

Running head: SAMPLING AND DNA-BASED SPECIES DELIMITATION

Rarity and Incomplete Sampling in DNA-based Species Delimitation

DIRK AHRENS^{1,2}, TOMOCHIKA FUJISAWA^{3,4}, HANS-JOACHIM KRAMMER¹, [JONAS EBERLE](#)¹, [SILVIA FABRIZI](#)^{1,2}, [ALFRIED P. VOGLER](#)^{2,4}

¹ *Centre of Taxonomy and Evolutionary Research, Zoologisches Forschungsmuseum Alexander Koenig Bonn, Adenauerallee 160, 53113 Bonn, Germany*

² *Department of Life Sciences, Natural History Museum, London SW7 5BD, United Kingdom*

³ *Department of Zoology, Graduate School of Science, Kyoto University, Sakyo, Kyoto 606 8502, Japan*

⁴ *Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot SL7 5PY, United Kingdom*

Corresponding author: [Dirk Ahrens](#); Centre of Taxonomy and Evolutionary Research, Zoologisches Forschungsmuseum Alexander Koenig Bonn, Adenauerallee 160, 53113 Bonn, Germany; E-mail: ahrens.dirk_col@gmx.de

Abstract

DNA-based species delimitation may be compromised by limited sampling effort and species rarity, including ‘singleton’ representatives of species, which hampers estimates of intra- vs. interspecies evolutionary processes. In a case study of southern African chafers (beetles in the family Scarabaeidae) many species and subclades were poorly represented and 48.5% of species were singletons. Using *cox1* sequences from >500 specimens and ~100 species, the Generalized Mixed Yule Coalescent (GMYC) analysis as well as various other approaches for DNA-based species delimitation (AGBD, PTP, Species Identifier, Statistical Parsimony), frequently produced poor results if analyzing a narrow target group only, but the performance improved when several subclades were combined. Hence, low sampling may be compensated for by “clade addition” of lineages outside of the focal group. Similar findings were obtained in reanalysis of published datasets of taxonomically poorly known species assemblages of insects from Madagascar. The low performance of undersampled trees is not due to high proportions of singletons *per se*, as shown in simulations (with 13%, 40% and 52% singletons). However, the GMYC method was highly sensitive to variable effective population size (N_e), which was exacerbated by variable species abundances in the simulations. Hence, low sampling success and rarity of species affect the power of the GMYC method only if they reflect great differences in N_e among species. Potential negative effects of skewed species abundances and prevalence of singletons are ultimately an issue about the variation in N_e and the degree to which this is correlated with the census population size and sampling success. Clade addition beyond a limited study group can overcome poor sampling for the GMYC method in particular under variable N_e . This effect was less pronounced for methods of species delimitation not based on coalescent models.

[Singletons; effective population size; Sericini; Coleoptera; southern Africa]

INTRODUCTION

DNA-based species delimitation and identification are now commonplace in studies of biodiversity ([Monaghan et al. 2009](#)). Single-locus data continue to dominate the DNA taxonomy literature and provide currently the only practical approach for studies of hundreds or thousands of species ([Fujisawa and Barraclough 2013](#); [Ratnasingham and Hebert 2013](#)), despite the undisputed theoretical superiority of multi-locus studies for inferring species boundaries (Yang and Rannala 2010, 2014). Methods for *de novo* species delimitation from single-locus DNA sequence data can broadly be classified as tree-based and distance-based ([Templeton 2001](#); [Meier et al. 2006](#); [Puillandre et al. 2012](#); [Zhang et al. 2013](#); [Ratnasingham and Hebert 2013](#)). Tree-based methods, including the Generalized Mixed Yule Coalescence (GMYC) model ([Pons et al., 2006](#); [Fontaneto et al. 2007](#)), estimate the species boundaries using differences in branching or mutation rate in a phylogenetic tree to discern different evolutionary processes of inter- and intraspecific diversification. The GMYC method calculates rates of branching for speciation (λ_s) and coalescence (λ_c) processes, and the shift in these rates on the root-to-tip axis defines the species boundary. For each node a threshold is estimated for the fit of the mixed Yule – coalescence model of two branching rates and its significance is tested against a uniform coalescence process ([Powell 2012](#); [Reid and Carstens 2012](#); [Fujisawa and Barraclough 2013](#)). The same principle is also implemented in the related Poisson Tree Processes method (PTP; [Zhang et al. 2013](#)), which relies on a measure of mutational changes for the internodes, i.e., without the ultrametricization used in GMYC for measuring the branching rates. In contrast, distance-based methods do not utilize the tree topology, but instead rely on the close clustering of haplotypes that is taken as indication of species limits, and hence the species boundaries also depend on the difference in variation of intra- and interspecific rates of diversification ([Puillandre et al. 2012](#)). Other distance methods are not

themselves partition-generating, but fit DNA data to existing hypotheses of species, e.g. from Linnaean species names or morphological diagnostic characters ([Meier et al. 2006](#)).

Robust estimation of species limits defined by intra- vs. interspecific variation requires adequate sampling of both levels to recognize these boundaries. However, in biodiversity studies of taxonomically poorly known groups and ecosystems ([Pons et al. 2006](#); [Monaghan et al. 2009](#); [Fujisawa and Barraclough 2013](#); [Tang et al. 2014](#)) species and genotype sampling may vary due to actual differences in species abundances or due to incomplete sampling. Many species are rare and may be represented by only one “singleton” exemplar ([McGill et al. 2007](#)). [Lim et al. \(2012\)](#) estimated that as many as 30% of formally described invertebrate species have been found only once or in a single location and this proportion may be even higher under the limited geographic sampling in DNA barcoding studies. Conversely, the taxonomist’s choice of exemplars may introduce a different sampling bias that eliminates these natural differences in abundance, e.g. when studying a set of closely related species to evaluate species boundaries.

As the power of species delimitation for most approaches is affected by the ability to measure the diversity within and between species, e.g. using the estimate of coalescence rates (λ_c) and speciation rate (λ_s) in the GMYC model, a high proportion of singletons may obscure the transition point if the rates are estimated from too few groups with available population-level sampling ([Lim et al. 2012](#)). Geographic structure may aggravate this problem: coalescence within local demes is rapid relative to coalescence among demes, and hence strong geographic variation may bias the Yule-coalescence transition threshold if sampling of demes is incomplete ([Lohse 2009](#)). Given the problem of undersampling, [Lim et al. \(2012\)](#) questioned the applicability of the method to samples from local biota.

Several studies have now established that the mean effective population size (N_e) of species relative to their divergence times is the main parameter affecting the accuracy of DNA-based species limits. Hence, the interplay of the population mutation rate (θ) and the rate of species formation (i.e., the depth of the shift from speciation to coalescent branching) are now well-established determinants of species boundaries ([Esselstyn et al. 2012](#); Reid and Carstens 2012; Fujisawa and Barraclough 2013). In addition, variability in N_e among species in a lineage may reduce the performance of the GMYC model, as coalescence times differ and therefore a uniform age of the coalescence model would be inappropriate ([Monaghan et al. 2009](#); [Fujisawa and Barraclough 2013](#)). The coalescent estimate is overlain by the variation in sampling density, in addition to true differences in species abundance. Combined, these factors lead to imbalances in species representation across the tree and the retrieval of singletons ([Reid and Carstens 2012](#); [Talavera et al. 2013](#)). This raises a key question about the correlation of N_e , which has great impact on the accuracy of the species delimitation, and the census population size, which affects the probability of species sampling and gathering singletons.

Given the limitations of sampling, improved estimates for intra- vs. interspecific rates of evolution may be obtained by extending the study of a focal clade to include a broader set of species. Specifically, if the proportion of singletons is extremely high, and the amount of data cannot be increased for the lineage under study, confidence intervals for λ_s and λ_c become large. However, subclade analyses of [Talavera et al. \(2013\)](#) already suggested that the inclusion of data from related lineages could compensate for shortcomings of a poorly represented sample, as the accuracy of parameter values can be improved. This possibility of adding related clades (“clade addition”) could overcome the limitations of low available sample sizes for the focal

group, but a detailed investigation is required under what conditions the addition of more taxa to a GMYC analysis provides the greatest benefit.

We use empirical and simulated data to address how various sampling regimes and the depth of sampling may affect the accurate estimation of the GMYC threshold. Our empirical samples include chafers (a subgroup of the beetle family Scarabaeidae) in the tribe Sericini collected across southern Africa (Fig. 1). They are targets for DNA-based species delineation because of their morphological homogeneity and high species diversity ([Ahrens & Vogler 2008](#)) and the need to associate the phytophagous adults with their soil-dwelling saprophagous or rhizophagous larvae in pest management ([Ahrens et al. 2007](#)). Based on morphological inspection, our sample includes a great proportion of rare species and singletons (49%). These specimens span a broad phylogenetic diversity and therefore we tested if the accuracy of Yule and coalescence rates improves with the inclusion of divergent subclades, compared to analyses of individual clades in isolation. Associated simulations emulating the empirical data (tree depth, variation in N_e and sample size) should help to elucidate how sampling density and species abundances as well as the mode of adding taxa affect species delimitation. In addition, several existing data sets from similar studies of DNA-based species delimitation of tropical insects were used to establish the wider significance of these observations.

MATERIALS AND METHODS

Taxon sampling, DNA extraction, and DNA sequencing

Specimens of Sericini were collected in 53 localities (Fig. 1; Supplementary Fig. 1; Supplementary Table 1) across southern Africa. DNA was extracted from thoracic muscle tissue using Qiagen DNeasy columns or Promega Wizard SV96 Plate extraction kits. Vouchers were deposited in the collection of the Zoologisches

Forschungsmuseum Alexander Koenig (ZFMK). Morphospecies were assigned based on the complex shapes of the male copulation organ that provide reliable diagnostic species traits (Ahrens 2004). In insects, species-specific divergence in male genitalia among close relatives is often dramatic ([Eberhard 1985](#)), and thus they are useful species-diagnostic characters in insect systematics ([Tuxen 1970](#)). Morphometric variation of male genital characters in the study group defines entities that are difficult to distinguish by external morphology (Warnock 2009; Warnock et al., in preparation). For the set of specimens under study here this analysis revealed the presence of 101 morphospecies (including 10 putative species represented only by females whose affinities could not be established). Eight species and four genera of the sister tribe Ablaberini were used as outgroups ([Ahrens and Vogler 2008](#)). For all individuals the 3' half of the cytochrome oxidase subunit 1 (*cox1*) was amplified using primers Pat and Jerry (Simon et al. 1994). Sequencing was performed on both strands using BigDye v. 2.1 and an ABI3730 automated sequencer. Sequences were edited using Sequencher v. 4.8 (Genecodes Corp., Ann Arbor, MI, USA).

In addition, we obtained from Genbank four previously published data sets of insect assemblages consisting of mitochondrial single gene data (*cox1*, *cob*) from 40-130 morphospecies and several hundred specimens each including mayflies (Ephemeroptera: Baetidae and Oligoneuridae), butterflies (Lepidoptera: Hesperiiidae, Papilionoidea, Pieridae, Nymphalidae, Lycaenidae, and Riodinidae), dung beetles (Coleoptera: Scarabaeidae: Scarabaeinae), and water beetles (Coleoptera: Dytiscidae and Hydrophilidae) that had been obtained from six sites across Madagascar ([Monaghan et al. 2009](#)).

Phylogenetic and population genetic analysis and GMYC modeling

Phylogenetic analysis was performed on the full data and on unique haplotypes using Maximum Likelihood (ML) implemented in PhyML v.3.0 ([Guindon and Gascuel 2003](#)) with a GTR+gamma model, as selected by Modeltest using AIC ([Posada and Crandall 1998](#)) with all parameters estimated from the data and four substitution rate categories. In chafers, the resulting trees were rooted with *Cyrtocamenta* sp. (Melolonthinae: Ablaberini), while in the analyses of dung beetles, water beetles, mayflies and butterflies the trees were rooted at the most basal ingroup node established in the analysis of [Monaghan et al. \(2009\)](#) whose outgroups were not included here. We measured branch support using the approximate likelihood ratio test (aLRT) as implemented in PhyML v.3.0 ([Anisimova and Gascuel 2006](#)). The tree of unique haplotypes was made ultrametric using Pathd8 ([Britton et al. 2007](#)) assigning the root an arbitrary age of one. This non-parametric method does not assume a molecular clock, but provides rapid rate smoothing suitable for use with the GMYC method ([Astrin et al. 2012](#); [Tang et al. 2014](#)) and performs similarly well as model-based approaches implemented in BEAST ([Monaghan et al. 2009](#); [Papadopoulou et al. 2013](#); [Talavera et al. 2013](#)). GMYC modeling was performed using the *splits* package ([Ezard et al. 2009](#)) for R (available from <http://r-forge.r-project.org/projects/splits/>). Only the single-threshold model was used, given limited improvements achievable with the multiple threshold model ([Fujisawa and Barraclough 2013](#)).

Subclade analysis was carried out by subdividing arbitrarily each original data set into 5-7 subsets according to the evident major clades in the ML trees. GMYC modeling was conducted on each of the resultant datasets, using the tree from a newly performed ML analysis under inclusion of an additional outgroup specimen selected from each subclade's sister clade. Lineage-through-time plots and the likelihood

surface of the GMYC model were produced for subclades and the full data sets using *splits*.

The same approach was used to analyze combinations of these subclades, to explore whether the number of subclades and order of clade addition affect the GMYC analysis. For the Sericini dataset, we created 14 additional data sets from the seven subclades (labeled A, C, G, E, M, Q, and R in Fig. 1) that were combined in a stepwise manner until all subclades were included, creating two chains that started with subclade A and subclade R, respectively. The order of adding single subclades was either by starting at the basal split and including new branches from the base upwards or by starting with a tip clade and including new branches going towards the base (see Supplement for details).

The accuracy of the GMYC model was assessed by the match ratio = $2 \cdot N_{\text{match}} / (N_{\text{GMYC}} + N_{\text{morph}})$, where N_{match} is the number of species with exact matches, i.e., all specimens of one morphospecies (and only these) belong to one GMYC entity, and N_{GMYC} and N_{morph} are the number of GMYC groups and morphospecies. The match ratio was also applied to test other methods of species delimitation. We also assessed the accuracy of GMYC by simple counts of GMYC entities vs. morphospecies, following other studies (e.g., [Astrin et al. 2012](#); [Reid & Carstens 2012](#); [Hamilton et al. 2014](#)), but this approach overestimates the accuracy as cases of splitting and lumping cancel out each other. The congruence with morphospecies provides a heuristic measure of accuracy against the best estimate of species limits in the empirical data. Finally, we assessed the precision of the GMYC estimate through the width of the AIC-based confidence set ([Fujisawa and Barraclough 2013](#)). Testing the model performance based on the GMYC significance alone (p_{LRT}) ([Talavera et al. 2013](#)) is problematic, as this parameter was not correlated with the match ratio (see Supplement for details).

Watterson's θ (Watterson 1975) was inferred for each GMYC group with multiple haplotypes available using DnaSP v. 5.0 (Librado and Rozas 2009). Values for θ were then tested for the degree of phylogenetic conservation on the *coxI* ML tree. We obtained Blomberg et al.'s K (Blomberg et al. 2003) and Pagel's λ (Pagel 1999) with the *phylosig* function of the R package *phytools* (Revell 2010, 2012). λ and K are quantitative measures of the phylogenetic signal of a trait. Both assess the trait variation relative to values expected under the Brownian motion model of trait evolution. While K is a scaled ratio of the variance among species over the contrast variance, λ is a scaling parameter for the correlations between species.

Alternative species delimitation approaches

Species delimitation was also conducted with the Poisson Tree Processes model (PTP; Zhang et al. 2013), statistical parsimony analysis (Templeton 2001), Automatic Barcode Gap Discovery (ABGD; Puillandre et al. 2012), and distance-based clustering (Meier et al. 2006). Each of these analyses was conducted on the full Sericini data set, the set of unique haplotypes, and the 7 separate subclade sets. In PTP, the Yule-coalescent transition points are modeled based on the change of substitution rates on the phylogenetic input tree. Statistical parsimony analysis as implemented in TCS v.1.3 (Clement et al. 2000) partitions the data into networks of closely related haplotypes connected by changes that are non-homoplastic with a 95% probability (Templeton et al. 1992); if applied to mtDNA the extent of the networks has been found to be largely congruent with Linnaean species (Hart and Sunday 2007). ABGD detects significant differences in intra- and interspecific pairwise distances (i.e., the barcoding gap) without an *a priori* species hypothesis (Puillandre et al. 2012). The ABGD analysis was performed using the Web interface at <http://www.wabi.snv.jussieu.fr/public/abgd/abgdweb.html> with default parameters and

using Kimura 2-parameter (K2P) distances. The minimum relative gap width was set to different values between 0 and 1. K2P distance is widely used for DNA Barcoding ([Hebert et al. 2003](#)), although it may be poorly justified as the model of choice for the character variation encountered in typical barcode datasets ([Srivathsan and Meier 2012](#)). Differences in distance between best model and K2P model estimates are generally small and identification success rates are largely unaffected by model choice ([Collins et al. 2012](#)). Distance-based clustering was performed on pairwise distances applying a pre-defined distance threshold with the Species Identifier module of the TaxonDNA software v.1.6.2 ([Meier et al. 2006](#)). The procedure clusters all individuals by clique (where all individuals are connected to each other by distance values below the threshold) or quasi-clique (where some individuals are connected to each other indirectly, i.e., some distances in the cluster infringe on the threshold). The threshold is chosen for the best match with the number of morphospecies.

Species sampling simulations

Simulations were conducted to investigate the impact of the sampling regime and the existence of rare species. These simulations consisted of three steps: species tree generation, subsampling of species trees, and coalescence simulations of the gene tree on the species tree. First, species trees were simulated in Phylogen (Rambaut 2002) under the pure birth (Yule) model with a root age of ca. 10^6 and a speciation rate of 2×10^{-6} (speciation events per generation) following Fujisawa and Barraclough (2013). In total, 100 trees with 100 extant species were generated. Terminals were sampled from these trees under three scenarios described below. Next, gene genealogies (i.e., coalescent histories) were simulated on these species tree using SIMCOAL ([Excoffier et al. 2000](#)), under a scenario that represents each species by a single population, and two populations (species) are merged at speciation events specified by the simulated

species tree. Root age = 10^6 generations and $N_e = 2 \times 10^3$ were chosen based on the simulation results of Fujisawa and Barraclough (2013), who showed that a root age = $500 \times N_e$ returns about 80-90% success of species recovery (i.e., correct coalescence of gene trees in accordance with the known, simulated species tree). These simulations may be a reasonable approximation of a regional set of insect lineages. Given a fixed number of species in the tree and a constant speciation rate, the only parameter determining the coalescent is the ratio of N_e and root age. Assuming that the root age is 100 Ma, which is the estimated age of Sericini ([Ahrens et al. 2014](#)), $500 \times N_e$ corresponds to $N_e = 200,000$, and N_e is less if the local radiations are younger. These numbers are in a realistic range given existing estimates for insect populations (e.g. [Keller et al. 2005](#), [Charlesworth 2009](#), [Phillipsen and Lytle 2012](#), [Keightley et al. 2014](#)). The youngest age of species in the simulation was ~2500 generations. This means the youngest species are quite likely to be non-monophyletic, as the expected time required for two species to be reciprocally monophyletic with 95% chance is $\sim 2.2 N_e$ ([Hudson and Coyne 2002](#)). An R custom script was used to generate a SIMCOAL input file from the species tree.

Species trees were subsampled to simulate the effect of completeness of sampling under three scenarios, including (a) random lineage sampling, choosing between 3 to 98 species at random from the 100-species trees and removing unsampled species; (b) clustered lineage sampling, whereby in addition to one randomly chosen species, 2 to 97 species are sampled with a probability inversely proportional to the distance from the first species, to simulate a scheme that samples preferentially the closely related species; (c) clade-wise sampling, in which the species tree is split into five sub-clades and, starting from a single clade, the closest sister clades are successively added to the tree. This simulation approximates the clade addition analysis conducted with the empirical data. Five replicated sampling steps per

tree were taken for simulation (a) and (b). In total, 500 trees per subsampling type were generated. Gene trees were simulated within each species tree of 100 species with SIMCOAL after subsampling in the three subsampling schemes, with the number of individual samples per species set to 5.

In these simulations N_e initially was set to 2×10^3 (see above), but allowing for variation in this parameter we also conducted simulations with variable N_e and variable sample sizes of individuals, whereby N_e of each species was drawn from a log-normal distribution with mean= 2×10^3 and standard deviation (sd)=1, 1.5, or 2. In a first round of simulations (“variable N_e ”), the number of individuals drawn per species was kept constant ($n=5$) to assess the effect of variable N_e alone.

Second, in order to assess the effects of sampling probability and/or species abundance, we varied the number of individuals sampled per species (“variable sampling”), while keeping constant $N_e=2 \times 10^3$. Sample sizes for each species were set to be proportional to random numbers drawn from a log-normal distribution with sd=1, 1.5 or 2 and mean=5, which led to proportions of singletons of 13%, 40% and 52%, respectively. In a final round of simulations (“variable N_e plus variable sampling”), both the N_e and the sample size were varied together, whereby both parameters were drawn from the same distribution as before. For each of these analyses we also tested the sensitivity to the sampling regime under the random, clustered and clade-wise scenarios.

The GMYC delimitation was performed on each simulated tree using the *splits* package and the results were compared to the known number of species and extent of species measured by the match ratio (see above; but using instead of N_{morph} the number of true simulated species N_{true}). The AIC-based confidence set was used to measure the confidence of delimitation ([Fujisawa and Barraclough 2013](#)). Statistical

analyses on the simulations were conducted using the *ape* package of R ([Paradis et al. 2004](#)). Code used for simulations is provided in the Supplementary information.

RESULTS

Empirical sequence data and phylogenetic analysis

A total of 584 individuals assigned to 101 morphospecies including 49 singletons (plus eight singletons of Ablaberini used as outgroup) were sequenced for *cox1* (Supplementary Table 2). The resulting matrix of 826 bp containing 410 constant and 383 parsimony informative characters produced 394 unique haplotypes (excluding outgroups: 386). Aligned data matrices, tree files and examples of simulated trees have been submitted to Dryad ([doi:10.5061/dryad.h6q11](https://doi.org/10.5061/dryad.h6q11)). In the ML tree monophyly of most genera was highly supported (aLRT >0.9), except for *Lepiserica*, *Neoserica* and *Maladera*, *Nedymoserica*, *Trochaloserica*, and *Ablaberoides*, which were not monophyletic (Fig. 1; see Supplementary Fig. 2 for a tree with terminals fully labeled). A separate run including all 584 sequences produced a highly similar topology (see Dryad Supplementary Files). Identical haplotypes corresponded to the same morphospecies, except in one case of *Pleophylla* that was also notable for the non-monophyly of five of the eight morphologically well-recognizable morphospecies.

The GMYC procedure established 103 putative species (Table 1; Supplementary Table 3), of which 85 precisely matched a single morphospecies, while 16 other morphospecies were incongruent and were represented by 18 putative GMYC species, for a match ratio of 0.83 (Table 2). These discrepancies affected in particular the genera *Euronycha*, *Pleophylla* and *Neoserica* (Fig. 1), which had low match ratios (Clade A, C, G in Table 2). The population mutation rate θ estimated for each GMYC cluster was found to be highly variable ranging from 0.001 to 0.04 (mean= 0.013) (Fig. 1) and θ was not correlated with the number of sampled haplotypes ($r^2=0.08$)

(Fig. 2). We found that θ was highly variable among close relatives (Fig. 1) and it was not conserved phylogenetically when assessed with Pagel's lambda ($\lambda=7.042022e-05$; $p=1$) nor Blomberg's K ($K=0.40$; $p=0.388$), neither of which were significant.

GMYC accuracy with subclades and cumulative subclade analyses

Compared to the analyses of the full dataset, separate GMYC analysis on individual subclades showed many more splits (Table 1), and the match ratio was lower in three of the seven subclades, while it improved in two (Table 2). The sum of entities within the AIC confidence intervals for the subclade analyses ranged from 82 to 159, compared to 97 to 114 species for the analysis of the full data set (Table 1). In the case of clade C (*Pleophylla*), the confidence interval of the separate GMYC was extremely large, ranging from 1 to 41. The generally reduced accuracy and precision of the subclade analysis was also evident from abrupt speciation-to-coalescence transitions in the lineage-through-time plots and the general lack of clear peaks in the likelihood plots along the time axis (Fig. 3, Supplementary Figure 3).

When subclades were sequentially combined to generate larger clades, the match ratio of morphospecies and GMYC entities did not grow proportionally with the total number of sampled species (Supplementary Fig. 4). The two series with opposite order of subclade addition showed very different correlations of match ratio vs. number of sampled morphospecies, although both were greatly affected by the addition of the poorly performing subclade C (*Pleophylla*). Furthermore, results of the cumulative analyses showed that the match ratio was not correlated with the p_{LRT} value (i.e., the significance of the GMYC model) (Supplementary Fig. 4).

Four data sets of [Monaghan et al. \(2009\)](#) were used to assess the generality of the findings from subclade analysis. These sets were similar to the Sericini data set, each harboring between 42 and 117 species and having similar numbers of singletons

(Table 3). As with the Sericini, these data sets were split into 5-7 well-supported subclades and subjected to GMYC analysis. Match ratios with morphologically recognized species were high for each group. In three of four cases the match ratio was greatly reduced when individual subclades were analyzed and the sum of lower and upper confidence intervals for total species numbers greatly increased over that from the full analysis (Table 4).

Subclade analyses with alternative species delimitation methods

Under all other species delimitation approaches, except for statistical parsimony analysis, species counts and match ratios also differed in the analysis of the full dataset compared to subclades. Statistical parsimony analysis produced the highest number of entities (147 vs. 103 with GMYC), and the match ratio was the lowest. In the PTP analysis the number of entities was increased to 127, whereas in the two clustering methods that number was decreased to just below 100 (Table 1). The analysis of subclades both decreased (ABGD) and increased (PTP, Species Identifier) the number of recognized entities over the combined analysis of all clades (Table 1). As with GMYC, the match ratios deteriorated in ABGD subclade analyses but improved in PTP and SpeciesIdentifier. For example, PTP on single subclades showed an improved match ratio for five of the seven subclades, although they generally were worse than the best estimate from GMYC (on the full dataset) (Table 2). ABGD was sensitive to duplicate haplotypes: while the dataset including only unique haplotypes (as used for all other delimitation analyses) resulted in 100 entities, the all-specimen dataset that included duplicate haplotypes resulted in 124 entities (Table 2).

Species Identifier had the highest match ratio overall, whereby the preferred distance threshold for the all-taxa analysis was an unexpectedly high *cox1* divergence of 6%. At the widely applied threshold of 3% ([Hebert et al. 2003](#)) only 71 clusters

matched the morphospecies (compared to 74 at the preferred 6% cut-off). However, for the analysis on individual subclades the method resulted in different optimal threshold settings between 1% and 8% (Table 1, Supplementary Fig. 5), and while separation into subclades had little impact on the number of matched entities (Table 2), the total number of entities differed depending on the sampling design and equally preferred distance cut-offs in some cases produced multiple different outcomes (Table 1).

Again, findings with the 4 sets of [Monaghan et al. \(2009\)](#) were quite similar, with best performance (in terms of match ratio) of GMYC, PTP, or parsimony networks in most cases also better than the distance-based clustering. However, unlike in the Sericini data set, ABGD, PTP, parsimony networks, and distance based clustering generally performed better in the subclade analyses, except for the mayflies data set (Table 4). The entity counts (Table 3) and match ratios (Table 4) between subclade analyses and complete data analysis for these approaches were slightly less different than with the GMYC approach.

Species sampling simulations

Simulations were used to test the effects of sampling density and proportion of singletons under various schemes of data addition (Fig. 4). Given the known species tree, match ratios were established by comparing coalescent genealogies for a subsample of individuals. We first tested the effects of sampling density under constant effective population size N_e and sample size for the random, clustered and clade-wise sampling (Fig. 4, $sd=0$). Under random sampling, the match ratio slowly declined with increasing density of sampling of the tree, as more species became represented by multiple individuals, until the expected failure rate under the given values for N_e and lineage age was reached. Under the clustered and clade-wise

addition schemes, a slight benefit was evident from greater sampling density, in particular under the clade-wise sampling, but only for up to ~50 sampled species. Next, we varied the species abundance distributions that resulted in a high proportion of singletons of 13%, 40% or 52% (Fig. 4, $sd=1, 1.5, 2$), but this hardly affected the GMYC performance.

In contrast, variable N_e among the species in the tree had a greatly negative impact. Already moderate variation in N_e ($sd=1$) resulted in high variation among replicates and a negative correlation between sampled species number and GMYC model accuracy, both under random and clustered sampling (Fig. 5). For the latter, there was sometimes still a benefit in adding more taxa up to 50 species when N_e was more variable ($sd>1$) (Fig. 5). Likewise, the clade-wise sampling showed clear benefits from adding taxa in particular when total numbers of sampled species were low. GMYC success further deteriorated if variable species sampling was introduced together with variable N_e (Fig. 5, “variable N_e plus variable sampling”). The recovery of accurate species decreased with increasing sd of species abundance distribution (Fig. 5, Supplementary Fig. 7), but regressions of the match ratios showed that clade-wise sampling was always positive for up to 30 species.

We also investigated if the deviations from the known species limits are due to splitting or lumping. The plots under uniform N_e and species abundance distributions showed a generally tight 1:1 ratio of the number of GMYC groups with the number of the true species (Supplementary Fig. 8). Under more extreme conditions (variable N_e plus variable sampling), there was a trend of lumping of true species into single GMYC entities (Fig. 6). However, there were frequent exceptions among all sampling approaches, in particular when only a small proportion of species was sampled (<20 species). The variability of lumping/splitting among each simulation was generally greater under variable abundance sampling (Fig. 6, Supplementary Fig. 8). However,

even in the extreme cases, the total number of GMYC species was not far off from the count of true species, and the differences of these counts were much lower than discrepancies of species limits as measured by match ratios.

Finally, the AIC confidence interval for species estimated by GMYC was assessed (Supplementary Fig. 9). The number of sampled species did not affect the confidence set if the sample size was constant. However, when sampling became more unbalanced, the confidence interval widened (the precision of the species number estimate decreased), in particular as the number of species in the tree increased (see the correlation of the size of the confidence set with species number; Supplementary Fig. 9), for any of the three sample addition schemes. Interestingly, the accuracy of the GMYC species estimate, as measured by the match ratio between GMYC entities and true species, seemed not to be uniformly linked to the AIC confidence sets (Supplementary Fig. 9, lower panel). While the confidence sets were nearly constant under a wide range of conditions, the match ratio was already variable, until both parameters became variable under more extreme values of sample size and N_e .

DISCUSSION

Our empirical data are typical examples of contemporary DNA studies affected by limited sampling due to true rarity and small geographic ranges of many species and due to the logistics of fieldwork. The specimens gathered from natural assemblages were greatly variable in number of individuals per species and included many singletons. We found that even with this high proportion of singletons estimates of species limits were still mostly accurate against the ‘known’ entities established based on morphology, but GMYC accuracy was reduced when analyzing narrow lineages (subclades) alone. The improvement of estimates with all subclades included was evident from the more uniform likelihood surface (Fig. 3), the narrower AIC

confidence sets (Table 1, 3), and the increased match of the GMYC species entities with the morphospecies assignments (Table 2, 3). We conclude that the data from various subclades contribute to the overall parameter estimation of the GMYC model and improve the estimate of species limits locally on the tree. Seemingly, shortcomings of low sample size can be overcome by simply increasing the phylogenetic scope of a study, as already suggested by [Talavera et al. \(2013\)](#).

Our simulations broadly backed this hypothesis. The overriding parameter determining the GMYC success was the variability of N_e in the simulations, as the match ratio between GMYC entities and true species decreased from 95% (constant N_e) to a mean of 54% (variable N_e drawn from a distribution with $sd=2$) for a set of 80-100 species in the tree. Under these conditions, clade-wise sampling clearly had a positive effect, showing an increased match ratio as species were added to the tree (Fig. 5). The positive effect of this sampling scheme was much smaller when only the number of individuals per species was varied, if under constant N_e . Hence there is a slight benefit under any kind of conditions, but under constant N_e these sampling effects were seen only when the number of included species was low (<20 of 100 total species in the tree) and the effect was small altogether.

The Sericini dataset showed a high variability in N_e as measured by θ (i.e., the product of N_e and mutation rate) (Fig. 1). Hence, these are the conditions under which the clade-wise sampling is of greatest benefit (Fig. 5), which is confirmed by the improved accuracy of the estimates from the full clade set over the individual subclades. The simulations also show great variation in the match ratio (Fig. 5) and to a smaller extent in the ratio of numbers of GMYC groups to true species (Fig. 6), in particular for small clades when affected by a combination of variable N_e and high variability of sampling. Thus, species limits in the Sericini data can be expected to be inaccurate on many occasions simply due to the stochastic effects of lineage sorting,

under which conditions the clade addition had the greatest benefit. However, in the empirical data the subclades differ beyond the stochastic differences obtained in simulations, and thus their impact on other subclades varies depending on which clade is included in the data (see Supplementary Fig. 4, two chains of clade addition). First, the subclades include distinctly different numbers of species, beyond what is expected based on their different ages, i.e., there is great imbalance in the species tree and hence in the product of clade age and N_e that determines the probability of correct lineage sorting. Second, the amount of intraspecific variation differed among subclades and thus the GMYC threshold values defining the Yule-coalescence transition are shifted (Supplementary Fig. 3). This great variability of parameter values may limit the predictability of clade addition for improved estimates of species limits in empirical data. Yet, the analyses show that in general there is a clear improvement in the accuracy (number of species and extent) and precision (range of threshold values) of species delimitation when analyzing multiple subclades together (Supplementary Fig. 4). Importantly, this finding is confirmed for three out of four similar datasets from the literature ([Monaghan et al. 2009](#)) that largely indicate the same reduction of GMYC accuracy in analyses limited to subclades (Table 3).

The simulations also served to separate the effect of N_e and low abundance, including the proportion of singletons ([Lim et al. 2012](#)). We established that, on its own, a high proportion of singletons has little impact on the accuracy of inferred species limits, and thus rarity (and singletons) should not be conflated with the much more pertinent population genetics parameters. However, the two parameters still may act together if low N_e is also correlated with small census population size. Indeed, census population size and effective population size seem to be closely correlated at low abundances ([Palstra and Fraser 2012](#)) and consequently rare species are expected to exhibit low N_e . However, this correlation is not universal, and in any case small N_e

at a given clade age would reduce the stochastic error from lineage sorting and thus improve species delimitation. Hence, rarity *per se*, even if correlated with low N_e , would have a limited impact on the accuracy of inferred species limits, unless N_e varies greatly among the species in a clade (Fig. 5).

In the simulations under $sd=2$ the variation of N_e includes a range of mean $N_e/50$ to $50 \times \text{mean } N_e$, which is supposedly far beyond the range of realistic data ([Esselstyn et al. 2012](#); [Fujisawa and Barraclough 2013](#)) and therefore the simulations paint an overly pessimistic picture of the effect of variable population mutation rates. Yet, even in the empirical data the θ values varied over more than an order of magnitude and, under the reasonable assumption of largely uniform mitochondrial mutation rates across this group, the variability of N_e in the South African Sericini may be sufficiently high to have an impact on the successful performance of the GMYC model. In addition, there was a low correlation of N_e and species abundance estimated from the sequenced specimens ($r^2=0.08$; Fig. 2), which in part may be because of low accuracy of the N_e estimates due to small sample sizes, and probably in common with similar practice in most DNA barcoding studies, rare taxa were sequenced preferentially. This non-random approach oversamples rare species and undersamples common species, and thus sampling bias becomes uncorrelated with N_e . This type of sampling probably avoids the more extreme values of our simulations that result from the variation in abundance and consequently the chance of encountering a species. However, even this more uniform sampling regime cannot overcome the negative effect of high variability in N_e on the performance of the GMYC, and hence should be in a range where clade addition is useful, which was confirmed by the subclade analysis.

A related concern is the consistent bias from a phylogenetically conserved θ . Adding more clades to a given sample may increase tree depth (in relation to

speciation rate), e.g. in the chain starting with subclade A (Supplementary Fig. 4), and therefore as the number of lineages increases, rate estimates become more reliable and delimitation is improved. However, if these wider samples differ from the focal clade in population genetic parameters, the clade addition may introduce systematic biases into the species delimitation. We found that θ in Sericini is not phylogenetically conserved, and evidence from different taxa, including plants, lizards and aquatic beetles ([Ricklefs 2010, 2011](#); [Losos 2011](#); [Fujisawa et al. 2014](#)), suggests that N_e is not correlated with the species phylogeny more generally. Hence, the variation in N_e is at the species level, rather than clade level, and thus clade-wise addition of species is unlikely to introduce confounding variables. In addition, the fact that θ is not correlated with the number of the sampled haplotypes (Fig. 2) shows that the analysis also avoided potential biases in sampling density among subclades that could lead to consistent differences in the estimated N_e . These findings further validate the clade addition approach.

The effect of N_e and sampling bias was established here primarily for the GMYC approach, but these parameters should affect all methods that estimate the species limits from gene genealogies, as the underlying coalescent processes affect the inter- and intraspecific variation in all cases ([Ross et al. 2008](#); [Zhang et al. 2013](#)). Among the three partition-generating methods applied here, GMYC and PTP methods use the sequence information by modeling speciation and coalescent processes. In the former, tree linearization may force very short terminal branches to be longer, and in combination with branches defined by higher substitution rates may artificially oversplit the GMYC entities. Thus PTP should be more robust by avoiding the need for tree linearization ([Tang et al. 2014](#)), and potentially also be less affected by restricted sampling as encountered in our study with subclade analyses (Table 4). Our results indicate that the latter may apply also to the ABGD method that uses pairwise

distances among all sequences and searches for a break in the variation to establish the barcoding gap, which equally is not affected by artifacts from ultrametricization. Similarly, the parsimony network analysis that builds networks by linking the closest relatives rather than using parameters estimated across the wider data, seems to profit from simpler calculations on a more restricted data set. The improved subclade performance of these methods would argue against the utility of the clade addition approach, unlike in the case of the GMYC method, and may even suggest the opposite, i.e., the analysis of narrowly defined clades. However, the exclusive focus on species counts and the match ratio does not consider other parameters, such as the narrower confidence interval and overall fit of the likelihood surface that provide additional measures of success from clade addition in the GMYC model. This additional information is useful for characterizing a data set when assessing the effects of coalescent depth, geographic structure, variable N_e and others, particularly in situations where molecular species delimitation is not successful with all methods (e.g. [Hamilton et al. 2014](#)).

While the approaches for evaluating the sequence variation differ profoundly, in all cases the assessed pattern is the result of evolutionary processes of coalescence and speciation in the history of these lineages. All methods detect more than 2/3 of entities that are also established based on morphology (Tables 2, 4), showing that a large portion of the data is uncontroversial in identifying the Yule-coalescent boundaries. Besides these uncontroversial entities, the molecular methods tend to increase the number of entities beyond the morphological groups to various degrees (Table 1, 3). However, the tendency for strong oversplitting was evident in the subclade analysis only for the GMYC method. We do not know the true error, as the morphological data used as a reference equally are based on a single character system that may not be perfectly reliable. Yet, the potential shortcomings of the reference system and the

differences among DNA-based methods should not distract from the fact that in all cases their primary aim is to evaluate the coalescent-to-phylogeny transitions. The statistical properties of these methods have been tested in the context of separating coalescence and speciation processes, in detailed studies for GMYC (Fujisawa and Barraclough 2013), PTP ([Zhang et al. 2013](#)) and ABGD (Puillarde et al. 2012) approaches. The basic findings of these studies about the factors determining the success of species delimitation are similar, and thus the parameters being varied in the simulations and the resulting conclusions about rarity and singletons are informative broadly and not unique to the use of the GMYC method. In contrast, distance-based clustering methods such as Species Identifier that do not generate *de novo* partitions of sequence variation are affected by the coalescence processes only insofar as these determine the evolution of species limits. The threshold cut-off is calibrated by optimizing the number of exact matches with the morphospecies ([Meier et al. 2006](#)), and this number apparently is not affected whether all species are analyzed together or in broken up into subclades (Table 2). However, those individuals that are not ‘conspecific’ (do not match with the morphospecies) are clustered according to the threshold producing the best match, and these average cut-off values differ between the all-taxa and subclade analyses, which produces some differences in cut-off values and the resulting number of entities (Table 1). These thresholds have an impact on how many singletons are recognized but singletons *per se* do not interfere with the determination of the threshold values.

CONCLUSIONS

The great proportion of “singletons”, i.e., the only representative of a species ever found, is a hallmark of the state of taxonomy in many small-bodied animal groups

([Lim et al. 2012](#)). Species rarity and great variation in effective population size, combined with the difficulty of comprehensive sampling, complicate the use of sequence-based species delimitation. We show that the overriding issue for DNA-based species delimitation is the variable θ , which may or may not be linked to species abundance (sampling probability) in empirical studies, and that sampling bias may aggravate the effect of variable θ , but not to a great extent and only under rather extreme conditions. Due to the likely correlation of census population size and effective population size at low abundances, a high proportion of rare species would result in many cases of low θ and thus produce low variation in θ , which is favorable to GMYC performance. In the empirical example used here, variability of θ was high but not phylogenetically conserved, and thus presented the conditions where the use of extended sampling beyond a focal clade can improve the estimates of species limits. Confidence of species delimitation requires the exploration of population genetic parameters (root age, θ) in a data set (see also [Carstens et al. 2013](#)). This knowledge underlines the need for an integrative taxonomy that is based on a diverse set of methods (e.g., [Padial et al. 2010](#); [Hendrixson et al. 2013](#); [Miralles and Vences 2013](#); [Carstens et al. \(2013\)](#); [Hamilton et al. 2014](#)), in which the GMYC model, as other automated molecular species delimitation methods, retains a crucial role as a circularity-free tool in an iterative process of hypothesis refinement ([Yeates et al. 2011](#); [Fujisawa and Barraclough 2013](#)). However, all quantitative and iterative species delineation approaches using statistical or model-based algorithms of grouping in one way or another rely on sufficiently comprehensive and unbiased sampling ([Hausdorf and Hennig 2010](#); [Edwards and Knowles 2014](#)). Therefore, both the vagaries of population processes and the incongruous sampling of species will continue to be a challenge to modern taxonomy.

SUPPLEMENTARY MATERIAL

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at <http://datadryad.org>, doi:10.5061/dryad.h6q11.

FUNDING

The study was supported by the German Science Association (DFG; AH175/1-2 and AH/175/3) and by the Leverhulme Trust (F00/696P to APV and DA).

ACKNOWLEDGEMENTS

We are grateful to [Ruth Müller](#), [David Kral](#), Miloslav Snizek, Karl Werner, and [James Harrison](#) for collecting additional specimens and to [M. Monaghan](#) and [T. Barraclough](#) for helpful advice. For providing us with research and collection permits, we thank the various governmental institutions and departments in Eastern Cape (Permit No.: WRO 122/07WR and WRO123/07WR), Gauteng (Permit No.: CPF6 1281), Limpopo (Permit No.: CPM-006-00001), Mpumalanga (Permit No.: MPN-2009-11-20-1232), and Kwazulu-Natal (Permit Nos OP3752/2009, 1272/2007, 3620/2006). We thank Frank Anderson and Mark Fishbein as well as the anonymous referees for their helpful comments on the manuscript.

REFERENCES

- Ahrens D. 2004. Monographie der Sericini des Himalaya (Coleoptera, Scarabaeidae). Dissertation.de - Verlag im Internet GmbH, Berlin, 534 pp.
- [Ahrens D., Monaghan M.T., Vogler A.P. 2007. DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers \(Coleoptera: Scarabaeidae\). Mol. Phylogenet. Evol. 44: 436-449.](#)

- [Ahrens D., Vogler A.P. 2008. Towards the phylogeny of chafers \(Sericini\): Analysis of alignment-variable sequences and the evolution of segment numbers in the antennal club. *Mol. Phylogenet. Evol.* 44: 436–449.](#)
- [Ahrens D., Schwarzer J., Vogler A.P. 2014. The evolution of scarab beetles tracks the sequential rise of angiosperms and mammals. *Proc. R. Soc. B* 281: DOI: 10.1098/rspb.2014.1470](#)
- [Anisimova M., Gascuel O. 2006. Approximate likelihood ratio test for branches: A fast, accurate and powerful alternative. *Syst. Biol.* 55: 539-552.](#)
- [Astrin J.J., Stüben P.E., Misof B., Wägele J.W., Gimmich F., Raupach M.J., Ahrens D. 2012. Exploring diversity in cryptorhynchine weevils \(Coleoptera\) using distance-, character- and tree-based species delineation. *Mol. Phylogenet. Evol.* 63: 1-14.](#)
- [Blomberg S.P., Garland T., Ives A.R. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57: 717–745.](#)
- [Britton T., Anderson C.L., Jacquet D., Lundqvist S., Bremer K. 2007. Estimating divergence times in large phylogenetic trees. *Syst. Biol.* 56: 741-752.](#)
- [Carstens B.C., Pelletier T.A., Reid N.M., Satler J.D. 2013. How to fail at species delimitation. *Mol. Ecol.* 22: 4369-4383.](#)
- [Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews* 10: 195-205.](#)
- [Clement M., Posada D., Crandall K.A. 2000. TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* 9: 1657-1659.](#)
- [Collins R.A., Boykin L.M., Cruickshank R.H., Armstrong K.F. 2012. Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods Ecol. Evol.* 3: 457–46.](#)
- [Eberhard W.G. 1985. Sexual Selection and Animal Genitalia. Harvard University Press, Boston, MA, USA.](#)

- Edwards D.L., Knowles L.L. 2014. Species detection and individual assignment in species delimitation: can integrative data increase efficacy? *Proc. R. Soc. B* 281: 20132765. <http://dx.doi.org/10.1098/rspb.2013.2765>
- Esselstyn J.A., Evans B.J., Sedlock J.L., Khan F.A.A., Heaney L.R. 2012. Single-locus species delimitation: a test of the mixed Yule–coalescent model, with an empirical application to Philippine round-leaf bats. *Proc. R. Soc. B* 279: 3678–3686.
- Excoffier L., Novembre J., Schneider S. 2000. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J. Hered.* 91: 506-509.
- Ezard T., Fujisawa T., Barraclough T. 2009. splits: SPecies Limits by Threshold Statistics. R package version 1.0-11/r29. Available from <http://R-Forge.R-project.org/projects/splits/>.
- Fontaneto D., Herniou E.A., Boschetti C., Caprioli M., Melone G., Ricci C., Barraclough T.G. 2007. Independently evolving species in asexual bdelloid rotifers. *PLoS Biol.* 5: 914–921.
- Fujisawa T., Barraclough T. (2013) Delimiting species using the Generalized Mixed Yule Coalescent (GMYC) approach: A revised method and evaluation on simulated datasets. *Syst. Biol.* 62: 707-724.
- Fujisawa T., Vogler AP, Barraclough TG. 2014. Ecology has contrasting effects on genetic variation within species versus rates of molecular evolution across species in water beetles. *Proc. R. Soc. B* 282: 20142476. <http://dx.doi.org/10.1098/rspb.2014.2476>
- Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52: 696-704.
- Hamilton C.A., Hendrixson B.E., Brewer M.S., Bond J.E. 2014. An evaluation of sampling effects on multiple DNA barcoding methods leads to an integrative

approach for delimiting species: a case study of the North American tarantula genus *Aphonopelma* (Araneae, Mygalomorphae, Theraphosidae). *Mol. Phylogen. Evol.* 71: 79-93.

Hart M.W., Sunday J. 2007. [Things fall apart: biological species form unconnected parsimony networks. *Biol. Lett.* 3: 509-512.](#)

Hausdorf B., Hennig C. 2010. Species delimitation using dominant and codominant multilocus markers. *Syst. Biol.* 59: 491–503. (doi:10.1093/sysbio/syq039)

Hebert P.D.N., Cywinska A., Ball S.L., DeWaard J.R. 2003. [Biological identifications through DNA barcodes. *Proc. R. Soc. B* 270: 313–321.](#)

Hendrixson B.E., DeRussy B.M., Hamilton C.A., Bond J.E. 2013. [An exploration of species boundaries in turret-building tarantulas of the Mojave Desert \(Araneae, Mygalomorphae, Theraphosidae, *Aphonopelma*\). *Mol. Phylogen. Evol.* 66: 327-340.](#)

Hudson R.R., Coyne J.A. 2002. [Mathematical consequences of the genealogical species concept. *Evolution* 56: 1557-1565.](#)

Keightley P.D., Pinharanda A., Ness R.W., Simpson F., Dasmahapatra K.K., Mallet J., Davey J.W., Jiggins C.D. 2014. [Estimation of the Spontaneous Mutation Rate in *Heliconius melpomene*. *Mol. Biol. Evol.* 32: 239–243.](#)

Librado P., Rozas J. 2009. [DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451-1452.](#)

Lim G., Balke M., Meier R. 2012. [Determining species boundaries in a world full of rarity: Singletons, species delimitation methods. *Syst. Biol.* 61: 165–169.](#)

Lohse K. 2009. [Can mtDNA barcodes be used to delimit species? A response to Pons et al. *Syst. Biol.* 58: 439-442.](#)

Losos J.B. 2011. [Seeing the Forest for the Trees: The Limitations of Phylogenies in Comparative Biology. *Am. Nat.* 177: 709–727.](#)

- Maddison W.P., Maddison D.R. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75 <http://mesquiteproject.org>
- McGill B.J., Etienne R.S., Gray J.S., Alonso D., Anderson M.J., Benecha H.K., Dornelas M., Enquist B.J., Green J.L., He F., Hurlbert A.H., Magurran A.E., Marquet P.A., Maurer B.A., Ostling A., Soykan C.U., Ugland K.I., White E.P. 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.* 10: 995-1015.
- Meier R., Shiyang K., Vaidya G., Peter K. L. NG. 2006. DNA Barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Syst. Biol.* 55: 715-728.
- Miralles A., Vences M. 2013. New metrics for comparison of taxonomies reveal striking discrepancies among species delimitation methods in *Madascincus* lizards. *PLoS ONE* 8(7): e68242.
- Monaghan M.T., Wild R., Elliot M., Fujisawa T., Balke M., Inward D.J., Lees D.C., Ranaivosolo R., Eggleton P., Barraclough T.G., Vogler A.P. 2009. Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Syst. Biol.* 58: 298–311.
- Padial J.M., Miralles A., De la Riva I., Vences M. 2010. The integrative future of taxonomy. *Front. Zool.* 7: 16.
- Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401: 877–884.
- Palstra F.P. & Fraser D.J. 2012. Effective/census population size ratio estimation: a compendium and appraisal. *Ecol. Evol.* 2: 2357–2365.
- Papadopoulou A., Cardoso A., Gomez-Zurita J. 2013. Diversity and diversification of Eumolpinae (Coleoptera: Chrysomelidae) in New Caledonia. *Zool. J. Linn. Soc.* 168: 473–495.

- [Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.](#)
- [Phillipsen I.C., Lytle D.A. 2012. Aquatic insects in a sea of desert: population genetic structure is shaped by limited dispersal in a naturally fragmented landscape. *Ecography* 35: 1–13.](#)
- [Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol.* 55: 595-609.](#)
- [Posada D., Crandall K.A. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14: 817-818.](#)
- [Powell JR. 2012. Accounting for uncertainty in species delineation during the analysis of environmental DNA sequence data. *Meth. Ecol. Evol.* 3: 1-11.](#)
- [Puillandre N., Lambert A., Brouillet S., Achaz G. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Mol. Ecol.* 21: 1864–1877.](#)
- [Rambaut A. 2002. PhyloGen: phylogenetic tree simulator package, v. 1.1. See <http://evolve.zoo.ox.ac.uk/software/PhyloGen/>.](#)
- [Ratnasingham S., Hebert P.D.N. 2013. A DNA-based registry for all animal species: The Barcode Index Number \(BIN\) system. *PLoS One* 8\(8\): e66213.](#)
- [Reid N.M., Carstens B.C. 2012. Phylogenetic estimation error can decrease the accuracy of species delimitation: a Bayesian implementation of the general mixed Yule-coalescent model. *BMC Evol. Biol.* 12: 196.](#)
- [Revell L.J. 2010. Phylogenetic signal and linear regression on species data. *Methods Ecol Evol* 1: 319–329.](#)
- [Revell L.J. 2012. phytools: an R package for phylogenetic comparative biology \(and other things\). *Methods Ecol. Evol.* 3: 217–223.](#)

- [Ricklefs R.E. 2010. Evolutionary diversification, coevolution between populations and their antagonists, and the filling of niche space. PNAS 107: 1265–1272.](#)
- [Ricklefs R.E. 2011. Applying a regional community concept to forest birds of eastern North America. PNAS 108: 2300–2305.](#)
- [Ross H.A., Murugan S., Li W.L.S. 2008. Testing the reliability of genetic Methods of species identification via simulation. Syst. Biol. 57: 216–230.](#)
- Simon C., Frati F., Beckenbach A., Crespi B., Liu H., Flook P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequence and compilation of conserved polymerase chain reaction primers. *Ann. Entomol. Soc. Am.* 87: 651-701.
- [Srivathsan A., Meier R. 2012. On the inappropriate use of Kimura-2-parameter \(K2P\) divergences in the DNA-barcoding literature. Cladistics 27: 190–194.](#)
- [Talavera G., Dincă V., Vila R. 2013. Factors affecting species delimitations with the GMYC model: insights from a butterfly survey. Meth. Ecol. Evol. 4: 1101-1110.](#)
- [Tang C.Q., Humphreys A.M., Fontaneto D., Barraclough T.G. 2014. Effects of phylogenetic reconstruction method on the robustness of species delimitation using single-locus data. Meth. Ecol. Evol. 5: 1086–1094.](#)
- [Templeton A.R., Crandall K.A., Sing C.F. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA-sequence data. 3. Cladogram estimation. Genetics 132: 619-633.](#)
- [Templeton A.R. 2001. Using phylogeographic analyses of gene trees to test species status and processes. Mol. Ecol. 10: 779–791.](#)
- Tuxen S.L. (Ed.) 1970. *Taxonomist's Glossary of Genitalia in Insects*. Scandinavian University Press, Copenhagen.

- Warnock R. 2009. A comparison between species delineation based on DNA sequences and genital morphometrics in beetles (Coleoptera). Unpublished master thesis, Imperial College London: 1-34.
- Watterson G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7: 256–276.
- Yang Z., Rannala B. 2010. Bayesian species delimitation using multi-locus sequence data. *Proc. Natl. Acad. Sci. U S A.* 107(20): 9264–9269.
- [Yang Z., Rannala B. 2014. Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.* 31: 3125–3135.](#)
- [Yeates D.K., Seago A. Nelson A., Cameron S.L., Joseph L., Trueman, J.W.H. 2011. Integrative taxonomy, or iterative taxonomy? *Syst. Entomol.* 36: 209–217.](#)
- [Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* 29: 2869-76.](#)

Table 1. Number of DNA-based entities in comparison to the number of morphospecies.

	GMYC		PTP	ABGD	Distance-based clustering		Parsim. networks	Morpho-species
Clade	N _{GMYC}	Cs			Threshold	N _{Clust}	N _{TCS}	N _{morph}
A	5 (7)	5-13	5 (12)	7 (6)	1/ 2.5%	13/ 9 (6)	13 (13)	6
C	4 (5)	1-41	6 (11)	7 (6)	4%	4 (4)	19 (19)	9
E	11 (11)	10-12	10 (12)	10 (11)	4%	11 (11)	15 (15)	8[+1 ♀]
G	27 (26)	25-35	28 (27)	33 (24)	3/ 8%	27/ 20 (24)	32 (32)	17[+8 ♀]
M	6 (6)	5-7	6 (8)	6 (6)	4%	6 (6)	8 (8)	6
Q	9 (5)	9-9	5 (5)	5 (4)	2%	5 (4)	5 (5)	5
R	41 (32)	27-42	42 (39)	37 (31)	6%	29 (29)	43 (43)	28[+1 ♀]
Σ	103 (92)	82-159	102 (114)	105 (88)		95/ 84 (84)	135 (135)	79[+10 ♀]
ALL [§]	103	97-114	127	100	6%	95	147	101

Note: The number of putative GMYC species (N_{GMYC}) is given with their AIC confidence set (Cs).

Values in parentheses represent the number of entities for that subclade from the ‘all data set’ analysis.

The outgroup was not considered in the count for the entities and confidence interval of the subclades ([§] includes Ablaberini and several species not contained in subclades). In case of resulting equally compatible thresholds for distance-based clustering, both values are given.

Table 2. Match ratio of DNA-based species delimitation methods on chafer data with reference to morphospecies.

Clade	GMYC _A	GMYC _S	PTP _A	PTP _S	ABGD _A	ABGD _S	Dist. Clust. _A	Dist. Clust. _S	Networks
A	0.46 (3)	0.73 (4)	0.56 (5)	0.73 (4)	0.5 (3)	0.46 (3)	0.5 (3)	0.53/0.67 (5/5)	0.53 (5)
C	0.46 (3)	0.5 (3)	0.21 (2)	0.43 (3)	0.4 (3)	0.38 (3)	0.46 (3)	0.46 (3)	0.14 (2)
E	0.80 (8)	0.80 (8)	0.67 (7)	0.84 (8)	0.8 (8)	0.74 (7)	0.8 (8)	0.8 (8)	0.5 (6)
G	0.90 (23)	0.88 (23)	0.88 (23)	0.87 (23)	0.73 (18)	0.76 (22)	0.9 (22)	0.88/0.8 (23/18)	0.77 (22)
M	1 (6)	1 (6)	0.57 (4)	1 (6)	1 (6)	1 (6)	1 (6)	1 (6)	0.57 (4)
Q	1 (5)	0.43 (3)	1 (5)	1 (5)	0.67 (3)	1 (5)	0.67 (3)	1 (5)	1 (5)
R	0.85 (26)	0.60 (21)	0.65 (22)	0.59 (21)	0.93 (28)	0.73 (28)	1 (29)	1 (29)	0.53 (19)
Σ	<u>0.82 (74)</u>	<u>0.71 (68)</u>	<u>0.67 (68)</u>	<u>0.74 (70)</u>	<u>0.78 (69)</u>	<u>0.72 (70)</u>	<u>0.86 (74)</u>	<u>0.86 (79/74)</u>	<u>0.56 (63)</u>
ALL[§]	0.83 (85)	-	0.7 (80)	-	0.8 (81)	-	0.87 (85)	-	0.6 (74)

Note: Summed match values are in bold, highest sum underlined; absolute number of exact matches in parenthesis for combined (_A) and subclade (_S) analysis ([§] includes Ablaberini and several species not contained in subclades).

Table 3. Results of species delimitation on the water beetle, dung beetle, mayfly and butterfly data for combined (A) and subclade (s) analyses.

Clade	GMYC _A	GMYC _s	CS	N _{morph}	PTP _A	PTP _s	ABGD _A	ABGD _s	DC _A	DC _s	T%	TCS _A	TCS _s	%
Water beetles														
A	6	6	[6-7]	6	6	6	8	8	6	6	1.5-7.5	6	6	40.0
B	5	10	[10-10]	4	6	6	6	6/8	4	4	2.5-4	5	5	40.0
C	32	33	[32-33]	32	31	31	33	30	30	34	1-1.5	32	32	37.5
D	9	8	[8-8]	7	8	8	13	8	7	8	2-7	8	8	12.3
E	27	27	[26-27]	24	26	27	24	26	24	26	2-4	24	26	45.8
F	29	78	[77-78]	29	28	29	36	26	22	26	2	41	30	41.4
Σ	108	162	[159-163]	103	105	107	114	104/106	71	104	-	116	107	
All [§]	117	-	[110-119]	111	117	-	134	-	104	-	2	116	-	39.6
Dung beetles														
A	na	na	na	3	3	2	1	na	2	3	0.5	2	2	100
B	11	11	[11-11]	11	11	11	11	11	11	11	1-6.5	11	11	41.7
C	9	9	[5-12]	10	9	9	10	8	7	8	2-3.5	10	10	37.5
D	3	3	[2-3]	3	2	3	3	3	3	3	1-7.5	3	3	0
E	19	11	[11-17]	15	15	13	21	12	15	16	2-2.5	14	17	46.2
Σ	42	34	[29-43]	42	41	38	46	34	38	41	-	40	43	
All [§]	47	-	[46-47]	40*	41	-	47	-	40	-	2-2.5	41	-	43.2
Mayflies														
A	7	7	[0-12]	7	8	5	6	6	7	7	2.5-3.5	7	7	42.8
B	20	20	[10-22]	20	18	15	20	20	20	20	2-2.5	20	20	57.9
C	6	19	[19-19]	2	7	6	7	7	7	3	4.5	7	6	0
D	7	4	[4-7]	7	7	7	7	7	7	7	0.5-4	7	7	71.4
E	8	8	[8-8]	9	9	9	9	8	9	9	2-10	9	9	50.0
F	4	4	[2-4]	4	5	4	4	4	4	4	0.5-9.5	4	4	75.0
Σ	52	62	[44-72]	49	54	46	53	52	54	49	-	54	53	
All [§]	65	-	[59-66]	57	63	-	61	-	63	-	2.5	63	-	64.9
Butterflies														
A	29	43	[43-43]	29	30	29	28	29	29	29	1.5-3.5	29	29	44.8
B	18	23	[23-23]	18	17	18	17	18	18	18	1-2.5	18	18	61.1
C	33	34	[33-35]	34	33	33	32	33	33	34	1	33	33	79.4
D	5	11	[11-11]	5	6	5	5	5	5	5	1.5-6.5	5	5	16.7
E	8	8	[8-8]	8	8	8	8	8	8	8	1-10	8	8	87.5
F	7	7	[4-8]	7	7	7	7	7	7	7	2-7	7	7	85.7
G	16	16	[5-17]	18	16	16	15	15	15	18	1	15	15	83.3
Σ	116	142	[127-145]	119	117	116	112	115	115	119	-	115	115	
All [§]	129	-	[126-131]	130	131	-	125	-	127	-	2-2.5	127	-	65.4

Note: Number of entities for GMYC (including confidence sets of GMYC analysis - CS), PTP, ABGD, Distance Clustering (DC, including the threshold (T%) in %), parsimony networks (NW) and morphospecies (N_{morph}), and percent singletons (%) ([§] - includes several species not contained in subclade analyses). Several morphospecies appear in different subclades, therefore the sum of subclade morphospecies may exaggerate total morphospecies number.

Table 4. Match ratio of DNA-based species delimitation methods on the water beetle, dung beetle, mayfly and butterfly data of Monaghan et al. (2009).

Clade	GMYC _A	GMYC _S	PTP _A	PTP _S	ABGD _A	ABGD _S	DC _A	DC _S	TCS _A	TCS _S
Waterbeetles										
A	1 (6)	1 (6)	1 (6)	1 (6)	0.71 (5)	0.71 (5)	1 (6)	1 (6)	1 (6)	1 (6)
B	0.67 (3)	0.14 (1)	0.4 (2)	0.4 (2)	0.4 (2)	0.67 (3)	1 (4)	1 (4)	0.67 (3)	0.67 (3)
C	0.88 (28)	0.92 (30)	0.86 (27)	0.76 (24)	0.8 (26)	0.74 (23)	0.71 (22)	0.82 (27)	0.81 (26)	0.81 (26)
D	0.62 (5)	0.8 (6)	0.8 (6)	0.8 (6)	0.4 (4)	0.8 (6)	0.71 (5)	0.8 (6)	0.8 (6)	0.8 (6)
E	0.90 (23)	0.82 (21)	0.68 (17)	0.78 (20)	0.71 (17)	0.84 (21)	0.79 (19)	0.8 (20)	0.75 (18)	0.84 (21)
F	0.86 (25)	0.22 (12)	0.6 (17)	0.66 (19)	0.4 (13)	0.67 (17)	0.59 (15)	0.62 (17)	0.46 (16)	0.51 (15)
Σ	0.86 (90)	0.58 (76)	0.72 (75)	0.73 (77)	0.62 (67)	0.75 (75)	0.73 (71)	0.78 (80)	0.69 (75)	0.72 (75)
ALL [§]	0.84 (96)	-	0.84 (93)	-	0.6 (72)	-	0.7 (76)	-	0.71 (81)	-
Dung beetles										
A	na	na	1 (3)	0.33 (1)	0 (0)	na	0.4 (1)	1 (3)	0.4 (1)	0.4 (1)
B	1 (11)	1 (11)	1 (11)	1 (11)	1 (11)	1 (11)	1 (11)	0.91 (10)	1 (11)	1 (11)
C	0.53 (5)	0.53 (5)	0.53 (5)	0.74 (7)	0.5 (5)	0.67 (6)	0.59 (5)	0.56 (5)	0.6 (6)	0.6 (6)
D	0 (0)	1 (3)	0 (0)	1 (3)	0 (0)	1 (3)	0.33 (1)	1 (3)	0 (0)	1 (3)
E	0.47 (8)	0.69 (9)	0.6 (9)	0.79 (11)	0.39 (7)	0.67 (9)	0.6 (9)	0.77 (26)	0.55 (8)	0.69 (11)
Σ	0.57 (24)	0.74 (28)	0.67 (28)	0.81 (33)	0.52 (23)	0.76 (29)	0.68 (27)	0.8 (33)	0.63 (26)	0.75 (32)
ALL [§]	0.67 (29)	-	0.67 (27)	-	0.62 (27)	-	0.68 (27)	-	0.63 (25)	-
Mayflies										
A	1 (7)	1 (7)	0.8 (6)	0.67 (4)	1 (7)	0.77 (5)	1 (7)	1 (7)	1 (7)	1 (7)
B	0.9 (18)	0.9 (18)	0.79 (15)	0.63 (11)	0.61 (11)	0.9 (18)	0.9 (18)	0.9 (18)	0.95 (19)	0.9 (18)
C	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
D	1 (7)	0.36 (2)	1 (7)	1 (7)	1 (7)	1 (7)	1 (7)	1 (7)	1 (7)	1 (7)
E	1 (9)	1 (8)	1 (9)	1 (9)	1 (9)	1 (9)	1 (9)	1 (9)	1 (9)	1 (9)
F	1 (4)	1 (4)	0.67 (3)	1 (4)	1 (4)	1 (4)	1 (4)	1 (4)	1 (4)	1 (4)
Σ	0.86 (44)	0.7 (39)	0.78 (40)	0.74 (35)	0.75 (38)	0.84 (43)	0.87 (45)	0.96 (85)	0.89 (46)	0.88 (45)
ALL [§]	0.90 (55)	-	0.83 (50)	-	0.80 (47)	-	0.92 (55)	-	0.92 (55)	-
Butterflies										
A	1 (29)	0.5 (18)	0.95 (28)	1 (29)	0.98 (28)	1 (29)	1 (29)	1 (29)	1 (29)	1 (29)
B	0.94 (17)	0.68 (14)	0.91 (16)	1 (18)	0.98 (14)	1 (18)	1 (18)	1 (18)	0.94 (17)	1 (18)
C	0.90 (30)	0.91 (31)	0.87 (29)	0.96 (32)	0.85 (28)	0.81 (27)	0.87 (29)	0.91 (31)	0.87 (29)	0.87 (29)
D	1 (5)	0.12 (1)	0.73 (4)	1 (5)	1 (5)	1 (5)	1 (5)	1 (5)	1 (5)	1 (5)
E	1 (8)	1 (8)	1 (8)	1 (8)	1 (8)	1 (8)	1 (8)	1 (8)	1 (8)	1 (8)
F	1 (7)	1 (7)	1 (7)	1 (7)	1 (7)	1 (7)	1 (7)	1 (7)	1 (7)	1 (7)
G	0.88 (15)	0.88 (15)	0.71 (12)	0.88 (15)	0.79 (13)	0.61 (10)	0.79 (13)	0.83 (15)	0.79 (13)	0.79 (13)
Σ	0.94 (111)	0.72 (94)	0.88 (104)	0.97 (114)	0.89 (103)	0.89 (104)	0.93 (109)	0.94 (103)	0.92 (108)	0.93 (109)
ALL [§]	0.99 (128)	-	0.9 (117)	-	0.91 (117)	-	0.93 (120)	-	0.94 (122)	-

Note: Results are given for combined (_A) and subclade (_S) analysis ([§] includes several species not contained in subclade analyses; summed values in bold, highest sum underlined). The absolute number of exact matches with morphospecies in parenthesis.

Figure legends

Figure 1. Ultrametric tree of the southern African Sericini species. The figure shows the assignments of each haplotype to morphospecies and to delineated entities from parsimony network analysis, GMYC, PTP, ABGD, and distance based-clustering. For the latter four, subclades analyses (A, C, E, M, Q, G, R) are shown in a separate column indicated by “S”. Color of cluster boxes indicate mismatch (red), agreement (green) or uncertainty (light blue, i.e., in case of females with uncertain ID) with the morphospecies assignment, diverse colors of morphospecies in clade C indicate identical non-monophyletic morphospecies. Black dots in the morphospecies column represent non-assignable female specimens. The right panel indicates θ for the respective GMYC entities. Sampling sites are indicated with dots on the map. Habitus images: *Pleophylla*, *Maladera*, and *Trochalus* (from top to bottom).

Figure 2. Plot of the lineage-specific population mutation rates (θ) vs. sampled number of haplotypes of the southern African Sericini. The dashed line indicates the linear regression line on the data.

Figure. 3. The fit of the GMYC model to *cox1* data of exemplarily selected subclades (A, M) and the complete Sericini data set (All). Top panel: LTT plot with GMYC threshold time. Bottom panels: likelihood surface of the GMYC model.

Figure 4. Simulations showing the effect of variable species abundance (but constant N_e) on GMYC species delimitation in relation to the number of sampled species. Each panel shows the match ratio of known to correctly inferred GMYC groups for subsamples drawn from the 100-species tree. Sampling from the species trees was

performed under random, clustered and clade-wise addition of GMYC groups (see Materials and Methods). Gene genealogies were then generated on the subsampled trees either under constant sample size of each species ($sd=0$), i.e., 5 individuals per species (left-hand panel), and under variable sample size of each species (right-hand panels), i.e., sampling a variable number of individuals from each species according to a log-normal distribution with $sd=1$, 1.5, and 2.0 (leading to an estimate of the proportion of singleton species of 13%, 40% and 52%, respectively). Regressions on plots of 30, 50 and 100 species are shown by red dotted, blue dashed, and black continuous lines, respectively.

Figure 5. Success of GMYC species delimitations in simulations under variable N_e in relation to the number of sampled species. Each panel shows the match ratio for subsamples drawn from the 100-species tree. Sampling from the species trees was performed under random, clustered and clade-wise addition of GMYC groups. Genealogies on the subsampled trees were simulated under increasingly greater skew of N_e (based on a log-normal distribution with $sd=1$, 1.5, and 2.0) but constant sample size (panel “constant”). In addition to variable N_e , the species sample size was variable (panel “variable”) with the numbers of samples proportional to the species' N_e that led to an increased number of singletons (13%, 40% and 52% singletons under $sd=1$, $sd=1.5$ and $sd=2$, respectively). Regressions on plots of 30, 50 and 100 species are shown by red dotted, blue dashed, and black continuous lines.

Figure 6. Lumping and over-splitting by the GMYC model in simulations in relation to the sampling bias: ratio of GMYC vs. true species compared with variable N_e for the different sampling schemes under constant and variable sample size.

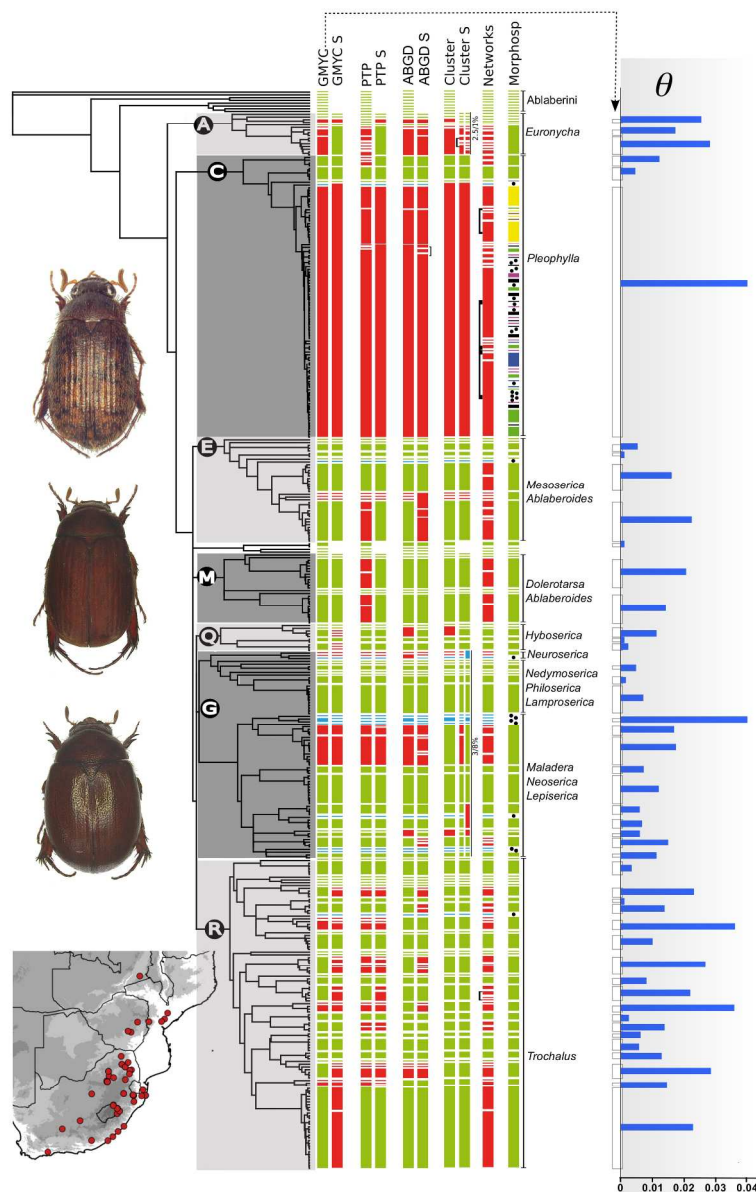


Figure 1. Ultrametric tree of the southern African Sericini species. The figure shows the assignments of each haplotype to morphospecies and to delineated entities from parsimony network analysis, GMYC, PTP, ABGD, and distance based-clustering. For the latter four, subclades analyses (A, C, E, M, Q, G, R) are shown in a separate column indicated by "S". Color of cluster boxes indicate mismatch (red), agreement (green) or uncertainty (light blue, i.e., in case of females with uncertain ID) with the morphospecies assignment, diverse colors of morphospecies in clade C indicate identical non-monophyletic morphospecies. Black dots in the morphospecies column represent non-assignable female specimens. The right panel indicates for the respective GMYC entities. Sampling sites are indicated with dots on the map. Habitus images: *Pleophylla*, *Maladera*, and *Trochalus* (from top to bottom).

297x422mm (300 x 300 DPI)

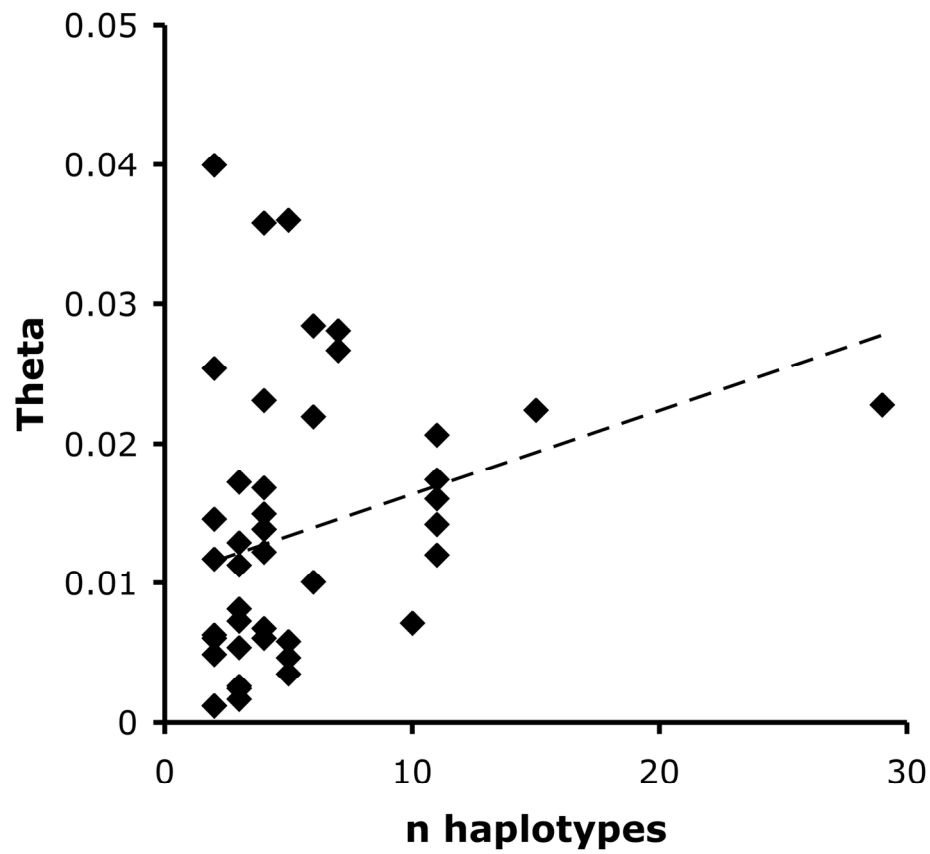


Figure 2. Plot of the lineage-specific population mutation rates () vs. sampled number of haplotypes of the southern African Sericini. The dashed line indicates the linear regression line on the data.
155x145mm (300 x 300 DPI)

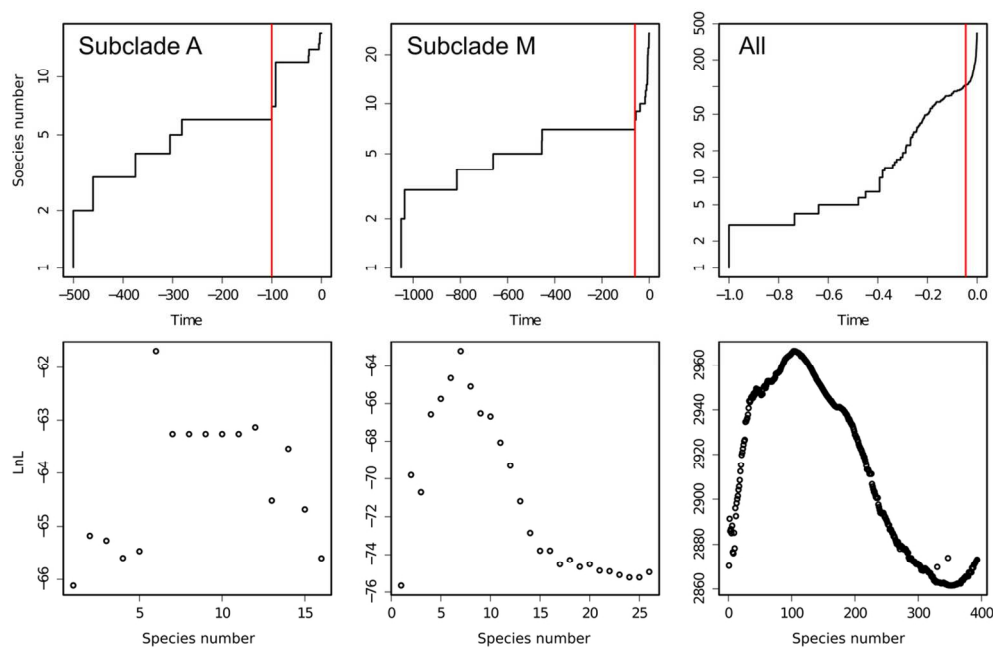


Figure. 3. The fit of the GMYC model to cox1 data of exemplarily selected subclades (A, M) and the complete Sericini data set (All). Top panel: LTT plot with GMYC threshold time. Bottom panels: likelihood surface of the GMYC model.
111x76mm (300 x 300 DPI)

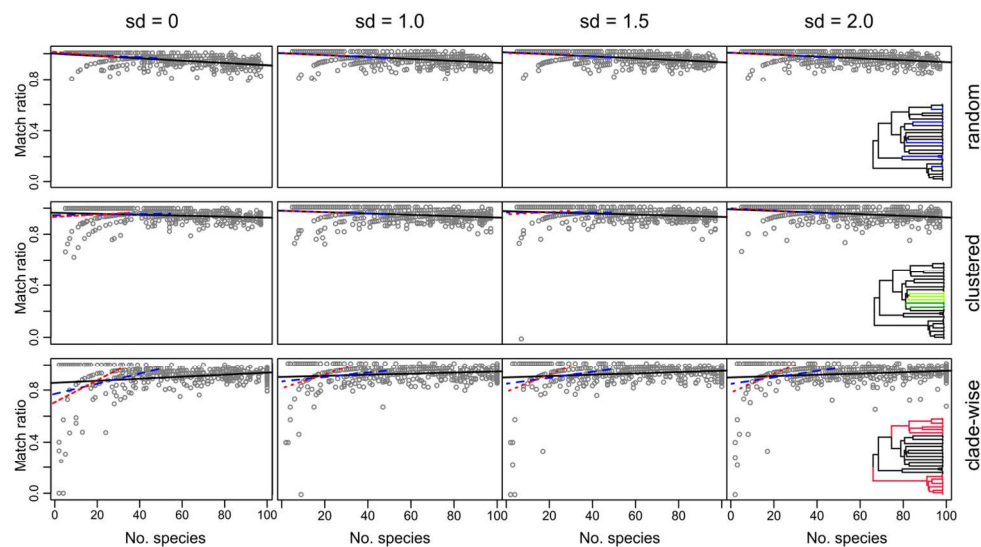


Figure 4. Simulations showing the effect of variable species abundance (but constant N_e) on GMYC species delimitation in relation to the number of sampled species. Each panel shows the match ratio of known to correctly inferred GMYC groups for subsamples drawn from the 100-species tree. Sampling from the species trees was performed under random, clustered and clade-wise addition of GMYC groups (see Materials and Methods). Gene genealogies were then generated on the subsampled trees either under constant sample size of each species ($sd=0$), i.e., 5 individuals per species (left-hand panel), and under variable sample size of each species (right-hand panels), i.e., sampling a variable number of individuals from each species according to a log-normal distribution with $sd=1$, 1.5, and 2.0 (leading to an estimate of the proportion of singleton species of 13%, 40% and 52%, respectively). Regressions on plots of 30, 50 and 100 species are shown by red dotted, blue dashed, and black continuous lines, respectively.

124x71mm (300 x 300 DPI)

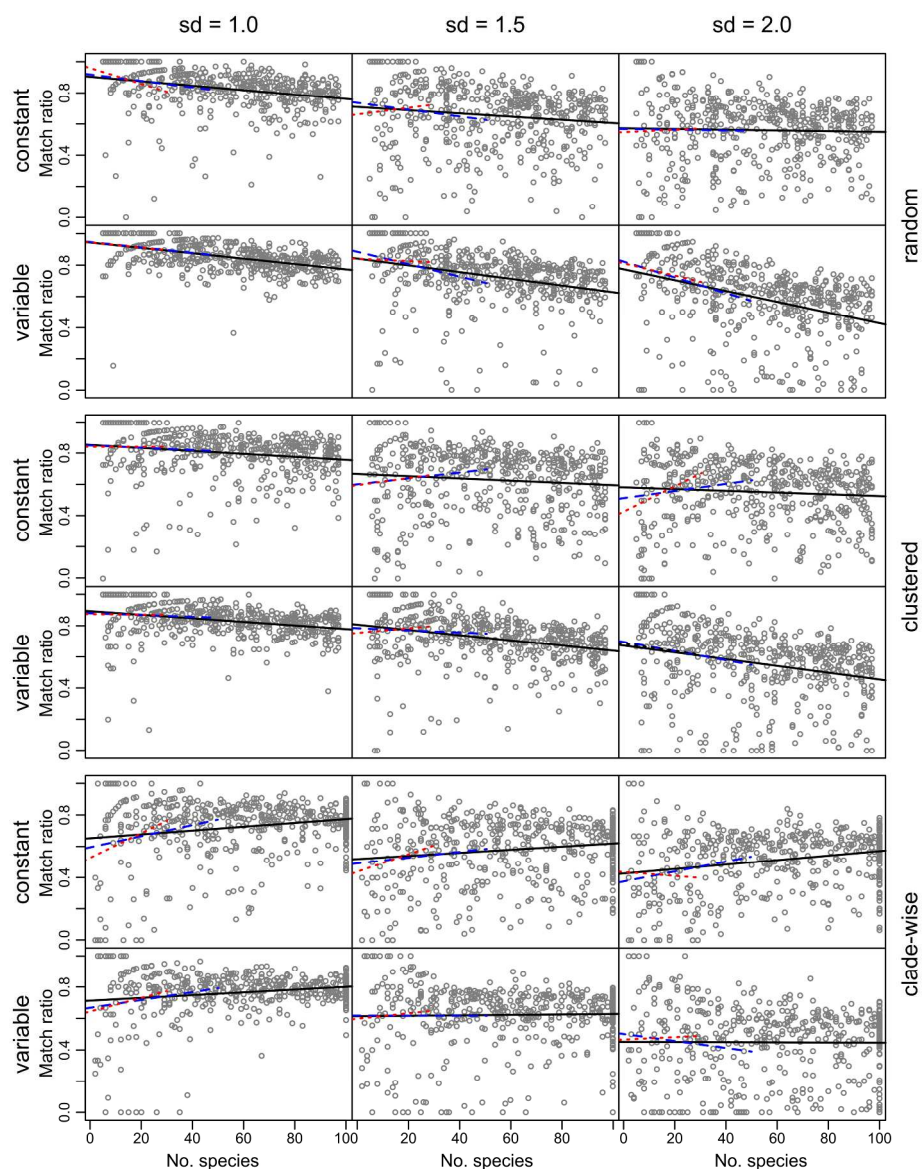


Figure 5. Success of GMYC species delimitations in simulations under variable N_e in relation to the number of sampled species. Each panel shows the match ratio for subsamples drawn from the 100-species tree. Sampling from the species trees was performed under random, clustered and clade-wise addition of GMYC groups. Genealogies on the subsampled trees were simulated under increasingly greater skew of N_e (based on a log-normal distribution with $sd=1$, 1.5 , and 2.0) but constant sample size (panel "constant"). In addition to variable N_e , the species sample size was variable (panel "variable") with the numbers of samples proportional to the species' N_e that led to an increased number of singletons (13%, 40% and 52% singletons under $sd=1$, $sd=1.5$ and $sd=2$, respectively). Regressions on plots of 30, 50 and 100 species are shown by red dotted, blue dashed, and black continuous lines.

228x293mm (300 x 300 DPI)

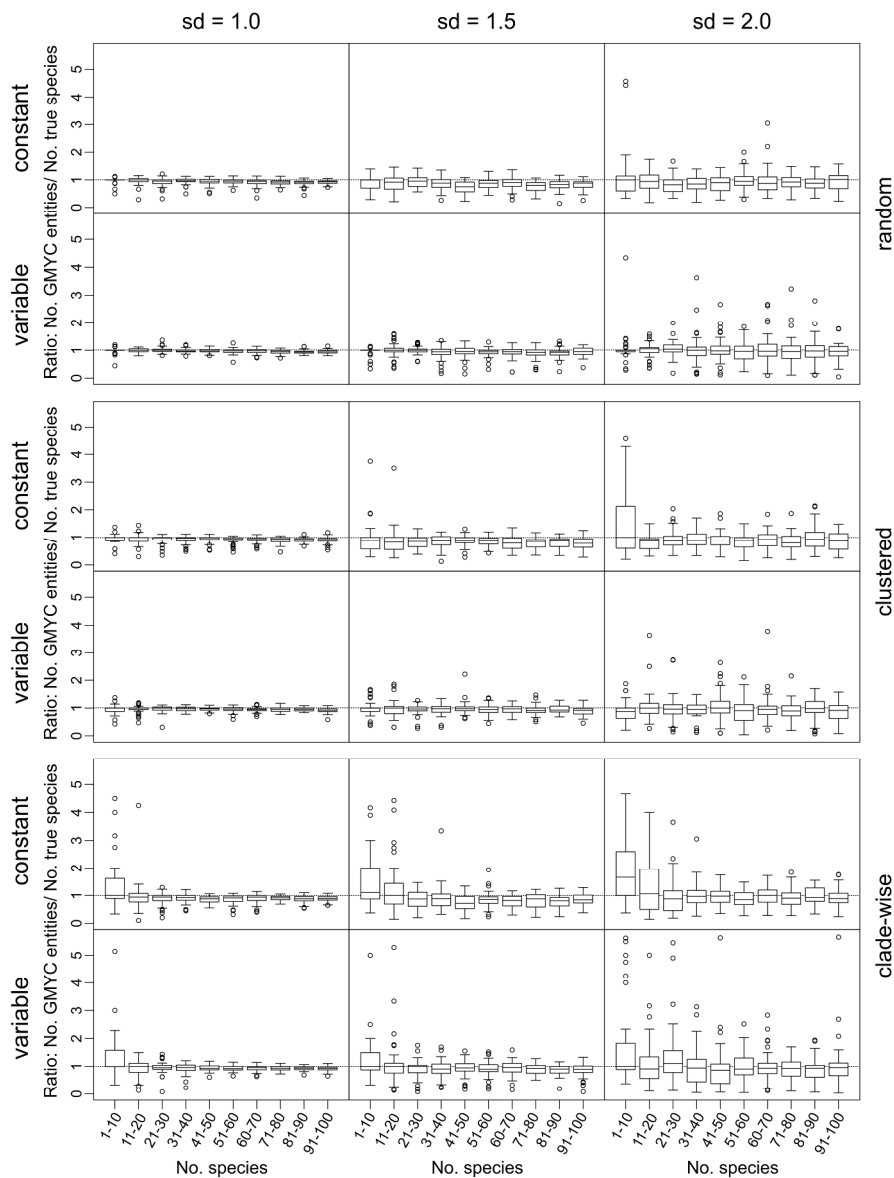


Figure 6. Lumping and over-splitting by the GMYC model in simulations in relation to the sampling bias: ratio of GMYC vs. true species compared with variable N_e for the different sampling schemes under constant and variable sample size.
230x303mm (300 x 300 DPI)