Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects

JOAN PONS,1,2,10 TIMOTHY G. BARRACLOUGH,2,3 JESUS GOMEZ-ZURITA,1,2,7 ANABELA CARDOSO,1,4,7 DANIEL P. DURAN,1,8 STEPHAN HAZZELL,1,2,9 SOPHIE KAMOUN,5 WILLIAM D. SUMLIN,6 AND ALFRIED P. VOGLER1,2

1Department of Entomology, The Natural History Museum, London SW7 5BD, United Kingdom; E-mail: a.vogler@nhm.ac.uk (A.P.V.)
2Division of Biology and NERC Centre for Population Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, United Kingdom
3Jodrell Laboratory, Royal Botanic Gardens, Kew TW9 3DS, United Kingdom
4Facultad de Ciencias da Universidade de Lisboa, Departamento de Biologia Animal, Centro de Biología Ambiental, Rua Ernesto Vasconcelos, 1749-016, Campo Grande, Lisboa, Portugal
5Department of Plant Pathology, Ohio State University, Ohio Agricultural Research and Development Center, Wooster, Ohio, 44691, USA
6Department of Entomology, Texas A&M University, College Station, Texas 77843, USA
7Present Address: Area de Biología Animal, Departamento de Zoología y Antropología Física, Facultad de Biología, Universidad de Murcia–Campus de Espinardo, 30071 Murcia, Spain
8Present Address: Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, 37235, USA
9Present Address: Division of Zoology, School of Animal and Microbial Sciences, University of Reading, Reading RG6 6AJ, United Kingdom
10Present Address: Unitat de Biologia Evolutiva, Facultat de Ciencias de la Salut i de la Vida, Universitat Pompeu Fabra, C/Dr. Aiguader 80, 08003 Barcelona, Catalonia, Spain

Abstract.—Cataloging the very large number of undescribed species of insects could be greatly accelerated by automated DNA-based approaches, but procedures for large-scale species discovery from sequence data are currently lacking. Here, we use mitochondrial DNA variation to delimit species in a poorly known beetle radiation in the genus *Rivacindela* from arid Australia. Among 468 individuals sampled from 65 sites and multiple morphologically distinguishable types, sequence variation in three mtDNA genes (cytochrome oxidase subunit 1, cytochrome *b*, 16S ribosomal RNA) was strongly partitioned between 46 or 47 putative species identified with quantitative methods of species recognition based on fixed unique (“diagnostic”) characters. The boundaries between groups were also recognizable from a striking increase in branching rate in clock-constrained calibrated trees. Models of stochastic lineage growth (Yule models) were combined with coalescent theory to develop a new likelihood method that determines the point of transition from species-level (speciation and extinction) to population-level (coalescence) evolutionary processes. Fitting the location of the switches from speciation to coalescent nodes on the ultrametric tree of *Rivacindela* produced a transition in branching rate occurring at 0.43 Mya, leading to an estimate of 48 putative species (confidence interval for the threshold ranging from 47 to 51 clusters within 2 logL units). Entities delimited in this way exhibited biological properties of traditionally defined species, showing coherence of geographic ranges, broad congruence with morphologically recognized species, and levels of sequence divergence typical for closely related species of insects. The finding of discontinuous evolutionary groupings that are readily apparent in patterns of sequence variation permits largely automated species delineation from DNA surveys of local communities as a scaffold for taxonomy in this poorly known insect group. [Phylogenetic species concept; coalescence; mtDNA; Cicindelidae; Australia; paleoclimate.]

Automated DNA sequencing procedures could greatly accelerate identification of known species (“DNA barcoding”; Hebert et al., 2003) and provide a framework for the taxonomy of poorly known groups (Tautz et al., 2003). However, sequencing-based approaches in taxonomy are useful only if an individual’s DNA sequence can unambiguously associate that individual with a larger group of organisms (a species or higher taxon). Establishing this association is complicated due to intraspecific variation and the possible incongruence of gene and species histories that might prevent the easy recognition of these groups, in particular where only a single fragment of mtDNA is used (Lipscomb et al., 2003; Mallet and Willmott, 2003; Will and Rubinoff, 2004). In current taxonomic DNA sequencing, species-level entities have been accepted to be known a priori, and their DNA signature is based on genotypes from a representative sample of well-identified individuals (Hajibabaei et al., 2006; Hebert et al., 2003; Meyer and Paulay, 2005).

“DNA barcoding” therefore accepts that species limits are established following traditional practices of taxonomy, usually based on morphology, and the DNA information is fitted into this system of predefined taxonomic groups. This approach is problematic because the correspondence of sequence variation with existing Linnean binomials is expected to be inexact. This may be because of the existence of unrecognized cryptic species (Hebert et al., 2004a, 2004b) or incongruence of gene histories confounding species recognition (Funk and Omland, 2003; Meyer and Paulay, 2005; Monaghan et al., 2006) or because procedures for species delimitation of the traditional taxonomy have been vague or species concepts have been applied inconsistently (Agapow et al., 2004; Cracraft, 1992; Meyer and Paulay, 2005; Wiens and Penkrot, 2002). To avoid these problems, a DNA-based taxonomic system should use the sequence information itself as the primary information source for establishing group membership and defining species boundaries.

Current procedures for delimiting species from sequence data aggregate populations lacking discrete differences into a single species (Cracraft, 1983; Davis and Nixon, 1992; Sites and Marshall, 2003; Wiens and Penkrot, 2002). These “smallest detected sample[s] of self-perpetuating organisms that have unique sets of characters” (Nelson and Platnick, 1981) can be recognized based on fixed nucleotide differences unique to
such (sets of) populations (the criterion of “diagnos-
ability”; Cracraft, 1983), implemented in population ag-
gregation analysis (PAA; Davis and Nixon, 1992). A
tree-based variant of this procedure (cladistic haplotype
analysis; CHA) also considers homoplastic character
states as diagnostic if they are shown to be uniquely de-

derved for a particular group (Brower, 1999). The Wiens-
Penkrot (WP) method (Sites and Marshall, 2003; Wiens
and Penkrot, 2002) is also a tree-based method for delim-

ing phylogenetic species but uses the topology (rather
than character polarity, as in CHA) to assess whether or
not clades are restricted to a (set of) populations to the
exclusion of clades elsewhere (criterion of “exclusivity”).

The prior definition of populations, usually defined
globally, is a critical step in all of these methods,
although the focus on individual organisms rather than

a preconceived interbreeding “group” would be more
objective for species delimitation (Vrana and Wheeler,
1992). Current methods for estimating population cohe-
rence from the sequences themselves rely on evidence for
recombination to determine species limits (Doyle, 1995)
but are impractical for large-scale analysis at present. An
alternative possibility to detect species separation is to
search for discontinuities in sequence variation associ-
ated with the species boundaries. For example, statistical
parsimony analysis (Templeton et al., 1992) separates
groups of sequences into different sequence networks
if genotypes are connected by comparatively long
branches that are affected by homoplasy. Although ho-

moplasious connections do not necessarily correspond to
species boundaries, in practice this algorithm often does
separate groups that are roughly coincident with named
species or species groups (e.g., Cardoso and Vogler, 2005;

Templeton, 2001; Wilder and Hollocher, 2003; Templeton
et al., 1995), and the WP method for species delimitation
is now recommended to be applied only in conjunction
with statistical parsimony analysis in order to reduce the
complexity of large data sets (Sites and Marshall, 2003).

These observations would suggest the possibility of
using analyses of branch lengths on a DNA tree for ex-
plicit tests of species boundaries, based on the difference in
branching rates at the level of species and populations.
Branch lengths between species are determined by spe-
ciation and extinction rates (macroevolution) (Nee et al.,
1994), whereas branch lengths within a species reflect co-
alescence processes at the level of populations (microevo-
lution) (Hudson, 1991; Rosenberg and Nordborg, 2002;

Wakeley, 2006). Well-developed approaches exist for an-
alyzing branching rates in either framework. Combining
equations that describe processes of lineage birth at the
species level with coalescence models within species, it is
possible to develop a statistical framework for estimating
the predicted shift in dynamics of branching associated
with the species boundary. Here we develop a method
that determines the locations of ancestral nodes that de-

fine putative species and applies a likelihood ratio test
to assess the fit of the branch lengths to a mixed lineage
birth-population coalescence model.

As an example of a group whose current taxon-
omy is incomplete, we attempted DNA-based species
delineation in tiger beetles in the genus Rivacin del a
(Coleoptera: Cicindelidae). These beetles are found in
saline habitats and temporal lakes of interior Australia.
To date, 24 species of Rivacin del a have been formally de-
scribed but because of the inaccessibility of their habi-
tat, many species still await discovery (Freitag, 1979;

Sumlin, 1997). Populations are restricted to habitat near
the water edges in an otherwise inhospitable desert en-
vironment. Although similar in their lifestyle as general
predators of small arthropods, species and populations
differ in color pattern, body shape, habitat preference,
and behavioral traits. Most species are flighted, but Rivacin del a includes several flightless lineages that are ex-
tremely fast runners, attaining speeds of up to 170 times
their body length per second (Kamoun and Hogenhout,
1996). The analysis of morphological variation in this
group is complicated due to low character variation and
nondiscrete differentiation. We hence conducted
comprehensive DNA sequencing across the geographic
range of Rivacin del a to provide the scaffold for assign-
ing species membership. We find that the various ex-
isting procedures for DNA-based species delimitation
identify biologically meaningful groups. The new like-
lihood method for analyzing branching rate produced
results closely similar to those obtained with character-
based methods but does not require prior assumption of
population boundaries and can provide a statistical mea-
ure of confidence to the assessment of species limits.

MATERIAL AND METHODS
Collecting Information and DNA Procedures
Specimens were sampled at the edges of temporal salt
lakes and salt flats in the Provinces of Western Australia,
South Australia, and the Northern Territory during two
expeditions in 2001 and 2003. Several of these sites were
surveyed for cicindelids for the first time, and collec-
tions produced many apparently undescribed species
and variants (Fig. 1). Specimens were placed directly into
absolute ethanol after cursory identification in the field.
Up to six individuals (more in a few cases) from each
locality were sequenced, plus an additional set of up
to six individuals if obvious morphologically different
forms were recognized during field work at a given site
or where individuals were obtained in different years or
at multiple adjacent sites on the larger lakes (Supplemen-
tary Table S1; http://systematicbiology.org). This sam-
pling regime was thought to be a good reflection of the
total diversity encountered at the various collecting sites
capturing species-level groups and their internal vari-
ation, given the absence of a thorough taxonomic treat-
ment, while keeping the sequencing effort to a minimum.
In total, 468 specimens assigned to 108 local sets of in-
dividuals (65 sites, plus morphologically distinct forms
and samples from repeat visits of a site) were included
in the analysis.

Nondestructive DNA extraction was performed us-
ing a Qiagen DNeasy kit. Whole specimens were soaked
overnight in extraction buffer at 37°C, with small per-
forations made to the side of the abdomen and DNA
extracted from the supernatant. Damage to specimens (some of them to be holotypes in future species descriptions) was minimal, and softness of specimens was maintained if transferred to 70% ethanol after the extraction. Vouchers will be maintained as pinned dry specimens in the W. D. Sumlin personal collection. Amplification of three mtDNA regions from the cytochrome oxidase subunit 1 (cox1), cytochrome b apoenzyme (cob), and 16S ribosomal RNA (rrnL) plus adjacent regions, was generally successful with well-established oligonucleotides (Pons et al., 2004) but required newly designed cox1 and rrnL primers in some cases, the latter to avoid coamplification of a nuclear paralog (Pons and Vogler, 2005). (The genetic nomenclature in this paper follows Boore, 2001). Primer combinations for each region are given in Supplementary Table S2. PCR fragments were sequenced in both directions on an ABI377 automated sequencer. The sequences reported in this paper have been deposited in GenBank under accession nos. AJ617921–AJ618351 (cox1), AJ618352–AJ618766 (cob), AJ619087–AJ619548 (rrnL). In addition, three species of Australian cicindelids established to be closely related to Rivacindela based on a wider survey of Cicindela sensu lato (unpublished) were used as outgroups (AJ831553–AJ831564). The data and tree files have been uploaded at Treebase under SN2798-10988.

Data Analysis
Statistical parsimony analysis was carried out with TCS version 1.3 software (Clement et al., 2000) using only individuals with complete sequence information for all three genes. Where a separate lake sample did not include at least one specimen with complete sequence information (listed in parentheses in Table S1), these were assigned to the networks based on their phylogenetic position in standard parsimony analysis (below). Statistical parsimony analysis partitions the data into independent networks of haplotypes connected by changes that are non-homoplastic with a 95% probability (Templeton, 2001). Intragroup and pairwise intergroup divergences were calculated with Mega version 2.1 (http://www.megasoftware.net) using p-distances. $F_{st}$ and $P$ values at 0.05 significance were calculated from pairwise differences in Arlequin 2.0 (http://lgb.unige.ch/arlequin).

Parsimony tree searches on the complete data set were conducted using PAUP*4.0b10 (Swofford, 2002) with
40 independent runs and 200 ratchet iterations (Nixon, 1999) each, with 15% of characters reweighted using the default settings of the PAUPRat script (http://www.ucalgary.ca/~dsikes/software2.htm). The preferred topology was selected among an initial set of parsimony trees, from which one was selected according to the Shimodaira-Hasegawa (SH) test under a GTR+I+Γ likelihood model and used as starting tree for an ML search in PAUP*4.0b10. The ML search was conducted on a desktop computer for 3 days testing a total of 18,622 arrangements. Where nodes were consistent with the ML tree, bootstrap values were given based on parsimony analysis obtained by searching 100 pseudoreplicated data sets (generated using SEQBOOT in Phylip 3.57) using the ratchet. A strict consensus tree was created from these trees for each pseudoreplicate, and the 100 strict consensus trees each representing a bootstrap replicate were used to create a 50% majority rule consensus tree providing the support values. Relative ages of nodes were estimated using the r8s software (Sanderson, 2002) by fitting branch lengths of an ML tree using penalized likelihood and a smoothing parameter of 10, chosen as optimal by cross-validation. Absolute ages were estimated setting the split of *R. aurifodina* (sample 204) and *R. salicursoria* (sample 208) to 3.2 Mya (Pons et al., 2004). This was based on an estimate from a phylogeny of worldwide lineages of *Cicindela sensu lato* calibrated from biogeographic evidence in North American lineages which included only these two species to represent *Rivacindela* (Barraclough and Vogler, 2002). Sequence divergence between individuals of these species was 6.0% (SD 0.14%) and hence the estimate corresponds closely to the widely used insect mtDNA molecular clock calibration of 2.3% per My (Brower, 1994). Lineage-through-time analyses were obtained as a semilogarithmic plot of the number of lineages against time since the first bifurcation (Nee et al., 1992).

Species Delimitation

Different quantitative methods for species delimitation were applied, implemented by visual inspection of the variable nucleotide positions and trees derived from these. These procedures were conducted only within the independent networks defined in the statistical parsimony analysis, greatly reducing the complexity of scoring separated groups. Population profiles of character variation were established according to Sites and Marshall (2003) as the basis for PAA (Davis and Nixon, 1992) and CHA (Brower, 1999), the latter by assessing variable characters on the likelihood tree shown below. The WP method also used this tree to delimit “exclusive” populations, defined as the monophyly of geographically restricted genotypes to the exclusion of clades elsewhere. F_{st} values were used for aggregating samples with non-significant pairwise F_{st} in analogy to the grouping of populations under PAA (Supplementary Fig. S1).

Analyses of mtDNA Branching Times

A statistical model was developed to test for the predicted change in branching rates at the species boundary. The overall aim of the procedure is to classify the observed branching time intervals defined by the nodes in a clock-constrained phylogram to either being the result of inter-specific (“diversification”) or intraspecific (“coalescent”) processes of lineage branching (Fig. 2). A full description of the model and its performance on simulated trees will be provided elsewhere (Barraclough, unpublished).

As raw data, the model uses the waiting times between successive branching events on the ultrametric DNA tree, \( x_i \). We combine standard models that separately consider branching within populations (Hudson, 1991; Wakeley, 2006) and branching between species (Nee, 1994, 2001; Nee et al., 1994). Under a neutral coalescent, the likelihoods of the waiting times within a single population with effective population size \( N_e \) and \( n_i \) lineages present during waiting time \( i \) are given by:

\[
L_{(x_i)} = \lambda n_i (n_i - 1) e^{-\lambda n_i (n_i - 1) x_i},
\]

where the birth rate

\[
\lambda = \frac{1}{2N_e}
\]
The simplest standard approach for considering branching between species is as a Yule model (Yule, 1924); i.e., a stochastic birth-only model. The likelihoods of the waiting times in a species phylogeny (one tip per species) of a clade with constant average speciation rate, $\lambda$, and no extinction are given by:

$$L_{(x)} = \lambda n_{i} e^{-\lambda n_{i} x_{i}}$$  \hspace{1cm} (3)

We combine the above equations describing population and speciation processes to consider a clade that has diversified into $k$ species, each of which can be treated as a single population with effective size, $N_{ij}$, $j = 1$ to $k$. Assuming a constant speciation rate without extinction, and neutral coalescence within each species, the likelihoods of waiting times in the entire tree under this mixed Yule coalescent (MYC) model are given by:

$$L_{(x)} = b e^{-b x_{i}}$$  \hspace{1cm} (4)

where

$$b = \lambda_{k+1} n_{i,k+1} + \sum_{j=1,k} (\lambda_{j} n_{ij} (n_{ij} - 1))$$  \hspace{1cm} (5)

where $k+1$ is the index assigned to the diversification process and $n_{ij}$ is the number of lineages in waiting interval $i$ belonging to process $j$. $\lambda_{k+1}$ is the speciation rate and $\lambda_{j}$ are the branching rates for each coalescent process as defined in Equation 2. We do not consider effective population sizes explicitly hereafter, rather the coalescent branching rate parameter for each species. The term $b$ is the probability that an event of any type happens at the end of the waiting interval and $e^{-bx}$ is the probability that no event happened during the waiting interval (see appendix 1 of Nee, 2001).

This MYC model can be fitted to an ultrametric tree by maximizing the sum of the log-likelihoods of waiting times across the entire tree. A key step is fitting the location of the switches from speciation to coalescent nodes; i.e., the most recent common ancestral node defining each species. The simplest approach is to assume that there is a threshold time, $T$, before which all nodes reflect diversification events and after which all nodes reflect coalescent events. Species in this model are thus delimited by the descendents of nodes branching crossing the threshold. $T$ can be optimized to find the maximum likelihood solution and hence to estimate the number of species. Note that this approach does not assume that all species have the same age of their most recent common ancestor, which would not be the expected even under an equal population size model, but rather that the most recent diversification event occurred before the oldest within-species coalescent event. Approximate 95% confidence intervals for the parameters can be calculated by finding solutions with 2 log-likelihood units of the maximum (Edwards, 1972).

To test whether there is significant evidence for the predicted transition in branching rates, the likelihood for the threshold model can be compared to that obtained assuming no threshold; i.e., a single branching process for the entire tree ($k = 0$). Assuming initially that all species have the same effective population size, the threshold version of the MYC model introduces two additional parameters compared to the null model: an additional branching rate parameter, $\lambda_{2}$, and $T$. A standard log-likelihood ratio test can be used to assess whether the alternative model provides a significantly better fit than the null model of no such shift in branching process: twice the difference in log-likelihood is expected to be chi-square distributed with 2 degrees of freedom (Goldman, 1993). Failure to reject the null model could have several explanations. First, the clade might in fact represent a single species (unlikely in the present case). Second, the observed branching rate within species depends on the number of individuals sampled per species: small samples per species will weaken the power to detect the transition. Conversely, incomplete sampling at the species level would reduce the apparent branching rate in $\lambda_{1}$. Third, some combinations of actual branching processes will also make it harder to detect the transition; for example, a combination of fast speciation rate and large population sizes.

Equations (4) to (6) make strict assumptions about both speciation and population processes that seem unlikely to hold in many real clades. As a general solution to this problem, we introduce additional parameters that allow for a qualitatively wide range of different models for both types of branching processes. Nee et al. (1994a) and Nee (2001) discussed transformations of waiting intervals that allow derived models to be treated as exponential processes with a single rate parameter. These accounts imply a simple general transformation of the MYC model by substituting $b$ in Eq. (4) with $b^{*}$:

$$b^{*} = \lambda_{k+1} (n_{i,k+1})^{p_{k+1}} + \sum_{j=1,k} (\lambda_{j} (n_{ij} (n_{ij} - 1))^p_{j})$$  \hspace{1cm} (6)

We call this the general mixed Yule coalescent (GMYC) model. The $p_{j}$ represent scaling parameters that can be optimized during model fitting. Interpretation depends on which class of branching events are considered. The scaling parameter for the diversification process, $p_{k+1}$, provides similar information to the gamma statistic of Pybus and Harvey (2000). $p_{k+1} = 1$ represents a constant speciation rate model with no extinction. $p_{k+1} > 1$ indicates an apparent increase in diversification rate towards the present, which might reflect, for example, a real increase in speciation rate or the effects of constant background extinction (Barraclough and Nee, 2001; Nee, 1994). $p_{k+1} < 1$ represents an apparent decrease in diversification rate towards the present, which might reflect, for example, a real slow-down in speciation or the effects of incomplete sampling of species within the clade (Nee et al., 1994b; Pybus and Harvey, 2000). The scaling parameters for the coalescent processes within each species are interpreted differently. $p_{j} = 1$ represents a neutral coalescent model. $p_{j} < 1$ indicates a relative deficit of
recent coalescent events, expected, for example, if pop-
ulations were growing in size or experiencing balancing
selection (Nee et al., 1994). \( p_i > 1 \) indicates a relative ex-
cess of recent coalescent events, expected, for example,
if all populations were declining in size, following a re-
cent selective sweep affecting the marker or if there is
further population structure within the species. Hence,
optimizing across possible values of \( p \) for both classes of
branching events relaxes the assumptions of the method
and can provide pointers for further analyses to explore
possible causes for departure from the simplest model.

Code implementing the model in R using functions
from the APE library is available from TGB. First, we
visualized the waiting time data by plotting the log of
the number of lineages through time (Nee et al., 1992). A
transition in branching rates should be visible as a sud-
den increase in slope of the plot towards the present. Sec-
ond, we ran the threshold version of the GMYC model on
the calibrated ultrametric tree of Rivacindela. Third, we
compared the likelihood to that obtained assuming a sin-
gle branching process for the tree. For the GMYC model,
assuming all species have the same parameter values, the
threshold model has five parameters \((\lambda_1, \lambda_2, p_1, p_2, T)\),
whereas the null model has two \((\lambda_1\) and \(p_1\)); hence, there
are three degrees of freedom for the comparison.

results

Sequencing of three gene regions of mtDNA for the se-
lected 468 specimens produced a data matrix of 1914 bps,
of which 514 bps were informative. The average uncor-
corrected \( p \)-distance between any two ingroup sequences
was 5.8% (7.8%, 9.4%, and 2.8%, respectively, for the
rected \( p \)-distance between any two ingroup sequences
of which 514 bps were informative. The average uncor-
lected 468 specimens produced a data matrix of 1914 bps,
ablative species of similar extent, and those obtained from
728 bp fragment of the \( cox1 \) gene most closely matched
the result from the full data set. Average pairwise diver-
gence within these putative species determined by
the WP method was 0.5%, much lower than the average
among-species divergence of 6.3%, and 2.2% between sister
species.

Population genetic analysis was also used for ag-
gregation of populations with non-significant pairwise
\( F_{st} \) (Supplementary Fig. S1), in analogy to PAA/CHA.
Among the 66 groups with four or more individuals
included in the analysis, under a significance level for \( F_{st} \)
\( P < 0.05 \), this analysis led to recognition of only five ad-
tional entities over those in the WP analysis (Fig. 4).
Bonferroni corrections for multiple comparisons would
increase the significance level for \( F_{st} \) to \( P = 0.00076 \),
which would lead to the recognition of fewer subdivided
groups, although this value is likely to introduce a large
ype 2 error, i.e., the lumping of groups separated at a
significant \( F_{st} \) level. Instead, using uncorrected pairwise
\( F_{st} \) might result in type 1 error of accepting groups that
appear separated simply due to chance, and therefore
the reported results (Fig. 4) constitute the upper bound
of the number of entities.

The existence of distinct mtDNA lineages was con-
firmed by analysis of branch lengths. A lineage-through-
time plot showed a steep upturn in branching rates
towards the present, presumably marking the transi-
tion from between-species to within-species rate of line-
age branching (Fig. 5). The position of the transition
was fitted using the GMYC model assuming a sin-
gle threshold: pretransition nodes were assumed to re-
fect speciation events and posttransition nodes were
assumed to reflect within species coalescence. The model
fitted a transition in branching rate occurring at 0.43
Mya, leading to an estimate of 48 putative species, 5
of which contain a single individual. Confidence limits
for the threshold ranged from 0.39 to 0.53 Mya and
for the estimated number of species from 47 to 51.
The GMYC model was preferred over the null model
of uniform branching rates \((\log L = 2641)\), compared to
null model \( \log L = 2606; \Delta L = 71.8, \chi^2 \) test, d.f. = 3,
\( P < 0.001 \). The groups of sequences identified by these
shifts matched closely with those identified through other
aggregation analyses (Fig. 4). There were six cases in
which the WP method delimited species embedded
within a paraphyletic species that were not detected by
the branch-length method. Conversely there were seven
cases of divergence (sometimes a single individual) de-
tected by the branch-length analysis but not by the WP
method.
Figure 3. ML tree depicting relationships of *Rivacindela* mtDNA haplotypes with branch lengths fitted assuming a molecular clock. The two main sister groups are shown separately (panel A, Eastern group; panel B, Western group). Localities for each of 47 putative species (“Clades”) obtained under the WP method are indicated by a two-letter code. Different shading is used for better visibility of the extent of groups. Gray bars and site names colours indicate widely distributed or paraphyletic species. The line style indicates whether branches were estimated as between-species branching (stippled) or within-species branching (solid) in the likelihood procedure. The dotted vertical line shows the maximum likelihood transition point of the switch in branching rates. The grey shading indicates the confidence limits for the transition point falling within 2 log-likelihood units of the ML solution (Material and Methods). Numbers above nodes represent bootstrap support values based on 100 pseudoreplicates and ratchet parsimony searches, shown only for the deep level clades. Bootstrap values under 50% are not shown. (Continued)
FIGURE 3. (Continued)
FIGURE 4. Species entities obtained by various procedures of grouping: Network statistical parsimony (Templeton et al., 1992), Population Aggregation Analysis (PAA) (Davis and Nixon, 1992), Cladistic Haplotype Analysis (CHA) (Brower, 1999), Wiens-Penkrot method (WP) (Sites and Marshall, 2003; Wiens and Penkrot, 2002), haplotype frequencies ($F_{ST}$), and the new likelihood method based in the analysis of mtDNA branching times. Grey blocks indicate the aggregation of samples from the sites listed at the top (a “+” sign after the site name indicates where more than one site was aggregated in a group). The darker grey fields indicate groups linked in a particular subset. WP all data refers to the groupings based on the full data set (numbers in each field correspond to the clades in Fig. 3). The results from WP analysis conducted separately on the three mtDNA subsets (cox1, cob, rrnL), are also shown. For $F_{ST}$, the figure shows the result of pairwise comparisons between all populations of four or more individuals, with fields left empty where fewer individuals were analyzed. For the likelihood analysis, the numbers in the grey fields indicate additional groupings that are recognized with this procedure, further subdividing groups shown as individual blocks. The bottom two lines of the graph show the number of individuals and number of sites included in each group.

The basic GMYC model assumes that all species have the same coalescent branching rates, $\lambda_j$, and scaling parameters, $p_j$. To check the robustness of the results to this assumption, we repeated the threshold optimization procedure but allowed all the putative species to have different values of these parameters. This yields a model with 89 parameters; therefore, we simplified this model by classifying the species into three categories based on their estimates of $p_j$ using the k-means clustering function in R. This simplified model yielded an estimate of 48 species, with confidence limits of 46 and 52 and a significantly better fit than the GMYC model that assumed constant parameter values across species ($LogL = 2652$, $\Delta L = 2652 - 2641, 2\Delta L = 21; \chi^2$ test, d.f. = 4, $P < 0.001$). Ignoring the putative species with single individuals sampled, 32 putative species fell in a category with $p_1 = 0.03$, 8 fell in a category with $p_2 = 0.3$, and 3 fell in a category with $p_3 = 1.0$. Hence, overall there was evidence for a deficit of recent branching events within most of the sampled species (constraining all species to obey neutral coalescent $p_1$ to $p_k$ fixed at 1 reduces $LogL$ to 2517, compared to three-category model, $2\Delta L = 270; \chi^2$ test, d.f. = 5, $P \ll 0.001$). For between-species branching, we obtained $P = 1.3$, signifying a slight increase in net diversification rate towards the tips relative to a constant speciation rate model.

FIGURE 5. Lineages-through-time plot based on the time calibrated tree obtained from all 468 haplotypes. The sharp increase in branching rate, corresponding to the transition from interspecies to intraspecies branching events, is indicated by the dotted line.
Most putative species identified in the analysis were restricted to a single lake, but four widely distributed species were present at between 5 and 13 sites, showing contiguous ranges that included all nearest neighbor sites (Fig. 1). Geographical structure was apparent also at deeper levels in the tree. For example, the basal node defined two major clades roughly coincident with eastern and western localities (Fig. 3). Sister species and larger sister clades tended to be found mostly at neighboring lakes or lake systems. A major divide separating the drainage systems (van de Graaff et al., 1997) was also apparent in the distribution of haplotypes. In western Australia, most populations of the same basin or paleo-river system were included in the same clade composed of closely related haplotypes. Haplotypes up to approximately 0.7 My of divergence were almost entirely confined to a single paleo-river, whereas populations from different drainage systems were not closely related, even where these lakes were in close geographic proximity (Fig. 6). Although most lakes harbored only one species, a maximum of three co-occurring species was observed, but these were not closely related in most cases.

**DISCUSSION**

The use of short mtDNA sequences in taxonomic studies, including “DNA barcoding” (Hebert et al., 2003), has been applied mostly to known species, but beyond novel tools for identification (and occasional evidence for cryptic species; Hebert et al., 2004a, 2004b; Meyer and Paulay, 2005), this does not directly bear on the taxonomic status of groups (Moritz and Cicero, 2004). Here we used quantitative analyses of sequence data to delimit putative species, whereby the sequence information itself provided a framework for alpha taxonomy. Species delimitation relied on two principal approaches, including established procedures based on diagnostic characters (Sites and Marshall, 2003) and a novel procedure based on detecting the shift in the rate of lineage branching. Other studies have observed similar shifts in branching rates as signatures of the species boundary (e.g., Acinas et al., 2004; Barraclough et al., 2003; Cardoso and Vogler, 2005; Hebert and Gregory, 2005; Hugall et al., 2002; Monaghan et al., 2005; Wiens and Penkrot, 2002). However, to our knowledge there have been no previous attempts to utilise this information in a quantitative procedure of species delimitation, despite the development of methods for clustering of sequences (Blaxter, 2004) and matching them to predefined taxonomic groups (Steinke et al., 2005; Matz and Nielsen, 2005).

The branch-length analysis relies on a probabilistic model separating species diversification (phylogeny) from coalescent processes (genealogy within species). In this sense, the shift in branching rate corresponds to the elusive boundary between Hennig’s (1966) phylogenetic and tokogenetic realms, indicating divergent and reticulate relationships. Our goal was to produce a method capable of fitting approaches of phylogenetics and population genetics, as the two major disciplines operating above and below this boundary (Brower et al., 1996). In its current implementation, the approach is based on simple assumptions such as a constant speciation rate model or neutral coalescence, although these are relaxed somewhat by incorporating the scaling parameter. The procedure has acceptable false-positive rates when applied to data simulated assuming no transition in branching rates and reasonable power for data simulated assuming distinct species (Barraclough, unpublished). The approach could be modified to specify detailed models of speciation, extinction, and population processes, and then to choose among competing models, but because of the bewildering range of possible scenarios (Barraclough and Nee, 2001; Charlesworth et al., 2003), this was not further explored here.

The analysis of branch length overcomes two principal problems of existing methods for quantitative species delimitation. First, the approach takes into account uncertainty of species limits by permitting confidence intervals when allocating species-defining nodes, a desirable property where species limits are weakly developed (Hey et al., 2003). Population aggregation methods do not allow for uncertainty, but rather optimize species limits based on a fixed criterion. Other recent population-based approaches take uncertainty into account (e.g., Matz and Nielsen, 2005) but have yet to be applied to species delimitation per se rather than identification of unknown sequences. Second, the branch length–based methods do not require that populations are defined a priori before being subjected to tests of aggregation (Davis and Nixon, 1992; Mayden, 1997). Definitions of population units remain problematic, as they frequently lack easily discernible spatial or genetic boundaries (Schaefer, 2006); e.g., where distributional ranges are insufficiently studied or where morphologically cryptic forms are found in the same geographical locality. For our study of Rivacindela we simplistically assumed that individuals collected at different sites were members of separate populations, whereas all individuals from a single site were considered a uniform population, unless morphological differences were evident. As population aggregation analyses are based on the a priori definitions of populations, this kind of informal treatment could introduce a major factor of uncertainty into species delimitation. The branching-based approach is independent of the recognition of population limits. In addition, the approach permits the inclusion of rare species represented by only a single individual, which is problematic in population based species concepts.

The length of time needed to attain the signatures used for species recognition differs between these methods. For example, CHA and PAA rely on the prediction that fixed nucleotide differences are unlikely between random samples of a single population, but expected to accumulate between isolated populations. Under a neutral model of mutation and genetic drift, for a mitochondrial marker the expected time to the first fixed difference, and hence ability of these methods to detect isolated species, is less than 0.5 Ne generations after populations were isolated (Hey, 1991). The WP method relies on the assumption that an isolated population will
eventually become monophyletic with respect to its ancestral or sister population. According to simulations of the probability to observe reciprocal monophyly, this is expected to occur at around $0.7 N_e$ generations; i.e., there is a 50% chance of a given locus being monophyletic at that time (table 1 in Hudson and Coyne, 2002). The branch-length analysis proposed here requires additionally that monophyletic groups are recognizable on longer
stem branches, expected to occur around $Ne$ generations after lineage separation (Hudson and Coyne, 2002). To have 95% of a given locus displaying a given signature, the time interval needed will be higher in each case, for example, over $2N_e$ generations, for changes in branching rate to arise. The final method we used, based on the inbreeding coefficient, detects significant between-population genetic subdivision from pairwise $F_{ST}$ estimates, which does not entail fixed differences or monophyly. Hence this method might detect more recent breaks in gene flow than the others, but risks that populations with partial gene flow are recognized as separate entities: it constitutes an extreme upper limit for species numbers in population-based analyses.

Whether these methods actually produce differing estimates of numbers of groups will depend on the relative separation of species versus within-species coalescence times. If many related species diverged more recently than the oldest coalescence times of the alleles they contain, the methods will underestimate species numbers, delimiting species complexes rather than individual species. This will affect methods relying on branch length more strongly than methods based on diagnostic fixed character states which have a greater chance to identify recently diverged or otherwise cryptic species because they are expected to be apparent sooner after divergence (Goldstein and Desalle, 2003). However, in the present case very similar estimates of species numbers and species limits were obtained with either method, demonstrating the robustness of the data to different quantitative approaches of species delimitation. Differences that did occur reflect the opposing effects of cryptic within-population divergence detected by the analysis of branching rates on the one hand, and the recognition of more recent divergence by population-based methods on the other. In addition, the analysis of branching rates conducted on monophyletic lineages will not recognize paraphyletic species, while also relying on the correct reconstruction of the tree. Arguably, the best final assignment would be to integrate branch-length approaches and the PAA method (as it is not based on topology and hence a “correct” tree) and to recognize 54 putative species.

Neither approach would recognize recently derived ecological variants. For example, *R. eburneola* is a flightless species from Lake Gilmore, which was not separated by any of the DNA methods as distinct from the co-occurring, but flighted and morphologically and ecologically distinct *R. nr. blackburni*. This form should clearly be added to the count of *Rivacindela* species. The example hence demonstrated that the mtDNA-based methods may be conservative compared to classical morphological techniques of species differentiation, contrary to widely held concerns (see Agapow et al., 2004). In addition, DNA delimited species were generally congruent when compared to the morphologically based species assignment in those cases where names had been assigned to the specimens in preliminary identifications (Table S1), which allays concerns that DNA-based taxonomies result in oversplitting of the existing variation.

Although *R. eburneola* was the only obvious case where major morphological differences did not coincide with the mtDNA defined groups, minor variation was seen in the informal morphological identification during field work that also disagreed with DNA (Table S1). This raises the possibility of incongruent gene and species histories, in particular in the light of reported high propensity for gene flow of mtDNA between otherwise separated groups (Funk and Omland, 2003; Hudson and Coyne, 2002; Seberg et al., 2003; Will and Rubinoff, 2004). Without further morphological analysis or sequencing of nuclear loci, the magnitude of this problem in *Rivacindela* is difficult to assess. However, the high fidelity of geographic distribution of DNA-based groups and ancient river basins (Fig. 6) provides clear evidence against mtDNA dispersal, in particular during the past $\sim 700$ Ky since most of these species were separated. Across all western Australian species we observe only two cases of well-sampled distinct clusters in the branch lengths analysis (individuals from site GY in clade 30, and individuals from site DE in clade 28; the DE individuals being a diagnosable group in the PAA) where one or more sequences were embedded in a “foreign” population; i.e., in a different river basin than all other sequences within the cluster. Hence, in *Rivacindela* the broad congruence of mtDNA distribution with biogeographical landscape features is evidence for their biological relevance.

The geographic pattern strongly indicates vicariant speciation as the cause of the range distributions, with only occasional dispersal between drainage systems, consistent with the paleoclimatic history of the region. Ephemeral salt lakes of interior Australia formed by fragmentation of paleo-drainage basins that existed prior to the aridification of the continent starting in the Miocene (van de Graaff et al., 1977). Proposed ancient basins remained well preserved because of tectonic stability and slow erosion and sedimentation in the area (van de Graaff et al., 1977). The date of the species-to-population transition (Fig. 4) coincides narrowly with the changes from lacustrine clays to evaporites and dune sediments estimated to have occurred between 400 and 700 Kya in a final shift to present-day aridity in regional climates (Pillans and Bourman, 2001). This would indicate that species formation in *Rivacindela* was a direct result of the fragmentation of their habitat near edges of the disappearing river systems.

**Conclusions**

The acceptance of DNA taxonomy in the literature has been slow (e.g., Lipscomb et al., 2003; Wheeler, 2004; Will and Rubinoff, 2004), partly because the biological status of entities established from sequence information alone is not easily verified, unlike the long tradition of taxonomic research based on morphological characters. However, in *Rivacindela* the species status of the mtDNA-based entities is well justified both evolutionarily and biologically: the levels of sequence divergence between species matched those in other insects (Brower, 1994), including taxonomically well studied North American
Cicindela (Vogler et al., 2005); geographic ranges are contiguous but clearly delimited and consistent with biogeography (Fig. 1); patterns of genetic diversity are consistent with paleoclimatic evidence for their vicariant origin (Fig. 6); and the tree exhibits the branching rate signature of the species-population transition (Fig. 5). Besides the question about species status, the mtDNA groups also provide a powerful system for taxonomic communication, identification, and convenient integration of newly generated information. For example, specimens found during our second (2003) expedition were readily identified in the existing database either as members of established species or representing unknown groups (Supplementary Table S1).

With an estimated $10^7$ species of insects on Earth (Erwin, 1982; Odegaard, 2000), there is no realistic hope for their in-depth taxonomic treatment using traditional methods. Greater sequencing throughput and algorithmic analysis of species membership, as applied here to a depauperate desert ecosystem, could be extended to more complex assemblages, including the spectacular diversity of insect species in tropical forests. The efficacy of such approaches will depend on whether similar branching rate shifts are found generally in insects and to what degree they separate closely related species. The prominent feature of the branch-length distribution of Rivacindela defined by the largely synchronized origin of species fitting a simplified model of rate shift may be a special case due to its unique paleoclimatic history. In addition, the extent of population sampling in a large clade may rarely be as complete as in the current study, and could also affect the recovery of the rate shift, suppressing or inflating rates in either portion of the tree and shifting their slope relative to each other. Finally, population subdivision within species would also blur the species-to-population transition of branching rates, as subdivision (metapopulation structure) tends to increase the age of the most recent common ancestor within species (Wakeley, 2000; Wakeley and Aliacar, 2001), although, fortuitously, if the common ancestor of diverged species was also subdivided, genetic divergence between sister species will tend to increase by a similar amount (Wakeley, 2000). Given the emerging evidence for strong clustering of sequence variation in most groups of insects and other invertebrates (Hajibabaei et al., 2006; Meyer and Paulay, 2005; Monaghan et al., 2005; Page et al., 2005; Smith et al., 2005) and their largely uniform coalescence time (Meyer and Paulay, 2005), for the great majority of lineages the signal of the clade-to-population transition is probably sufficiently robust to be detectable even under a range of confounding biological factors and sampling regimes. This supports the wide applicability of mtDNA-based species delimitation in poorly known groups generally.

As well as establishing a taxonomic framework, the high-throughput mtDNA profiles simultaneously provide information on the evolutionary and ecological history of a group. In Rivacindela, we established that macroecological patterns (endemism, beta-diversity, range sizes) and dynamics of speciation result largely from the history of climate change, whereby separated populations have persisted in situ for $10^5$ to $10^6$ years. This provides a context for comparative studies, e.g., with the subterranean diving beetles dependent on the same aquifers as Rivacindela (Leys et al., 2003), or with patterns of diversity in the 1000 species of Cicindela s.l. on other continents (Pearson and Vogler, 2001). While providing a framework for taxonomy widely accessible without specialist knowledge, the comprehensive mtDNA profiling of faunas across biogeographic regions (see also Monaghan et al., 2006) at the same time provides a formidable resource for evolutionary biology, population genetics, and studies of global biodiversity patterns.

ACKNOWLEDGMENTS

We thank David Brzoska, Fabio Cassola, and Jay Shetty for help with collection of specimens, and Saskia Bode and Fatos Kopliku for sequencing. Charles Godfray, Ryan Gregory, Paul Harvey, Andy Purvis, Sean Nee, and Richard Thomas provided helpful comments on an earlier version of the manuscript. We are grateful to Marshal Hedin, Rod Page, Mark Fishbein, and an anonymous referee for their detailed and insightful comments in the review process. A permit for export of specimens from Australia was held by WDS. This work was funded by the Natural Environment Research Council (UK) grant NER/A/S/2000/00489 and NER/A/S/2001/01133.

REFERENCES


First submitted 16 October 2005; reviews returned 31 December 2005; final acceptance 7 April 2006

Associate Editor: Marshal Hedin
Copyright of Systematic Biology is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.