. The distances d_C at which the partitions are instantiated are represented in a phenetic tree (top part). Each node is a group, each horizontal dashed line is a partition. For each newly created partition, ASAP also computes a probability of panmixia (p-val) and a relative gap width metrics (W). Then using their respective ranks (given in parenthesis), ASAP computes an ad-hoc ASAP-score: the lower the score, the better the partition.

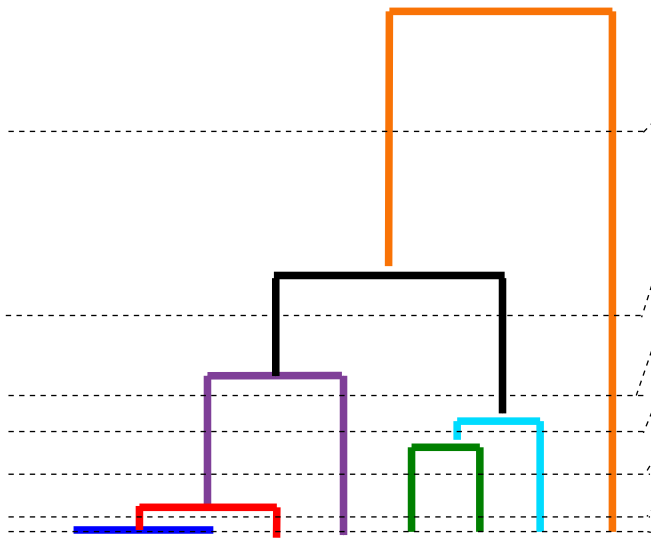
Figure 2. The computation time of ASAP as a function of the number of species. Illustrating the linear relationship, we estimate that on the current webserver, computation time is seconds 0.07 seconds per species.

Figure 3. Performance of ASAP as a function of the speciation rate. For two alternative models of speciation (Radiation and Yule), we report the fraction of replicates where ASAP find the four correct species (top panels). We considered either only the partition with the best *asap-score* (ASAP-1) or the partitions ranked first and second (ASAP-1/2). Obviously, the later has better performance. We also report the average number of predicted species, regardless they are correct or not (bottom panels). Each point is evaluated on 500 replicates.

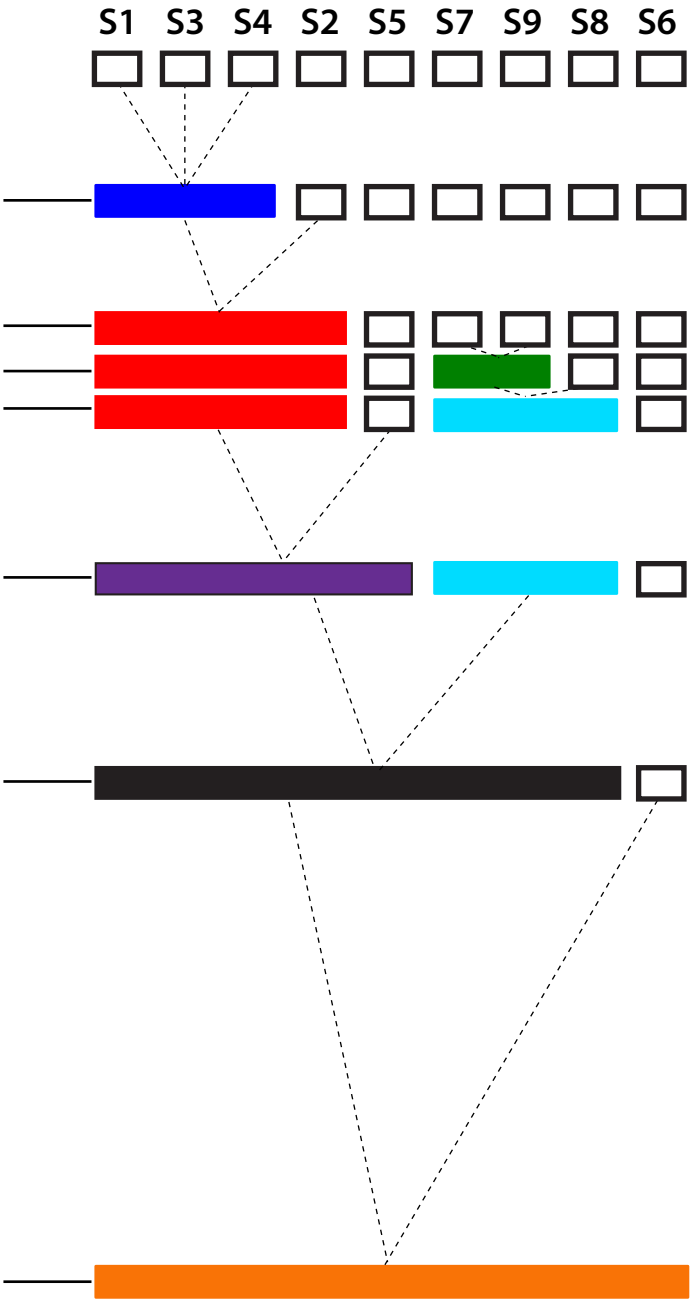
Figure 4. Power of ASAP, ABGD, PTP and GMYC to predict the correct number of species among 200 sequences. We vary the number of true species from 4 to 60 in the Radiation and in the Yule model. Each point is an average of 500 replicates and vertical error bars mark the standard deviation.

Supplementary Material 1: Computation of the relative barcode gap width.

Supplementary Material 2: Graphical output of ASAP.

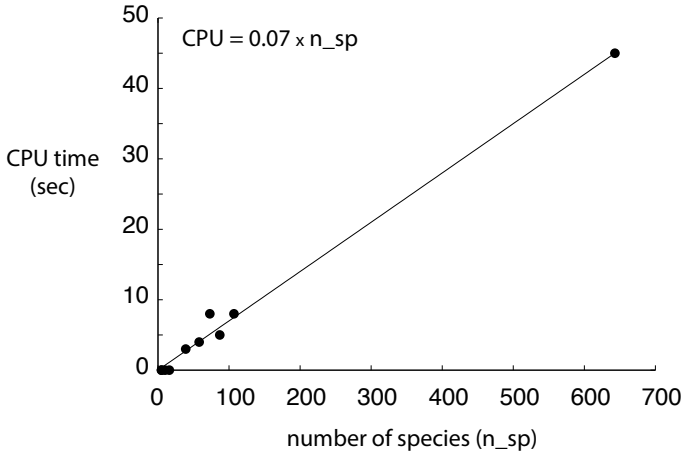


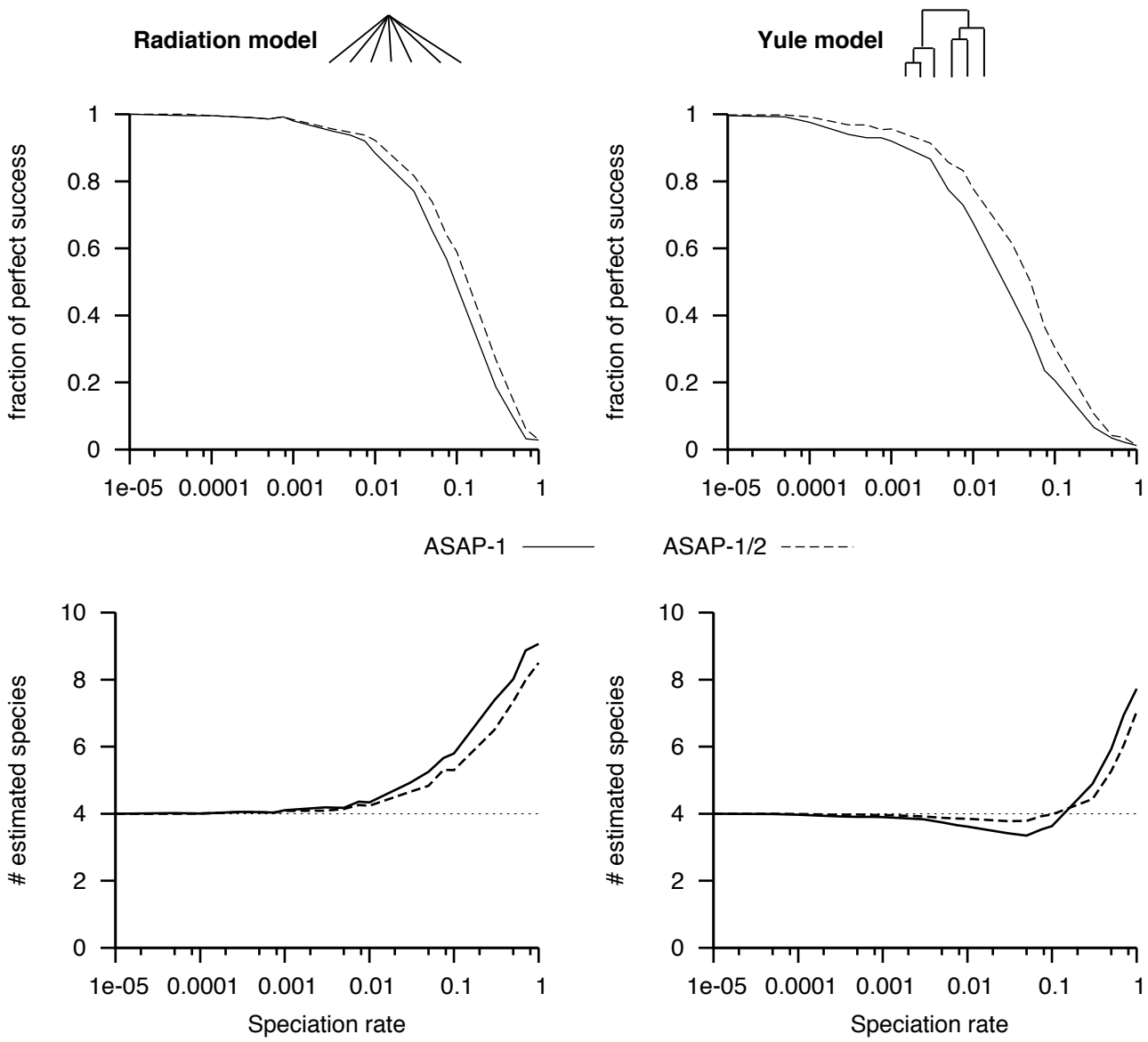
For the partitions			
d_T	p-val	Wx100	asap score
0.1301	0.76 (2)	0.57 (2)	2
0.0665	0.76 (3)	0.28 (7)	5
0.0405	0.22 (1)	0.57 (3)	2
0.0289	0.90 (5)	1.11 (1)	3
0.0231	1.00 (6)	0.50 (4)	5
0.0058	0.81 (4)	0.32 (5)	4.5
0.0029	1.00 (6)	0.09 (6)	6

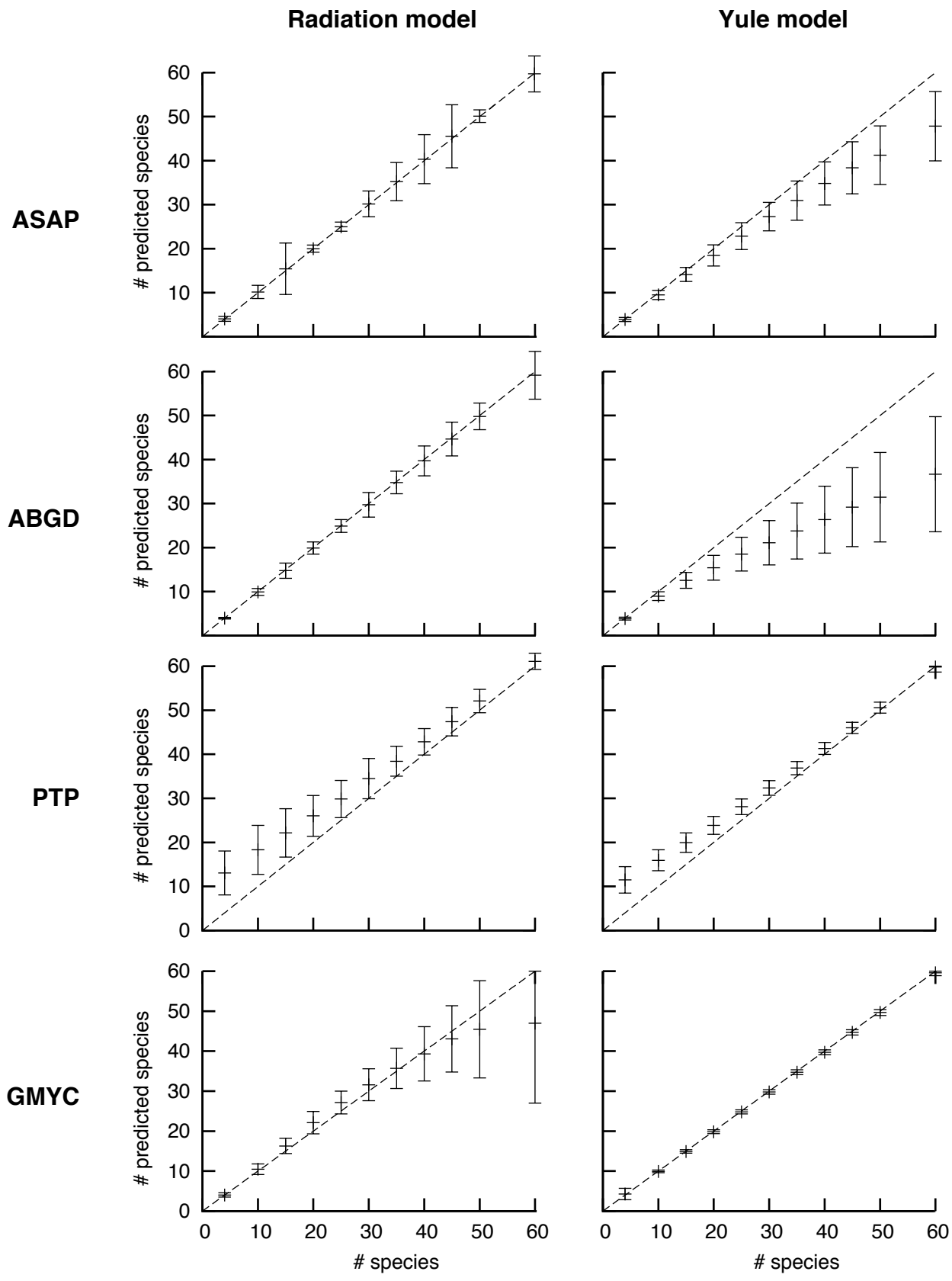


ranked distances		
1	S1,S3	0.0029
2	S1,S4	0.0029
3	S3,S4	0.0029
4	S1,S2	0.0058
5	S2,S3	0.0058
6	S2,S4	0.0058
7	S7,S9	0.0231
8	S7,S8	0.0289
9	S1,S5	0.0405
10	S3,S5	0.0405
11	S4,S5	0.0405
12	S8,S9	0.0405
13	S2,S5	0.0434
14	S1,S7	0.0665
15	S3,S7	0.0665
16	S4,S7	0.0665
17	S5,S7	0.0665
18	S2,S7	0.0694
19	S1,S9	0.0723
20	S3,S9	0.0723
21	S4,S9	0.0723
22	S5,S9	0.0723
23	S2,S9	0.0751
24	S5,S8	0.0751
25	S1,S8	0.0809
26	S3,S8	0.0809
27	S4,S8	0.0809
28	S2,S8	0.0838
29	S6,S7	0.1301
30	S6,S9	0.1329
31	S1,S6	0.1416
32	S3,S6	0.1416
33	S4,S6	0.1416
34	S2,S6	0.1445
35	S5,S6	0.1445
36	S6,S8	0.1474

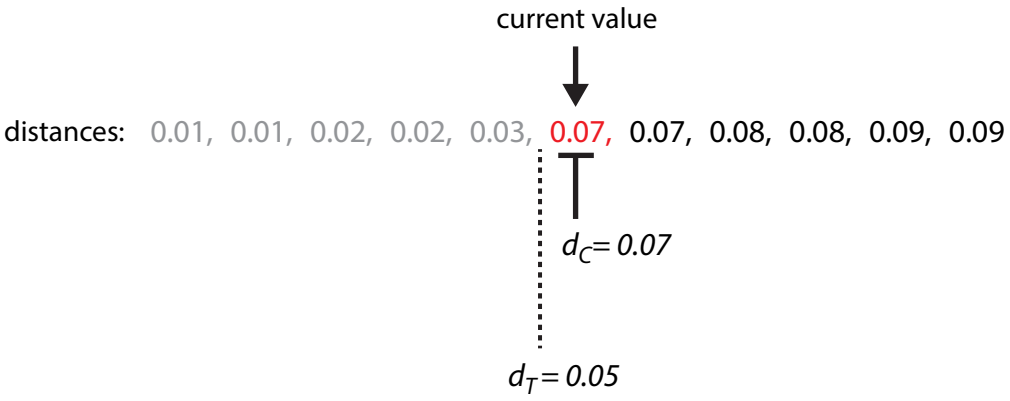
For the nodes		
π_{intra}	π_{inter}	p-val
-	0.0029	n/a
0.0029	0.0058	0.81
-	0.0231	n/a
0.0231	0.0347	0.90
0.0044	0.0412	0.22
0.0218	0.0734	0.76
0.0495	0.1409	0.76


CPU = 0.07 x n_sp
CPU time
(sec)
50
40
30
20
10
0
0
100
200
300
400
500
600
700
number of species (n_sp)

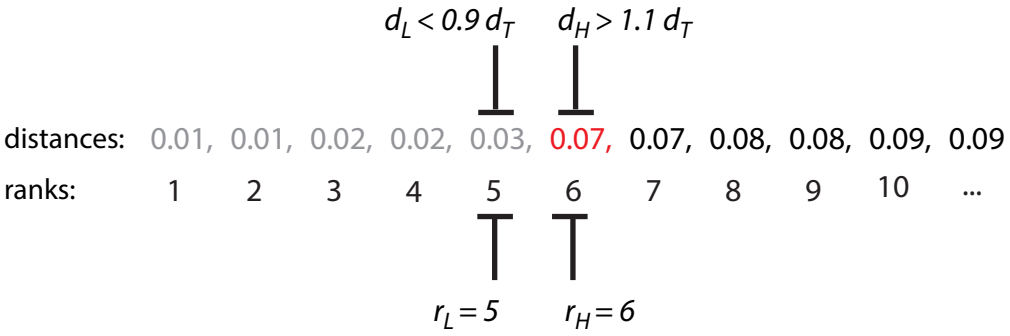

Radiation model
Yule model
fraction of perfect success
1
0.8
0.6
0.4
0.2
0
1e-05
0.0001
0.001
0.01
0.1
1
ASAP-1
ASAP-1/2
estimated species
10
8
6
4
2
0
Speciation rate


Radiation model
Yule model
ASAP
ABGD
PTP
GMYC
predicted species
species
0
10
20
30
40
50
60

Step 1 - computing d_C and d_T



Step 2 - Finding r_L and r_H



Step 3 - Computing W

$$W = [(d_H - d_L) / (d_H + d_L + 1)] / (r_H - r_L)$$
$$= [(0.07 - 0.03) / (0.07 + 0.03 + 1)] / (6 - 5)$$
$$= 0.036$$

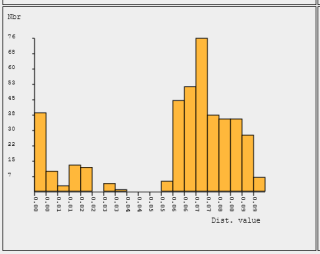
ASAP Web results using the JC69 Jukes-Cantor model for 29 sequences in the file Dataset pour nouvelle figure.fas

Split probability: 0.010000 sequence length: 658

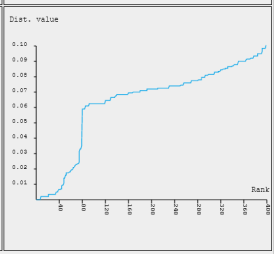
10 best partitions found by ASAP (see FAQ for more details)

Nb of groups	ASAP score	Proba/ranking	Slope/ranking	Max Dist. for grouping	Text
* 7	1.50	1.16e-01 (1)	2.001266e-01 (2)	0.024263	list csv
* 5	3.00	1.60e-01 (2)	1.048301e-01 (4)	0.046165	list csv
* 8	4.00	3.41e-01 (3)	3.865965e-02 (5)	0.015651	list csv
* 9	4.50	6.87e-01 (8)	2.010442e-01 (1)	0.010307	list csv
* 10	5.00	3.93e-01 (4)	2.325606e-02 (6)	0.005104	list csv
23	7.50	4.81e-01 (6)	1.762736e-02 (9)	0.000761	list csv
15	8.50	7.50e-01 (9)	2.133008e-02 (8)	0.002371	list csv
12	9.00	4.77e-01 (5)	5.378038e-03 (13)	0.003255	list csv
4	9.00	9.74e-01 (11)	2.259335e-02 (7)	0.059045	list csv
* 6	9.00	1.00e+00 (15)	1.787349e-01 (3)	0.032616	list csv

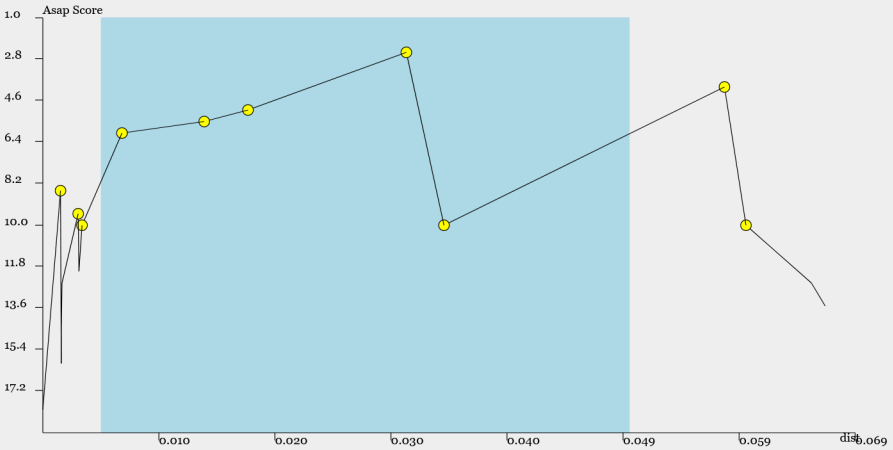
Histogram of distances [save]



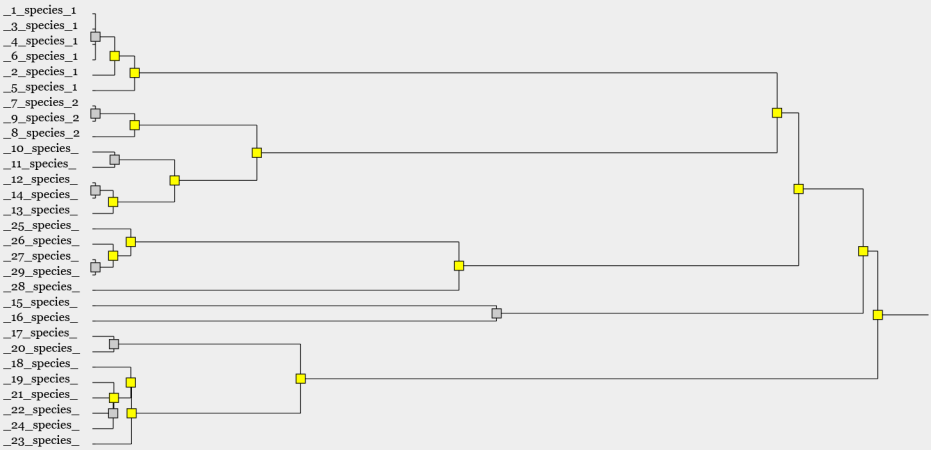
Ranked distances [save]



View/Save Boxed species graph here



Legend: <0.001 <0.05 <0.1 >0.1 N/A



View/save curves and dendrogram here

Time for running: 0 min 1 seconds