Factors affecting species delimitations with the GMYC model: insights from a butterfly survey

Gerard Talavera¹*, Vlad Dinca¹,² and Roger Vila¹*

¹Institut de Biologia Evolutiva, CSIC-Universitat Pompeu Fabra, Passeig Marítim de la Barceloneta, 37, 08003 Barcelona, Spain; and ²Department of Zoology, Stockholm University, 10 691 Stockholm, Sweden

Summary

1. The generalized mixed Yule-coalescent (GMYC) model has become one of the most popular approaches for species delimitation based on single-locus data, and it is widely used in biodiversity assessments and phylogenetic community ecology. We here examine an array of factors affecting GMYC resolution (tree reconstruction method, taxon sampling coverage/taxon richness and geographic sampling intensity/geographic scale).

2. We test GMYC performance based on empirical data (DNA barcoding of the Romanian butterflies) on a solid taxonomic framework (i.e. all species are thought to be described and can be determined with independent sources of evidence). The data set is comprehensive (176 species), and intensely and homogeneously sampled (1303 samples representing the main populations of butterflies in this country). Taxonomy was assessed based on morphology, including linear and geometric morphometry when needed.

3. The number of GMYC entities obtained constantly exceeds the total number of morphospecies in the data set. We show that c. 80% of the species studied are recognized as entities by GMYC. Interestingly, we show that this percentage is practically the maximum that a single-threshold method can provide for this data set. Thus, the c. 20% of failures are attributable to intrinsic properties of the COI polymorphism: overlap in inter- and intra-specific divergences and non-monophyly of the species likely because of introgression or lack of independent lineage sorting.

4. Our results demonstrate that this method is remarkably stable under a wide array of circumstances, including most phylogenetic reconstruction methods, high singleton presence (up to 95%), taxon richness (above five species) and the presence of gaps in intraspecific sampling coverage (removal of intermediate haplotypes). Hence, the method is useful to designate an optimal divergence threshold in an objective manner and to pinpoint potential cryptic species that are worth being studied in detail. However, the existence of a substantial percentage of species wrongly delimited indicates that GMYC cannot be used as sufficient evidence for evaluating the specific status of particular cases without additional data.

5. Finally, we provide a set of guidelines to maximize efficiency in GMYC analyses and discuss the range of studies that can take advantage of the method.

Key-words: cryptic species, DNA barcoding, generalized mixed Yule-coalescent model, Lepidoptera, phylogenetic community ecology, phylogenetic systematics, sampling strategy, species delimitation

Introduction

DNA techniques have provided hope to accelerate global biodiversity exploration by massively sequencing living organisms. Two alternative strategies are generally followed: massive single-locus sequencing projects allow the fast discovery of unknown diversity, while comprehensive genomic approaches enable a finer way to study already known organisms. Although it is generally agreed that multiple genetic markers allowing coalescence-based studies are optimal for studying species boundaries when combined with independent sources of evidence (morphological, ecological, karyological, etc.) (Rubinoff & Holland 2005; Will, Mishler & Wheeler 2005; Knowles & Carstens 2007), the single-locus modus operandi usually allows broader surveying and has flourished in recent years mainly due to DNA barcoding (Hebert, Cywinska & Ball 2003). Although DNA barcoding was not conceived as a tool to delimit species, it established the base for derived DNA-based methods for species delimitation (Davis & Nixon 1992; Pons et al. 2006; Cummings, Neel & Shaw 2008; Masters, Fan & Ross 2011). The use of the generalized mixed Yule-coalescent (GMYC) model (Pons et al. 2006; Fontaneto et al. 2007; Fujisawa & Barraclough 2013) has become one of the most popular approaches for species delimitation based on phylogenetic data. This method does not require any previous information about species and has been specifically developed for single-locus data, making it particularly promising for
phylogenetic community ecology studies or to explore groups with uncertain taxonomy. The GMYC method classifies branches in a gene tree as intra- or interspecific by maximizing the likelihood of a GMYC evolution model. Branching events between species are modelled with a Yule model, that is, assuming a constant speciation rate and no extinction (Nee, May & Harvey 1994; Barraclough & Nee 2001), and branching events within species are modelled using a neutral coalescent process (Hudson 1990).

While GMYC species delimitation has been broadly used in a variety of organisms, the performance of the method needs more empirical evaluation on a solid taxonomic framework. Most of the work performed to evaluate GMYC performance has been based on simulations: Papadopoulou et al. (2008) and Lohse (2009) contributed with interesting discussion regarding the influence of the sampling scheme, Esselstyn et al. (2012) assessed the significance of different coalescent parameters on individual gene trees, Reid & Carstens (2012) tested the effect of tree depth, the importance of allele sampling within species and the relevance of the DNA fragment length, and Fujisawa & Barraclough (2013) fully evaluated alternative scenarios of patterns of population variation and divergence between species. Although some insights have been obtained from empirical studies (e.g. Pons et al. 2011; Esselstyn et al. 2012), the nature and magnitude of many potential biases have not been evaluated in real cases. Ideally, a data set to test species delimitation performance needs to be extensive (including a substantial number of species), homogeneous (sampling intensity and range should be as similar as possible across taxa) and taxonomy needs to be resolved and assessed independently (to avoid circular reasoning). Thus, we decided to specifically address this question using a highly suitable empirical data set. This consisted of a complete and fairly homogeneous DNA barcode library for the butterfly fauna of an entire country (Romania) (Dincă et al. 2011a). This country includes about one-third of the European butterfly fauna. Since the European butterflies are arguably the best studied invertebrates in the world, the detailed morphological study of all 1387 samples (180 species) in Dincă et al. (2011a), including external and internal characters and linear and geometric morphometry when necessary, makes the taxonomy highly precise for this particular data set.

Whereas some factors that may affect GMYC performance can be controlled by the researcher, some others are characteristics of the system under study. Moreover, the limitations of the genetic marker used determine the maximum percentage of potentially recognized species. We focused on three aspects potentially shaping the branching pattern of the tree and likely to be determinant in GMYC results: (1) tree reconstruction method, (2) taxon sampling coverage/taxon level and (3) geographic sampling intensity/geographic scale. First, tree reconstruction and ultrametrization algorithms based on evolutionary models produce differences in relative branch length estimations. GMYC users typically choose less time-consuming programmes like RAXML ( Stamatakis 2006) plus a subsequent branch length transformation with PATHd8 (Britton et al. 2007) or r8s (Sanderson 2003). Software directly producing ultrametric trees (e.g. BEAST) avoids intermediate steps, but it is computationally demanding for large data sets, thus limiting applicability (Puillandre et al. 2012). Secondly, variability in rates of genetic change, more common at wider taxon level, might affect species delimitation. GMYC has been indistinctly used at several taxonomical levels, from an entire phylum (Barraclough et al. 2009), or a superfamily (Monaghan et al. 2009), to species complexes (Fontaneto, Boschetti & Ricci 2007; Papadopoulou et al. 2009a,b; Gebiola et al. 2012). Lastly, the sampling intensity across a given territory has been discussed as a limitation for the GMYC approach (Lohse 2009; Papadopoulou et al. 2009a,b; Pons et al. 2011; Reid & Carstens 2012). While it has been reported, based on simulations, that a complete representation of the genetic population structure within the study region for each taxon might be essential to avoid artifactual delimitations, mainly in terms of oversplitting (Lohse 2009), others argued that in real data sets this is not such a major concern (Papadopoulou et al. 2009a,b; Reid & Carstens 2012). Moreover, increasing the geographic scale of the study could result in oversplitting as well because intraspecific variation increases and interspecific divergence decreases (Bergsten et al. 2012). Based on the Romanian butterfly data set, we study the GMYC accuracy and evaluate the consequences of these factors. With the evidence obtained, we provide new insights into the proper conditions for the efficient application of the GMYC method.

Materials and methods

Methods are described in greater detail with full references in the Supporting Information.

DATA SET CHARACTERISTICS AND PHYLOGENETIC APPROACHES

Our data set consisted on a 634 base pair COI alignment of 1303 specimens representing 176 butterfly species extracted from Dincă et al. (2011a). Maximum likelihood (ML) and neighbour joining (NJ) ultrametric phylogenetic trees were obtained following different strategies in r8s (Sanderson 2003), PATHd8 (Britton et al. 2007), ChronoPL and chronos functions in the APE package in R (Paradis, Claude & Strimmer 2004) and Bayesian inference in BEAST 1.6.0 (Drummond & Rambaut 2007). The GMYC species delimitation tool (Pons et al. 2006; Fontaneto et al. 2007; Monaghan et al. 2009) was tested using the splits package (Ezard, Fujisawa & Barraclough 2009) implemented in a statistical software for the single- and multiple-threshold options. The multimodel GMYC approach (Powell 2012) to account for delimitation uncertainty was applied to one of the best performing phylogenetic approaches (BEAST with coalescent prior + strict clock). Unless otherwise stated, all the analyses were conducted on haplotype trees with no repeated haplotypes.

SUBCLADE TESTS ON GMYC SIGNIFICANCE AND PERFORMANCE

The GMYC is a method to objectively establish a divergence threshold to delimit species in a phylogenetic tree. To do so, it uses a likelihood approach to analyse the timing of branching events and detects significant switches between a Yule (interspecific) and a coalescent
(intraspecific) branching structure. To evaluate the behaviour of the significance of the GMYC versus the null model (no switches) when the number of species was reduced, all possible subclades within family-level phylogenetic trees were analysed. For each subclade showing significance, the GMYC performance to identify species was subsequently assessed under the single-threshold approach and compared with results for that group of species in the family tree.

**INTRASPECIFIC SAMPLING TESTS**

Taxon sampling coverage effects were studied following three approaches. First, an intraspecific gap was created for each of the 18 species with intraspecific genetic distances >0.5% that were correctly recovered as entities by GMYC by retaining only the two most genetically distant haplotypes in the data set. Secondly, we forced an unbalanced sampling by enlarging the data set with samples from outside the study area (Romania) for two species with available sequence data: *Polyommatus icarus* (Lycaenidae) and *Papilio machaon* (Papilionidae) (Table S1). Subsequently, the total number of sequences was randomly reduced to 50%, 20% and 10% (10 replicates each) for each species, and new tree inferences were carried out with BEAST (coalescent prior + strict clock). Lastly, to simulate possible effects derived from low intraspecific variability or poor sampling, we progressively increased the percentage of singletons (species represented by a single haplotype) in 40%, 50%, 70%, 90%, 95% and 100% of the 176 species (10 replicates each by random selection of the species and one of their haplotypes). The original data set had 48 (27%) species with a single haplotype and, as a control we also evaluated a data set of 128 species after excluding all singletons.

**Results**

**PHYLOGENETIC METHODS PERFORMANCE**

The number of resulting GMYC entities was similar (from 180 to 188) for most combinations of phylogenetic methods explored (Fig. 1a), with confidence intervals (CI) ranging from 178 to 191 (Table S2). The number of correctly identified species ranged between 140 (79.54%) and 143 (81.25%) from a total of 176 (Fig. 1a). However, three cases recovered notably poor numbers of correctly identified morphospecies: a distance-based NJ tree + Pathd8 (40.9%) and ML trees when the APE functions chronos (34.1%) and ChronoPL (55.7%) were used (Fig. 1a). The percentage of GMYC success was close to 80% for all the family trees, with the exception of the Papilionidae (100% in this case). No substantial differences were detected between Bayesian models using a coalescent or a Yule prior, and between a relaxed or strict molecular clock (Table S3, Fig. 1b). Generally, methods had problems in delimiting the same set of species, and the same effects for each taxon were recurrently observed. Twenty species were typically recovered as conspecific (10 cases of

![Fig. 1.](image-url)
lumping pairs), and 16 species were split (into pairs except for a triplet; Table S4).

GMYC performance when using sequence trees instead of haplotype trees produced almost identical values (Fig. 1c, Table S5). In all the approaches, the GMYC multiple threshold did not fit the data significantly better than the single-threshold option (Tables S2, S3 and S5). The inferred single threshold minimized split + lump percentages and was near to the maximum species identification performance, although it did not exactly reach the optimum for the data set. Indeed, the function of the success percentage displayed a rather flat plateau (Fig. 2). The multimodel GMYC approach (Powell 2012) provided the best-fitting model based on AIC criteria (method: multiple; AICc: -8269.386; entities: 189) and probabilities associated with each GMYC cluster (Table S6, Fig. S1).

SENSITIVITY TO DATA SET CHARACTERISTICS

Testing GMYC significance

We analysed all 298 possible subclades nested within family trees, and 65 clades (21.8%) produced significance for GMYC (likelihood ratio test P-value < 0.05). While results were slightly variable according to the taxa involved, a clear pattern of GMYC significance decrease was observed with reduced number of species and depth of the subclades (Fig. 3, Fig. S2). Subclades with less than three to five species were not significant (depending on the taxa involved), and none with less than six species was strongly significant (P < 0.001). GMYC was significant up to a 95% of singletons (Fig. 4). At this percentage, LRT P-values abruptly increased and only six of the ten replicates analysed achieved GMYC significance.

GMYC performance at lower taxon levels

Of the total 65 significant subclades in the family trees (Fig. 3), only 12 (18.5%) presented discordancess in the inferred number of GMYC entities when compared to the same clade in the entire tree (Fig. S2, Table S7). The observed differences generally involved one or very few taxa [usually of the Carcharodus clade (Hesperidae) and the Brenthis + Argynnis clade (Nymphalidae)], they were not necessarily related to the level of GMYC significance, and no clear tendency of splitting or lumping was observed.

Removal of intermediate haplotypes

The removal of intermediate haplotypes for the taxa with the highest intraspecific divergence (p-distance >0.5%) among those that were correctly recovered as single entities did not cause any changes to GMYC estimations (Table S8). Thus, it was meaningless analysing species with lower p-distance.

Unbalanced geographical range sampling

GMYC results were different in the two cases studied. For Polyommatus icarus, a maximum of five GMYC entities were recovered when a total of 110 specimens (44 haplotypes) from a much large geographical area than the rest of the data set were used (Table S1). The number of GMYC entities progressively decreased when sequences were randomly removed (Fig. 5). Such an effect was not detected in the case of Papilio machaon, where a single entity was recovered in all the analyses. However, for this species (25 specimens, 16 haplotypes), the area covered and the increase in maximum intraspecific divergence were much smaller than in the case of P. icarus: from a p-distance of 1.6% within Romania to 1.9% in the extended data set for P. machaon, compared with an increase from 1.1% to 3.5% for P. icarus.

Increasing the percentage of species with unique haplotype

An increase in the singleton percentage supposed a slight progressive improvement in the percentage of GMYC success (Fig. 4a). Note, however, that the challenge for the method decreased as higher proportions of singletons reduced the possibilities of splitting. A tree without singletons (excluding the original 48 singletons) was also evaluated, resulting in a GMYC performance (83%) similar to that of the initial data set.

Discussion

WHAT IS THE GMYC SUCCESS RATE?

Overall, the GMYC threshold practically reached an optimal among the possible range of single-threshold delimitations (Fig. 2). Thus, the method performs well in establishing a threshold that recovers the maximum correct information, but the particularities of the data set severely limit performance. A certain GMYC lumping tendency has been proven by simulation tests and attributed to a poor ability of the model to identify incomplete lineage sorting or clades undergoing rapid radiation (Esselstyn et al. 2012; Reid & Carstens 2012). In our data set, no such tendency was observed. On the contrary, the
total number of entities estimated by GMYC (180–188; CI: 178–191) was constantly higher than the number of morphospecies in the data set (176).

Cases of oversplitting may represent either real failures of GMYC or gaps in our taxonomical knowledge. Indeed, it has been claimed that GMYC is a tool to identify potential cryptic

Fig. 3. Generalized mixed Yule-coalescent (GMYC) significance was tested for all possible subclades within family haplotype trees (left column). Nodes are labelled according to the level of GMYC significance (no symbol = \( P > 0.05 \), * = \( P < 0.05 \), ** = \( P < 0.01 \), *** = \( P < 0.001 \)) obtained for the subclade that they define. The black part of the trees represents the subclades where differences between Yule and coalescent portions are not significant and are thus not suitable for being analysed independently under the GMYC model. Plots show GMYC significance (likelihood ratio test \( P \)-value) vs. the number of species included in the subclades (central column) and vs. the subclade tree depth (average divergence from root to tips) (right column).

Fig. 4. (a) Number of generalized mixed Yule-coalescent (GMYC) entities, number of correct hits and percentage for different percentage (0%, 27%, 40%, 50%, 70%, 90% and 95%) of singletons (species represented by a single haplotype). Horizontal grey lines represent the correct number of species. (b) GMYC significance (likelihood ratio test \( P \)-value) vs. percentage of singletons.

species when morphospecies are divided into two or more GMYC entities (Fontaneto et al. 2009; Ceccarelli, Sharkey & Zaldívar-Riverón 2012). Although the European butterfly fauna has been studied in depth, several unexpected discoveries of cryptic species have been recently reported (e.g. Nazari & Sperling 2007; Dapporto 2010; Dinca, Dapporto & Vila 2011; Dinca et al. 2011b). While it is not impossible that a few of the 16 cases of splitting may actually represent cryptic species, there is currently no relevant additional data supporting this hypothesis, and we consider them as failures of the GMYC method due to pronounced intraspecific genetic variability.

Several cases of failure stem from non-monophyly (c. 6%), some of which constitute suspected cases of introgression (Lysandra coridon – L. bellargus and P. napi – P. bryoniae). For the reasons mentioned and based on Fig. 2, it is evident that most cases of failure stem from the data set itself. Leaving aside data set limitations, the GMYC approach can represent an adequate tool to objectively select an optimal divergence threshold for roughly assessing biodiversity based exclusively on DNA sequence data and in the absence of taxonomical knowledge. It is worth noting that most studies use an arbitrary divergence threshold (a usual one is a 2% for DNA barcoding studies; Hebert, Ratnasingham & deWaard 2003; Hebert et al. 2004; Hajibabaei et al. 2006; Mutanen et al. 2012), but this may not be adequate to all taxon groups and data sets.

The best possible GMYC result is determined in each data set by the number of cases of non-monophyly plus the degree of overlap between intra- and interspecific divergences. The multiple-threshold model could in theory overcome part of the data set limitations. However, we find that the multiple-threshold model never fits our data set significantly better than the single-threshold one. This could be explained because the cases of lumping and splitting are highly mixed in the tree, among other reasons (Fujisawa & Barraclough 2013). Indeed, our cases of lumping and splitting are scattered along the tree, and both tendencies occur in closely related taxa. For example, the subfamily Satyrinae, and the genus Hipparchia in particular, include cases of both lumping and splitting. This phenomenon is expected to be common in many data sets and may be the main reason for the limited applicability of the multiple-threshold model.

**EFFECTS OF PHYLOGENETIC RECONSTRUCTION ON GMYC PERFORMANCE**

**Relevance of the method of obtaining ultrametric trees**

Optimization of branch lengths in a phylogenetic tree to convert it to ultrametric is a prerequisite for the GMYC method, but this process can be computationally demanding for large data sets. As a consequence, the fastest methods to obtain ultrametric trees are most frequently used, as for example the algorithms implemented in the software PATHd8. In our study, BEAST inferences were the most computationally demanding, followed by Garli ML inferences + PL in r8s. Most of the fastest inferences produced poorer results, but success percentages of the RAxML + PATHd8 algorithm were comparable to
those of BEAST genealogies, and this would be the preferred choice among the ones tested if time or computation capacity represents constraints, as is frequently the case. Some studies recommend BEAST as input for GMYC and show that the performance of the coalescent tree prior is better than that of the Yule prior (Monaghan et al. 2009) at least when the analysis is based on the COI marker (Ceccarelli, Sharkey & Zaldívar-Riverón 2012). Given that the GMYC uses coalescence as a null model, the coalescent tree prior is considered a more adequate option and appears to fit better the majority of the data sets in model comparisons (Monaghan et al. 2009). Regarding the molecular clock models, Monaghan et al. (2009) observed in their data set that a relaxed molecular clock with Yule prior resulted in a greater number of GMYC entities than other methods. Surprisingly, we observed very small differences among the different BEAST options (Tables S2, S3 and S5).

All specimens or haplotype collapsing?

Collapsing sequences to haplotypes is a common approach before running GMYC, as the model cannot handle polytomies and zero-length terminal branches. However, when a genealogy-based inference approach is employed, such as BEAST, identical sequences will be treated as different alleles coalescing back to their most recent common ancestor, which will insert non-zero branch lengths possibly affecting GMYC calculations. For our data set, the GMYC performance of genealogies obtained using the models implemented in the package BEAST was not found to be different from that of the collapsed data sets (Fig. 1c, Table S5). For this reason, the use of less demanding methods would be preferred, since large-scale biodiversity surveys may include vast numbers of specimens and the inference of BEAST genealogies will exponentially increase computational time when not removing duplicate haplotypes.

Effects of data set characteristics on GMYC performance

Imbalances between the Yule and coalescent portions of the tree may be crucial for the performance of the method and will depend on each specific data set. Several factors can shape the internal branching pattern favouring one portion or the other. Our empirical results show that both GMYC significance and performance are remarkably stable over a wide array of conditions, but results differ significantly from morphology-based species delimitation.

Taxon level effects

Scientists’ interests in biodiversity range from detailed studies of particular species-groups, to large-scale biodiversity surveys of all the taxa occurring in a specific area, such as in phylogenetic community ecology. The inclusion of sister taxa will have a lot of relevance because narrowing the interspecific divergences may potentially lead to underestimations. On the other hand, deep taxon level surveys are, for example, under the risk for including strong heterogeneity in speciation rates or in intraspecific population structure across taxa. Our data set is in an intermediate position between the two extremes discussed: It represents a rather diverse superfamily survey for a substantial geographical area and includes both deeply diverged lineages and extremely recent speciation events. We show that a surprisingly low number of species is required to produce GMYC significance (a minimum of between three and five depending on the taxa). Furthermore, within the boundaries imposed by GMYC significance, taxon level does not have a strong influence on species recognition. However, a wide applicability of the method with regard to taxon level in our case does not imply that all data sets are adequate, and care should be taken in small data sets (e.g. Fontaneto, Boschetti & Ricci 2007; Papadopoulou et al. 2009a,b; Gebiola et al. 2012). In these cases, adding outgroup taxa will increase the Yule portion of the tree and will presumably help balancing the Yule-coalescent equilibrium.

Impact of singleton percentage

Singletons in a data set (a single haplotype for a species) may be due to either a single specimen collected or multiple specimens without genetic variation. Our data set illustrates this: 48 species (27%) are represented by singletons. Of these, only four are due to a single specimen collected, and the other 44 are due to several specimens that share the same haplotype. Thus, a proportion of lineages will be singletons according to the intensity and success of the collecting work, but also to the genetic structure of the taxa within the area studied. The higher the number of singletons, the lower the coalescent portion in the tree providing information to the GMYC model (Lim, Balke & Meier 2011). However, some studies have found a good GMYC performance with proportions of singletons close to 60% (Monaghan et al. 2009; Ceccarelli, Sharkey & Zaldívar-Riverón 2012; Fujisawa & Barraclough 2013). Based on simulations, Reid & Carstens (2012) concluded that GMYC works efficiently with singletons as long as other taxa better represent the Yule-coalescent equilibrium.

Importance of intermediate haplotypes

Lohse (2009) suggested that undersampling may oversplit entities in GMYC analysis when <20% of demes are sampled. Our results indicate that intermediate haplotypes of a given species (those that are not the most diverged pair) are basically irrelevant, in the sense that their removal does not produce a split of that species. The test was performed with the species that displayed intraspecific divergences nearest to...
the GMYC divergence threshold, which are those theoretically most susceptible to this bias. This suggests that gaps in intraspecific sampling for a few species are rather unimportant as long as they do not affect the maximum intraspecific divergence. Thus, a sampling strategy directed to obtain the most divergent populations might be a good approach for eventually applying GMYC, especially when collecting complete series of demes is practically impossible. It is worth noting that the generalized presence of gaps in the data set might however influence the Yule-coalescent threshold estimated by GMYC.

A suitable number of samples per species cannot be simply a function of the size of the region studied, since the genetic richness and structure within the surveyed area are important as well (Zhang et al. 2010). Because butterflies are a group with relatively high dispersal rates, one may ask, for example, what happens in organisms with higher levels of population isolation or specialization, or displaying low population sizes (see Fujisawa & Barr aclough 2013). In this context, plotting the number of haplotypes versus the number of specimens or populations for each species (Fig. S3) may help evaluating the sampling effort with respect to the area and

---

### Table 1. Summary of factors tested, results obtained and practical recommendations for biodiversity studies employing the generalized mixed Yule-coalescent (GMYC) model

<table>
<thead>
<tr>
<th>Factors tested</th>
<th>Results</th>
<th>Practical recommendations for GMYC studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylogenetic methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree inference method</td>
<td>Similar results for all ML and Bayesian methods tested. NJ leads to poor results. Pathd8 and r8s produce similar results. ChronoPL and chronos work substantially worse.</td>
<td>Not necessary to test several options. Do not use NJ. Rapid algorithms such as PATHd8 are recommended to accelerate computing. Do not use ChronoPL and chronos.</td>
</tr>
<tr>
<td>Ultrametric tree obtention method after ML</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haplotype vs. sequence trees</td>
<td>No substantial differences.</td>
<td>It is possible to use haplotype trees to accelerate computing.</td>
</tr>
<tr>
<td>Taxon level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superfamily to family reduction</td>
<td>No substantial differences. A substantial Yule proportion is present in both cases.</td>
<td>Higher taxonomic levels can be equally operative, at least for taxa with a good number of species. Avoid studies with very few species. If only interested in one or few species, add outgroup taxa and always test for GMYC significance.</td>
</tr>
<tr>
<td>Reduction in clades to the minimal GMYC significance</td>
<td>Low Yule portion levels destabilize GMYC when less than 5 species are included. Within GMYC significance, variation of performance is small.</td>
<td>Sampling should be designed not to miss extreme haplotypes: try to cover isolated or phenotypically differentiated populations and the widest variety of habitats. Plotting intraspecific divergence accumulation curves may help monitoring the sampling strategy.</td>
</tr>
<tr>
<td>Reduction in intermediate haplotypes retaining maximum divergence within a species</td>
<td>No induction of splits at our geographical range. The most critical for a given species is the maximum intraspecific divergence, not so much the presence of intermediate haplotypes.</td>
<td>The extension of sampling area needs to be uniform over all species for comparable results.</td>
</tr>
<tr>
<td>Unbalanced geographical range</td>
<td>Species may suffer from oversplitting when new geographically distant haplotypes increase intraspecific divergence.</td>
<td>A high percentage of singletons may not affect the percentage of correctly identified species, but may considerably reduce the meaningfulness of the analysis.</td>
</tr>
<tr>
<td>Percentage of singletons</td>
<td>A high singleton percentage can be accommodated by GMYC (up to 95%). Threshold and percentages are not widely affected within GMYC significance.</td>
<td></td>
</tr>
</tbody>
</table>

---

Fig. 6. Summarized scheme and practical recommendations for the different steps used regarding the data set, the phylogenetic approach and the GMYC species delimitation. Solid black arrows indicate recommended steps.
the characteristics of our taxa (Zhang et al. 2010). For our data set, it can be seen that above c. 10 specimens or eight populations per species, the average number of haplotypes does not increase and is very variable depending on the species, which suggests that saturation is being reached at least for some species. Given that maximum intraspecific divergence is apparently more important than sheer haplotype numbers, accumulation curves showing how maximum intraspecific divergence grows with sampling effort for each species might be an even better approach to planning and monitoring the sampling strategy.

Geographic range balance among organisms surveyed

Because the GMYC method is based on an internal comparison of tree structure, the fact that all the specimens included in a tree originated from a geographic area of comparable size could be of importance for a reliable result. By enlarging the study area of *Polyommatus icarus* and *Papilio machaon*, we created a highly unbalanced sampling for these two species with respect to the rest of their families data set. This resulted in multiple splitting for *P. icarus* by contrast to the lack of change in *P. machaon*. Indeed, sampling expansion in the case of *P. icarus* was much more pronounced than for *P. machaon* in terms of area and specimens, as well as maximum intraspecific divergence. In conclusion, area sampling imbalances across taxa might indeed result in splitting when a sufficient increase in intraspecific divergence occurs with respect to the general tree shape. Many data sets combine available sequences mixing sampling at different scales for different species: we argue that in this case the results may not be directly comparable.

Conclusions

Our results demonstrate that GMYC performance is little affected in most of the conditions tested, including the phylogenetic methods used to obtain the ultrametric tree, the taxon levels explored and the quality of coverage for each taxon. However, a minimal set of conditions needs to be fulfilled and, as a rule, this space of relative stability is defined by GMYC significance values. We report a species identification performance rate close to 80% and show that failures are mostly caused by data limitations such as non-monophyletic lineages (c. 6%) or the impossibility to avoid both species splitting and lumping with a single threshold. The number of entities obtained invariably overestimates the total number of morphospecies. While our results are obtained with a specific data set and could not be directly applicable to other cases, they suggest that the GMYC may be a valuable tool to objectively establish a divergence threshold, which may be useful, for example, to highlight potential cryptic taxa. A note of caution is nevertheless necessary: a 20% of identification failure implies that the method is not reliable to determine species level without additional supporting data, and thus, GMYC entities cannot be directly considered equivalent to ‘species’, but to ‘potential species’. A synopsis of the factors tested results obtained, and practical recommendations are summarized in Table 1 and Fig. 6.

Acknowledgements

We wish to thank Anna Papadopoulou, Leonardo Dapporto and three anonymous reviewers for fruitful discussions and suggestions. Support for this research was provided by the Spanish MICINN (project CGL2010-21226/BS and predoctoral fellowship BES-2008-002054 to G.T.) and by a postdoctoral scholarship from the Wenner Gren Foundation Sweden to V.D.

References


Reid, N.M. & Carstens, B.C. (2012) Phylogenetic estimation error can decrease the accuracy of species delimitation: a Basque implementation of the general mixed Yule-coalescent model. BMC Evolutionary Biology, 12, 190.


Received 31 January 2013; accepted 25 July 2013
Handling Editor: Emmanuel Paradis

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Table S1. List of Polyommatus icarus and Papilio machaon sequences used in this study apart from the Romanian dataset in Dincă et al. (2011a).

Table S2. GMYC results for the entire dataset and using different phylogenetic approaches.

Table S3. GMYC results for the five family trees using different phylogenetic approaches.

Table S4. Cases of GMYC failure (species that are either split or lumped) when analysing the entire dataset.

Table S5. GMYC results for the five family genealogies including all surveyed specimens and using BEAST with a coalescent prior and a strict molecular clock.

Table S6. Statistics for the GMYC best-fit and other models (within dAICc = 3) for a BEAST tree with a coalescent prior and a strict molecular clock.

Table S7. Comparison between GMYC number of entities for the full family tree and each significant subtree.

Table S8. Among the species that were recovered as an entity when using family-level trees, those displaying the highest intra-specific divergence (p-distance > 0.5%) were used to test for the importance of intermediate haplotypes.

Figure S1. Probabilities (at nodes) that all tips nested within a node exist in a GMYC cluster according to model-averaged GMYC predictions.

Figure S2. Family trees with numbers assigned to nodes (corresponding to those of Table S6) and arrows indicating nodes that are significant for the GMYC model.

Figure S3. Plots showing the number of haplotypes vs. the (a) number of populations and (b) number of specimens for each of the species in the dataset.