PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 2987–2996
doi:10.1098/rstb.2008.0066
Published online 3 June 2008

Speciation and DNA barcodes: testing the effects of dispersal on the formation of discrete sequence clusters

Anna Papadopoulou[1,2], Johannes Bergsten[1,2], Tomochika Fujisawa[2], Michael T. Monaghan[1,2], Timothy G. Barraclough[2,3] and Alfried P. Vogler[1,2,*]

[1]*Department of Entomology, Natural History Museum, London SW7 5BD, UK*
[2]*Division of Biology, Imperial College London, Silwood Park Campus, Ascot SW7 2AZ, UK*
[3]*Jodrell Laboratory, Royal Botanic Gardens, Kew TW9 3DS, UK*

Large-scale sequencing of short mtDNA fragments for biodiversity inventories ('DNA barcoding') indicates that sequence variation in animal mtDNA is highly structured and partitioned into discrete genetic clusters that correspond broadly to species-level entities. Here we explore how the migration rate, an important demographic parameter that is directly related to population isolation, might affect variation in the strength of mtDNA clustering among taxa. Patterns of mtDNA variation were investigated in two groups of beetles that both contain lineages occupying habitats predicted to select for different dispersal abilities: predacious diving beetles (Dytiscidae) in the genus *Bidessus* from lotic and lentic habitats across Europe and darkling beetles (Tenebrionidae) in the genus *Eutagenia* from sand and other soil types in the Aegean Islands. The degree of genetic clustering was determined using the recently developed 'mixed Yule coalescent' (MYC) model that detects the transition from between-species to within-population branching patterns. Lineages from presumed stable habitats, and therefore displaying lower dispersal ability and migration rates, showed greater levels of mtDNA clustering and geographical subdivision than their close relatives inhabiting ephemeral habitats. Simulations of expected patterns of mtDNA variation under island models showed that MYC clusters are only detected when the migration rates are much lower than the value of $Nm = 1$ typically used to define the threshold for neutral genetic divergence. Therefore, discrete mtDNA clusters provide strong evidence for independently evolving populations or species, but their formation is suppressed even under very low levels of dispersal.

Keywords: DNA taxonomy; lotic-lentic; aquatic beetles; tenebrionid beetles; Aegean Islands; coalescence

1. INTRODUCTION

Sequence variation in mtDNA of most animal groups is discretely partitioned into clusters of closely related haplotypes that are widely separated from other such clusters (Hebert & Gregory 2005; Meyer & Paulay 2005; Blaxter *et al.* 2005; Pons *et al.* 2006; Vogler & Monaghan 2007). These mtDNA clusters often overlap with Linnaean species names and/or morphologically separable groups and are usually congruent with groups defined by nuclear markers, indicating that they broadly mirror the species category. This has led to the increasing use of mtDNA to assign individuals to clusters and, by proxy, species (Hebert *et al.* 2003). To date, much of the use of DNA barcodes relies on operational criteria to separate clusters (e.g. the '10× rule') and has thus been criticized, in part, because the evolutionary theory of speciation would not predict such uniform levels of divergence between species (Hudson & Coyne 2002; see also Mallet 2008). However, the presence of distinct genetic clusters remains a striking feature of mtDNA datasets that needs to be explained. In particular, from both operational and evolutionary perspectives, it is important to understand what processes explain variation in the strength of mtDNA clustering in different taxa.

It has long been recognized that population subdivision is critical for the origin of 'phylogroups' of closely related mtDNA haplotypes, i.e. mtDNA clusters (Avise 1989). Under various models of speciation, genetic drift in isolated populations descended from a common ancestor will eventually lead to the reciprocal monophyly of neutral markers (Neigel & Avise 1986; Carstens & Knowles 2007). In an allopatric setting, these monophyletic groups will be confined to different areas. The degree of distinctness of these groups is affected by the strength of the isolating barrier (i.e. migration rates between populations) and the time since a barrier has formed (Avise 1989; Avise & Ball 1990). For example, under an infinite island model populations are unlikely to develop significant genetic differentiation due to drift if $Nm > 1$, where N is the population size in an island and m is the migration rate (i.e. the probability

*Author and address for correspondence: Department of Entomology, Natural History Museum, London SW7 5BD, UK (a.vogler@nhm.ac.uk).

One contribution of 12 to a Theme Issue 'Speciation in plants and animals: pattern and process'.

of an individual from one population to breed in the other; Wright 1931; Slatkin 1985). While other processes might cause genetic divergence, for example, adaptation to distinct ecological niches, the above scheme makes a clear prediction: mtDNA variation should be more strongly structured in taxa with lower migration rates, all else being equal.

Here, we explore whether migration rates can explain broad differences in the degree of mtDNA clustering between clades. Two groups of beetles were chosen, both of which contain lineages occupying habitat types that differ in their stability and consequently presumed selective regimes for dispersal. The first group comprises aquatic diving beetles (Dytiscidae) that are either confined to flowing (lotic) or standing (lentic) water bodies. Lentic habitats are comparatively short lived, in particular the small ponds usually visited by aquatic beetles, which fill in due to sedimentation on time scales of decades. Running waters, although changing their course, are continuous over longer geological time spans (Bishop 1995). Persistence of populations in lentic habitats therefore requires dispersal by flight, rendering these beetles more dispersive than their lotic counterparts. This is supported by, on average, greater species ranges and a lower geographical species turnover in lentic species (Ribera & Vogler 2000; Ribera *et al.* 2003; Hof *et al.* 2006; Marten *et al.* 2006). The second example comprises flightless terrestrial darkling beetles (Tenebrionidae) from the Aegean Islands, in which closely related lineages occupy either sandy coastal areas or inland areas of soil. Sandy coastal areas are exposed to erosion from wind and waves, promoting passive dispersal or requiring high dispersal rate, whereas interior areas comprise more stable habitat of later successional stages. Both groups are potential examples of habitat-induced differences in dispersal ability, which could be reflected in their degree of geographical subdivision, genetic clustering and frequency of speciation.

In the absence of direct measures of dispersal or migration rates in these beetles, we use simulations of coalescence under population subdivision with varying migration rates to determine what levels of difference in migration rate could generate the observed differences in the degree of genetic clustering. The recently developed mixed Yule coalescent (MYC) model (Pons *et al.* 2006; Fontaneto *et al.* 2007) was employed to delimit genetic clusters and to compare the strength of clustering between beetle groups as well as in the simulated gene trees produced under different levels of migration.

2. MATERIAL AND METHODS

(a) *Taxon sampling, DNA sequencing and phylogenetic analysis*

Target groups from Dytiscidae and Tenebrionidae were chosen to represent lineages living in habitats of different stability. As representatives of lotic and lentic Dytiscidae, we examined three species of the genus *Bidessus*, including two lentic (*Bidessus goudotii* Laporte de Castelnau 1835 and *Bidessus unistriatus* Goeze 1777) and one lotic (*Bidessus minutissimus* Germar 1824) lineage (Ribera *et al.* 2003). A total of 33, 30 and 18 individuals, respectively, were sampled as part of a standardized transect across the western Palaearctic from 15 localities in Spain, France, England, Germany and Sweden (figure 1*a*). An individual of *Hydroglyphus geminus* Fabricius 1792 was used as an out-group. As representative of flightless Tenebrionidae occurring on different soil types, we selected the eastern Mediterranean species complex *Eutagenia smyrnensis sensu lato*, which was sampled from 74 localities on 18 islands of the central Aegean archipelago and 5 localities on the west coast of Turkey (figure 1*b*). Preliminary analyses revealed two clearly distinct mtDNA lineages with overlapping ranges but differing in their habitat preferences. One lineage occurs on coastal sand dunes and sandy beaches ('sand clade') and the other occurs on non-sandy soils ('soil clade'). Specimens collected include 46 sand-clade and 48 soil-clade individuals. *Stenosis syrensis* Koch 1936 from the same tribe Stenosini was used as an out-group.

Total genomic DNA was extracted from thorax or leg tissue using Wizard SV 96-well plates (Promega, UK). A fragment of approximately 800 bp from the 3′ end of the cytochrome oxidase I (*cox1*) gene was amplified using primers Jerry and Pat (Simon *et al.* 1994) and sequenced in both directions using a BigDye v. 2.1 terminator reaction kit. Sequences were analysed on an ABI 3730 automated sequencer and forward and reverse strands were assembled in Sequencher v. 4.6. Alignments included 708 characters (122 parsimony informative) for the *Bidessus* dataset and 829 characters (288 parsimony informative) for *Eutagenia*. Neither alignment was length variable. The *Bidessus* dataset consisted of 81 sequences including 25 unique haplotypes, while the *Eutagenia* dataset of 94 sequences and 57 unique haplotypes. All 177 sequences used in the analysis have been submitted to the EMBL Nucleotide Sequence Database under accession numbers AM947686–AM947862. Phylogenetic trees were obtained with Bayesian analyses in MrBayes v. 3.1.2 (Ronquist & Huelsenbeck 2003) for 5 million generations with two parallel searches using three heated and one cold Markov chain. A separate GTR+I+Γ model was applied to each of the three partitions corresponding to the codon positions. We used penalized likelihood as implemented in r8s v. 1.7 (Sanderson 2003) to obtain ultrametric trees. The optimal smoothing parameter was 1 for both empirical datasets, which was determined by cross-validation of values between 0.01 and 1000.

The number of haplotypes and segregating sites (S) were computed for each morphologically defined species of *Bidessus* or, in the case of *Eutagenia*, the two distinct mtDNA lineages occupying different habitat types. Mean nucleotide diversity π (Nei & Li 1979) was calculated for each lineage, geographical region and MYC cluster in empirical and simulated datasets (as described below). All calculations were done using Arlequin v. 3.1 (Excoffier *et al.* 2005) with the default settings. The batch mode was used for analysis of simulated datasets.

(b) *Simulated datasets*

We simulated mitochondrial genealogies and sequences for 16 populations, varying migration rate among populations in SIMCOAL (Excoffier *et al.* 2000). The simulations start from a given sample of individuals found in a number of demes, which are connected by a particular pattern of migration. The program first simulates the gene genealogy of the sample independently from the mutational process going backwards in time. Once the genealogy is obtained, mutations are assigned randomly to each branch of the tree, starting from the most recent common ancestor (MRCA) and assuming a uniform and constant Poisson process. Here we simulated an 800 bp sequence fragment with a mutation rate of 2% per million generations (comparable with the standard insect

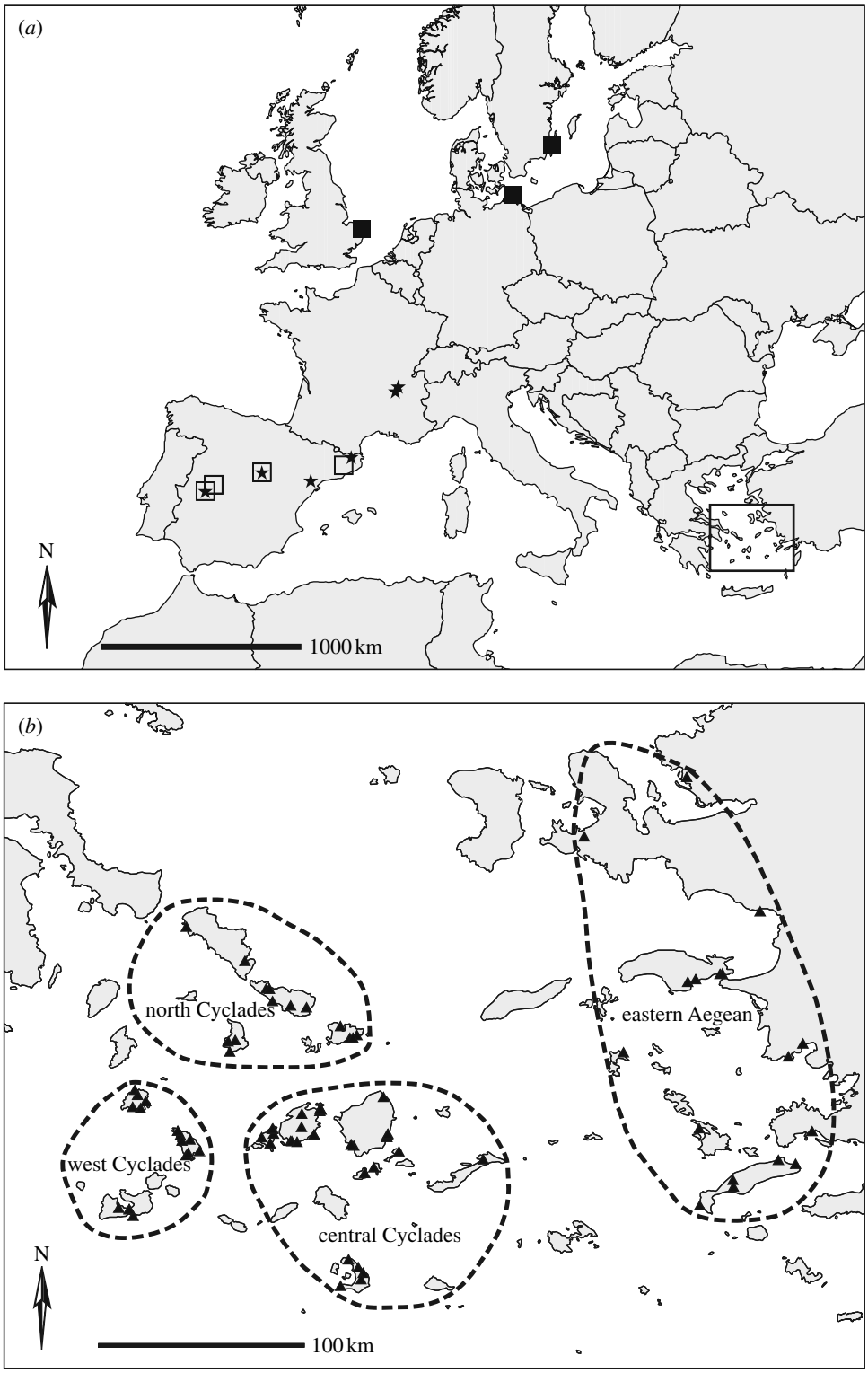

Figure 1. Sampling maps showing the collecting localities for (*a*) *Bidessus* diving beetles in western Europe and (*b*) *Eutagenia* darkling beetles in the Aegean archipelago in the eastern Mediterranean. Sampling sites for different species: filled squares, *B. unistriatus*; open squares, *B. goudotii*; stars, *B. minutissimus*; triangles, *E. smyrnensis s.l.*

molecular clock; Brower 1994) and a mutation model based on parameters estimated from the *Eutagenia* dataset with MrBayes: a transition/transversion ratio = 10 and a gamma-distributed among-site rate heterogeneity with $\alpha = 0.12$. Divergence among demes was simulated under the following scenario: a single panmictic population of a constant effective size ($N = 160\,000$) splits simultaneously into 16 daughter populations (= demes) each panmictic and of a constant size $N = 10\,000$, connected by symmetrical exchange of migrants (i.e. applying an island model scenario of migration). Each simulation continued for 1 million generations and 10 individuals were sampled from each of the 16 daughter populations. The migration rate (proportion of migrants exchanged with other demes per generation) is the only parameter that varied among the simulations and we examined eight different rates between $m = 10^{-8}$ and 5×10^{-5}, corresponding to absolute number of migrants per generation $Nm = 0.0001$–0.5. From the simulated gene trees, we calculated the percentage of daughter populations (demes) from which individuals were recovered as monophyletic, and the

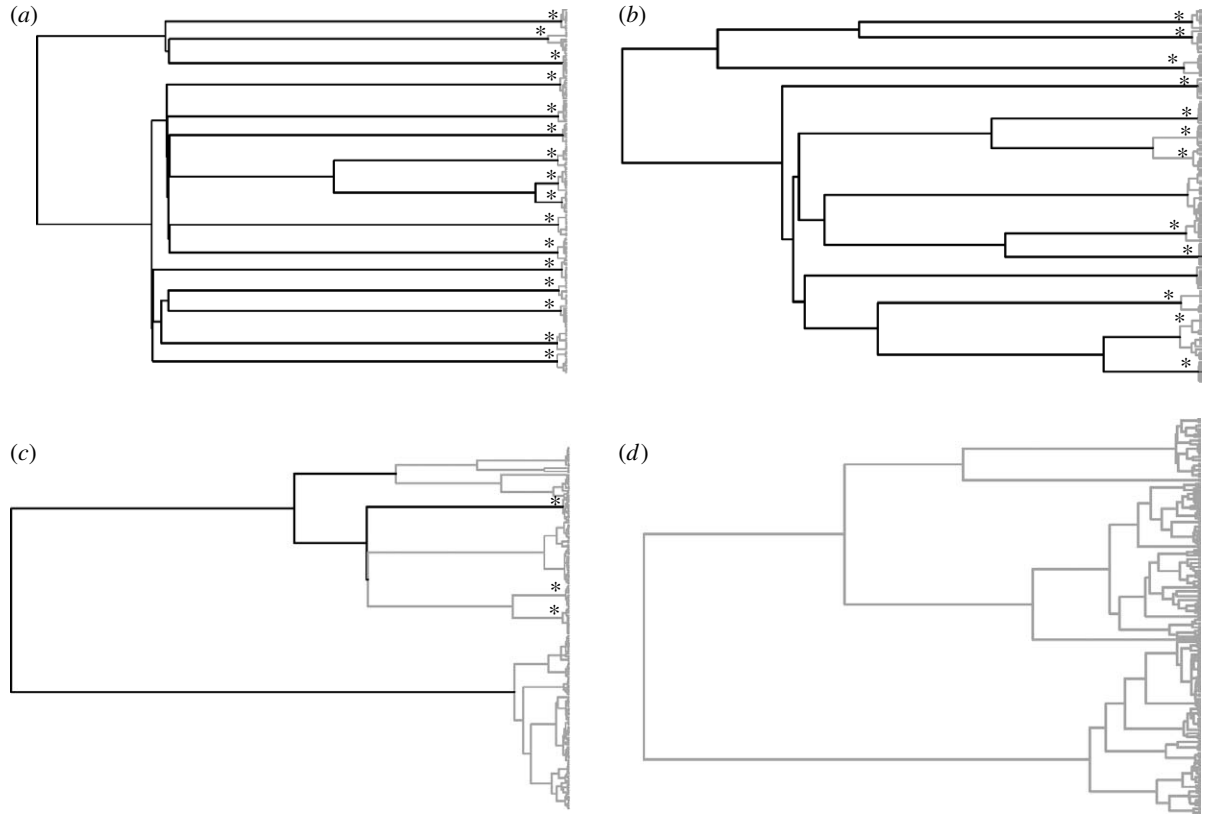

Figure 2. Examples of simulated genealogies at four out of the eight migration rates (m) examined, increasing from $Nm=$ (*a*) 0.0001, (*b*) 0.001, (*c*) 0.01 to (*d*) 0.1. Grey shading indicates branches allocated to the coalescent by the MYC model: (*a*) number of coalescent branching clusters (N_{MYC})=16, lambda ratio (λ_D/λ_C)=0.272; $\chi^2 p$=0.000; (*b*) N_{MYC}=13, λ_D/λ_C=0.474, $p<0.001$; (*c*) N_{MYC}=4, λ_D/λ_C=0.001, p=0.016 and (*d*) N_{MYC}=1, p=1.000. Nodes marked with asterisks correspond to demes that were recovered as monophyletic.

number of clusters obtained using a likelihood optimization of the MYC model as described below. Moreover, the nucleotide diversity π of each deme was estimated from the simulated sequence datasets. All simulations were carried out using 100 replicates, and the mean and median for the above parameters were calculated. Limited tests using 1000 replicates did not greatly alter these results.

(c) *Delimiting mtDNA clusters*

We used the MYC method of Pons *et al.* (2006) to delimit mtDNA clusters based on the transition from slow to faster apparent branching rates on the gene tree expected at the species boundary (Acinas *et al.* 2004). The method optimizes a threshold age, T, such that nodes before the threshold are considered to be diversification events (i.e. reflect cladogenesis generating the isolated species) and nodes subsequent to the threshold reflect coalescence occurring within each species. Waiting times between diversification events are modelled using a stochastic branching rate model, equivalent to a Yule model with branching rate λ_D (Yule 1924; Nee *et al.* 1992) but with an additional scaling parameter, p_D, which allows for smooth changes in per lineage branching rate over time, as might be expected under background extinction models or if species samples are incomplete (Barraclough & Nee 2001). Waiting times between coalescent events within species are modelled using a separate coalescent process for each species (Hudson 1991; Wakeley 2006), with branching rate, λ_C, but again modified by including a scaling parameter, p_C, that relaxes the strict assumption of neutral coalescence and constant population size by allowing smooth changes in branching rate over time, as might arise if population size has increased or declined through time or if there have been recent selective sweeps (Pons *et al.* 2006; Fontaneto *et al.* 2007). Therefore, branches crossing the threshold define k genetic clusters each obeying an independent coalescent process but with branching rate, λ_C, and scaling parameter, p_C, assumed to be constant across clusters. The likelihood of this model is calculated using eqn 6 from Pons *et al.* (2006) and compared with a null model that the entire sample derives from a single species (i.e. can be fitted by a single neutral coalescent model for all individuals). There are five parameters for the MYC model (T, λ_D, p_D, λ_C and p_C) and two for the null model (a single branching rate and scaling parameter). Full details are provided by Pons *et al.* (2006). A script implementing the method in R is available from T.G.B. A modified script was used to apply the method simultaneously to multiple simulated replicates and obtain a summary of the model parameters for each tree.

3. RESULTS

(a) *Simulated datasets*

Trees obtained from simulated datasets (figure 2) showed a general trend of decreased monophyly and increased nucleotide diversity within demes as migration rate increased. Changes in key parameters were confined mainly to a window between $Nm=5\times10^{-3}$ and 5×10^{-2}, with generally stable patterns above and below these values (figure 3). At the lowest migration rate ($Nm=1\times10^{-4}$), each of the 16 daughter populations formed deeply separated clades of closely similar haplotypes. With increasing migration rate, the separation of deep clades was delayed (i.e. occurred further back in the tree) while variation within these

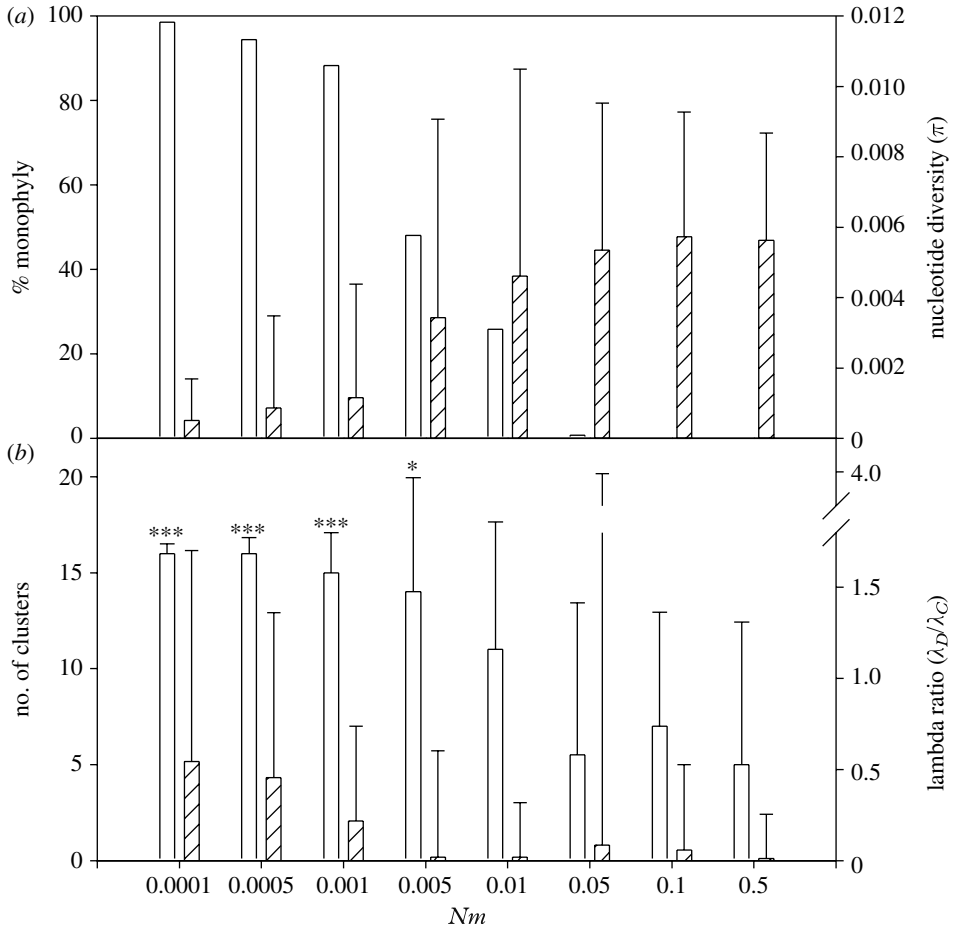

Figure 3. Summary statistics of simulated datasets presented as mean values (error bars = 1 s.d., $n = 100$, 1 million generations; see text). (*a*) Proportion of starting (local) demes recovered as monophyletic on the simulated trees and the average nucleotide diversity within demes (π). Open bars, monophyly; hatched bars, nucleotide diversity (π). (*b*) Number of clusters detected by the MYC model of lineage branching and lambda ratio, identified based on the point of transition from species-level (λ_D) and coalescence (λ_C) branching rates. Open bars, no. of clusters; hatched bars, lambda ratio. Statistical significance was evaluated with a likelihood ratio test comparing the MYC model with one of uniform coalescent branching. The lambda ratio (λ_D/λ_C) provides a convenient measure of the degree to which long and short branches can be distinguished, but at high Nm the MYC model has poor support; thus high λ_D/λ_C results only from artefactual delimitation of MYC groups. $^{*}p < 0.05$, $^{***}p < 0.001$.

clades increased slightly (figure 2). When the number of migrants per generation was greater than $Nm = 10^{-3}$, the clustering became generally weak. At these higher migration rates sequences from a single site (i.e. starting population) tend not to be monophyletic.

Summary statistics from 100 simulated genealogies at each migration level provided a quantitative assessment of the effects of migration on cluster formation. At the lowest migration rate, 98.5% of all populations (demes) were monophyletic after 1 million generations. Monophyly decreased substantially at $Nm = 5 \times 10^{-3}$ and was virtually zero under high ($Nm \geq 5 \times 10^{-2}$) migration rate (figure 3*a*). The migration rate also had a great effect on the genetic diversity of demes. In a pattern complementary to the level of deme monophyly, mean nucleotide diversity of demes was lowest ($\pi = 0.0050$) at the lowest migration rate and started to increase sharply at $Nm = 5 \times 10^{-3}$ to final values an order of magnitude higher at greater migration rates (figure 3*a*). At these rates, daughter populations were no longer separated and consequently the increased value for π corresponds to an increasingly greater number of individuals (up to $16 \times N$) maintaining higher nucleotide diversity.

The signature of distinct clusters in isolated populations was also assessed for the simulated genealogies with the MYC model. At the lowest migration rates ($Nm = 1 \times 10^{-4}$ and 5×10^{-4}), the model was strongly favoured and identified 16 clusters precisely corresponding to the individual demes in nearly all simulations (low standard deviation and a median of 16 clusters). Generally, there was a great overlap between the demes and the genealogically delimited groups. For example, under the lowest migration rates all but 1.5% of all demes were monophyletic, while the MYC model recovered 16 clusters on average with a low variance. Under slightly higher migration rate, the number of MYC groups decreased to 15 and 14, in step with the increase in non-monophyletic demes (figure 3). Increasing the migration rate further showed a great decline of clustering and the MYC model was no longer favoured over a simple coalescent model at rates $Nm > 10^{-2}$ (figure 3*b*). Poor model fit led to a decrease in the mean number of clusters (to low numbers of 5–7) and also resulted in increased variation in the number of clusters (figure 3*b*). Generally, the existence of MYC clusters was only supported when they coincide with the original demes.

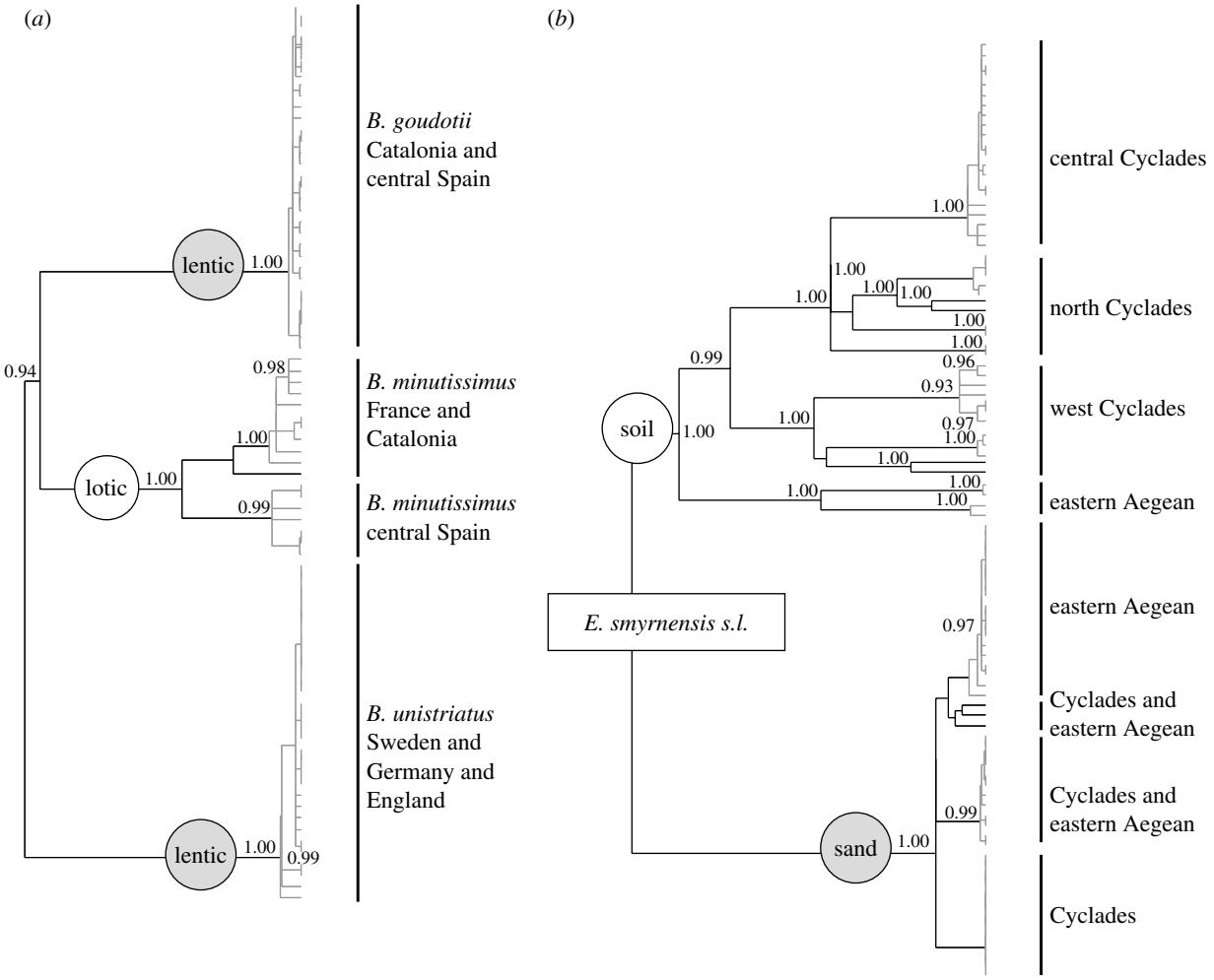

Figure 4. Bayesian trees for two empirical datasets. (*a*) *Bidessus* (Dytiscidae) including two lentic and one lotic species and (*b*) *E. smyrnensis s.l.* (Tenebrionidae) from sand and soil habitats. Grey shading indicates branches allocated to the coalescent by the MYC model. Numbers on the nodes correspond to posterior probabilities (only values above 0.90 are shown). These trees include all sequenced haplotypes, but in order to apply the MYC, only unique haplotypes were used.

The trees were then examined with regard to the degree of clustering, i.e. the degree to which shallow clusters are separated by long branches from other such clusters. This was approximated by comparing the rates of lineage branching in the phylogeny (λ_D) and coalescence (λ_C) portions of the tree, as their relative rates would provide a measure of how the branching is distributed along the root-to-tip axis. If the rates in λ_C are high in comparison with those in λ_D branching will be predominantly at the tips, indicating stronger clustering of haplotypes. At the lowest migration level, branching rates in the coalescence portion of the trees (λ_C) indeed exceeded the values for λ_D by about a factor of 2, indicating greater tip level branching and strong clustering (figure 2*a*,*b*). However, the ratio of the two rates decreased with higher migration rate as λ_C continued to increase, demonstrating that the clustering became less pronounced and the two modes of branching were less clearly recognizable (figure 3*b*). Once the migration rates exceeded the value under which clear clustering is observed, the λ_D/λ_C ratios were small and standard deviation highly variable among different migration rates, both of which are likely due to the recognition of artefactual MYC clusters.

(b) *Empirical datasets*

DNA sequence datasets obtained for the two model groups showed mtDNA diversity to follow roughly the predictions from simulated data (figure 4). In both groups, the lineages from the more stable habitat type (lotic for *Bidessus* and soil habitat for *Eutagenia*) exhibited increased rates of lineage branching deeper in the tree than the corresponding lineages from the less stable habitats. Lineages from these habitat types also had greater levels of haplotype diversity, greater geographical structure, greater number of segregating sites and higher nucleotide diversity (table 1). All of these conform to expectations of lower rates of dispersal in stable habitats. Specifically, the lotic (stable habitat) *B. minutissimus* from central Spain and Catalonia form distinct clades in contrast to the lentic (ephemeral habitat) *B. goudotii* from the same regions that are intermixed in a single cluster. In the MYC analysis, the lotic group was subdivided into three MYC clusters (one represented by a singleton only), with greater mean nucleotide diversity per MYC cluster than the single MYC group detected in each of the lentic species. The *Eutagenia* lineages showed similar trends: the 'soil' lineage was highly subdivided in accordance with the main geographical regions in

Table 1. Mitochondrial DNA (*cox1*) sequence variation in the two empirical datasets (*Bidessus* and *Eutagenia*) examined. (*K*, number of haplotypes; *n*, number of sequences; *S*, number of segregating sites; N_{MYC}, number of clusters delineated with MYC model; M_d, percentage of geographically diagnosable groups recovered as monophyletic.)

taxon	habitat	*K*	*n*	*S*	N_{MYC}	M_d
B. minutissimus	lotic (stable)	11	18	25	3	0.33
B. goudotii	lentic (unstable)	3	33	2	1	0
B. unistriatus	lentic (unstable)	4	30	3	1	0
E. smyrnensis s.l.	soil (stable)	40	48	154	12	1.00
E. smyrnensis s.l.	sand (unstable)	17	46	32	6	0

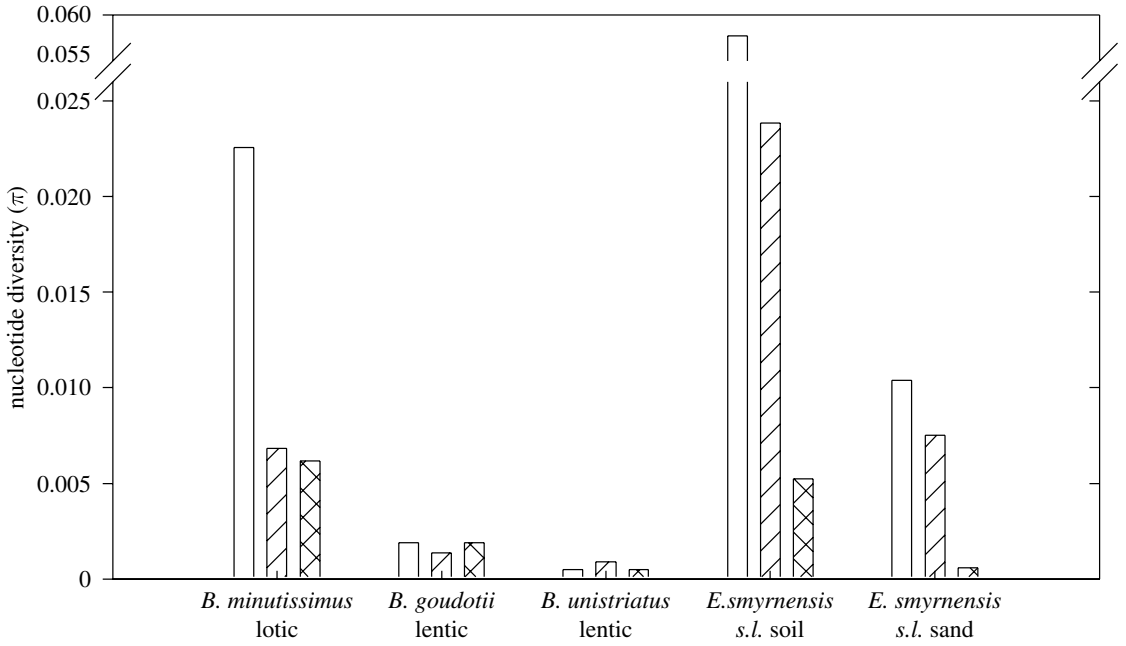

Figure 5. Mean nucleotide diversity (π) when sequences are grouped into morphologically defined species (open bars), geographically defined populations (hatched bars) and MYC clusters (cross-hatched bars).

the archipelago (central, north, west Cyclades and eastern Aegean) in contrast to *Eutagenia* from sand habitats where individuals from all four geographical regions were part of a single cluster. In addition, the stringent structure of geographical distribution and MYC clustering is not maintained in this group, as the three MYC clusters of the 'sand' clade are partly sympatric (figure 4).

The greater structure of the lotic versus lentic and the soil versus sand habitat species was also evident from the structure of nucleotide variation π at the geographical and MYC levels (figure 5). The genetic variation in the low-dispersal lineages (lotic and soil) dropped greatly when calculated separately for each of the geographical regions or MYC clusters. The effect was particularly clear for the MYC groups that showed greatly reduced nucleotide variation compared with the broader lineage from which they are drawn. When lineages were subdivided based on geographical regions, the genetic variation for each of the local groups was also greatly reduced compared with the entire clade in the low-dispersive *Bidessus* species. However, geographically defined groups in *Eutagenia* displayed a high π that was close to that for the entire clade, contrary to expectations for the low-dispersive lineages. Only when the MYC groups were considered did each of these show the expected low π values (figure 5). This discrepancy between geographically and genetically (MYC) defined groups suggests a problem with our *a priori* delimitation of demes, rather than contradicting the wider conclusions about the role of dispersal in the formation of geographically defined clusters. The geographical extent of local groups was delimited by individual islands, but these may not constitute the correct units for analysis if populations are geographically further subdivided *within* an island.

4. DISCUSSION

(a) *MYC groups and relevance to species formation*

Just as in traditional taxonomy, where perceived 'gaps' in morphological character variation are used to recognize species, mtDNA data generally exhibit strong clustering. Whether or not these clusters correspond to species-level entities in the Linnaean taxonomy is critical to the proposition of DNA barcoding for use in species identification. Our results confirmed the strong clustering for two lineages of beetles that have been broadly sampled throughout

their geographical range. The tenebrionids used here have been assigned to a single named species, *E. smyrnensis*, but appear clearly separated into many more independently evolving groups in the MYC analysis. Morphological studies now suggest that the two ecological types should be separated into different taxonomic units (B. Keskin 2007, unpublished data), but this would still leave a much higher level of MYC splitting than that suggested by the morphological taxonomy. The MYC clusters may indicate cryptic (morphologically not recognizable) species. Even if they would not explicitly be considered separate species, the MYC clusters of the soil clade were confined to groups of adjacent islands, and hence constitute historically separated entities and meaningful units from an evolutionary perspective (see Comes *et al.* 2008 for an example of plant taxa showing differentiation among ecologically similar habitats on different Aegean Islands).

The simulation results support the evolutionary significance of the MYC clusters. Under the conditions used here (10^6 generations and $N_e = 10^4$), the level of gene flow that permits the formation of MYC groups is much lower than the $Nm < 1$ at which neutral genetic differentiation is expected under an island model (Wright 1931; Slatkin 1985). As the migration rate increases to approximately $Nm = 5 \times 10^{-7}$, i.e. a point at which $Nm = 5 \times 10^{-3}$, the number of MYC clusters is reduced and recovery of monophyletic groups confined to a single daughter population (deme) rapidly goes to zero. Therefore, the MYC approach appears to be conservative, only detecting the products of population isolation when the levels of gene flow are much lower than those traditionally regarded as sufficient for neutral population divergence. The detection of MYC clusters does not seem to be consistent with only partial isolation among populations, contrary to previous suggestions (Fontaneto *et al.* 2007).

Equally, the formation of MYC groups requires sufficient time since divergence for reciprocal monophyly and long stem branches to evolve, a process that may be incomplete if lineage sorting is delayed (Maddison 1997; Hudson & Coyne 2002; Funk & Omland 2003; Hickerson *et al.* 2006; Knowles & Carstens 2007). Our simulations considered populations isolated for a sufficiently long time that reciprocal monophyly is highly likely to have evolved; at a constant effective population size of $N_e = 10\,000$, the probability of reciprocal monophyly between two isolated populations after 1 million generations is effectively 1 (table 1 in Hudson & Coyne 2002; Rosenberg 2003). Hence, all cases of non-monophyly of populations in our simulations, including the 1.5% across all groups at the lowest migration rate ($Nm = 1 \times 10^{-4}$; figure 3), are due to migration events. Shorter divergence times or larger effective population sizes would reduce the detectability of MYC groups even when migration rates are low enough to permit neutral divergence. For example, with either $N_e = 10\,000$ gene copies and a divergence time of 20 000 generations, or $N_e = 1$ million gene copies and a divergence time of 2 million generations, the probability of reciprocal monophyly between two isolated populations is approximately 50% (table 1 in Rosenberg 2003). The formation of discrete haplotype clusters is a general outcome of the separation of populations, provided that migration is suppressed below a certain level and enough time has passed.

(b) *Interpretation of habitat stability and dispersal*

Our interpretations assume that dispersal rates and hence gene flow among populations are lower in the more stable habitats than the less stable habitats. This idea has a long history in the ecological literature (Southwood 1977, 1988). Dispersal rate correlates negatively with habitat persistence in modelling (Travis & Dytham 1999) and empirical studies, for example, in brachypterous versus macropterous (reduced versus fully winged) insects occurring in natural or disturbed agricultural habitats, respectively (Denno *et al.* 1991). In aquatic habitats, a meta-analysis of lotic and lentic species of invertebrates has recently confirmed the lower average F_{ST} values in the latter (Marten *et al.* 2006), consistent with our hypothesis of shorter persistence of ponds, bogs and lakes due to sedimentation compared with streams and rivers (Ribera & Vogler 2000). Although less well established for the tenebrionids, it is plausible to assume similar effects also for the different soil types due to the greater erosion through wind and water of exposed sandy habitats. Lentic diving beetles and sand-inhabiting darkling beetles are therefore subject to selection for a higher migration rate.

The predictions from the differential habitat hypothesis were supported in both groups studied: in both aquatic dytiscids and terrestrial tenebrionids the stable-habitat lineages showed greater levels of population subdivision and geographical structure, had a greater number of separately evolving groups recognized by the MYC model, and lineage branching occurred more deeply in the tree than observed in their relatives inhabiting more unstable habitat types (table 1; figure 4). The high nucleotide diversity (π values) at the geographical level in both sand and soil *Eutagenia* (figure 5) does not contradict this conclusion but instead demonstrates the importance of establishing the extent of demes correctly. Here the geographical entities in our analysis were based on individual islands or groups of islands, but this may be simplistic owing to the complex geographical structure and geological history which may lead to further subdivision *within* islands. This again demonstrates the difficulties of defining demes in empirical data. However, this is a critical and possibly subjective step; *a priori* recognition of populations, which is a prerequisite for species delimitation under most species concepts (Sites & Marshall 2004), may affect the interpretation of patterns and underlying evolutionary processes. The MYC approach provides a test that is independent of these *a priori* defined groups.

(c) *Relevance to DNA barcoding*

The causes of strong clustering in sequence data from local assemblages and the presence of a 'barcoding gap' between intra- and interspecific pairwise sequence divergences (Meyer & Paulay 2005) still require an evolutionary explanation. Here we show that these

groups readily arise in simulated genealogies under a constant mutation process modelled after realistic data and sample sizes typically seen in mtDNA studies. Both the simulations and the side-by-side comparisons of ecological types support the strong effect of restricted gene flow. The different levels of dispersal (or broadly gene flow between demes) greatly affect the presence of discrete groups and their divergence, resulting in reduced interspecific variation and increased intraspecific variation, which may combine to cause the shrinking of the barcoding gap, i.e. the less pronounced separation of DNA clusters, which may not be detectable at all if variation in the markers sequenced is too low. In addition, the variance on all of these parameters is high (figure 2) affecting the degree of separation in any given case. A certain proportion of 'failure' of DNA-based approaches in taxonomy (Meyer & Paulay 2005; Hickerson *et al.* 2006; Meier *et al.* 2006; Elias *et al.* 2007) is therefore inevitable from dispersal, even at very low rates. More importantly, the *a priori* groups in barcoding studies using Linnaean taxonomy sometimes have cryptic subgroups as seen here in *Eutagenia*. In addition, the very low levels of migration at which discrete groups form and the high variance in the parameters mean that the MYC is quite conservative for the detection of evolutionarily separated groupings. A comparison of results from this technique with measures of p-distances in the barcoding literature may be an obvious next step. Applying conservative grouping procedures such as the MYC may be too strigent to detect recently diverged species but provides a starting point for broad surveys of diversity patterns.

We are grateful to the people who helped in collecting specimens of water beetles (I. Ribera, J. Geijer, G. Foster and L. Hendrich) and tenebrionids (I. Anastasiou, B. Keskin). We thank I. Ribera (Madrid) for discussions. Funding was provided by the NERC (grant NE/C510908/1), the BBSRC (grant BBS/B/04358) and a PhD studentship from the Greek State Scholarships Foundation to A.P.

REFERENCES

Acinas, S. G., Klepac-Ceraj, V., Hunt, D. E., Pharino, C., Ceraj, I., Distel, D. L. & Polz, M. F. 2004 Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**, 551–554. (doi:10.1038/nature02649)

Avise, J. C. 1989 Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* **43**, 1192–1208. (doi:10.2307/2409356)

Avise, J. C. & Ball, R. M. 1990 Principles of genealogical concordance in species concepts and biological taxonomy. *Oxf. Surv. Evol. Biol.* **7**, 45–68.

Barraclough, T. G. & Nee, S. 2001 Phylogenetics and speciation. *Trends Ecol. Evol.* **16**, 391–399. (doi:10.1016/S0169-5347(01)02161-9)

Bishop, P. 1995 Drainage rearrangement by river capture, beheading and diversion. *Progr. Phys. Geogr.* **19**, 449–473. (doi:10.1177/030913339501900402)

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R. & Abebe, E. 2005 Defining operational taxonomic units using DNA barcode data. *Phil. Trans. R. Soc. B* **360**, 1935–1943. (doi:10.1098/rstb.2005.1725)

Brower, A. V. Z. 1994 Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution. *Proc. Natl Acad. Sci. USA* **91**, 6491–6495. (doi:10.1073/pnas.91.14.6491)

Carstens, B. C. & Knowles, L. L. 2007 Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. *Syst. Biol.* **56**, 400–411. (doi:10.1080/10635150701405560)

Comes, H. P., Tribsch, A. & Bittkau, C. 2008 Plant speciation in continental island floras as exemplified by *Nigella* in the Aegean Archipelago. *Phil. Trans. R. Soc. B* **363**, 3083–3096. (doi:10.1098/rstb.2008.0063)

Denno, R. F., Roderick, G. K., Olmstead, K. L. & Dobel, H. G. 1991 Density-related migration in planthoppers (Homoptera, Delphacidae)—the role of habitat persistence. *Am. Nat.* **138**, 1513–1541. (doi:10.1086/285298)

Elias, M., Hill, R. I., Willmott, K. R., Dasmahapatra, K. K., Brower, A. V. Z., Malllet, J. & Jiggins, C. D. 2007 Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proc. R. Soc. B* **274**, 2881–2889. (doi:10.1098/rspb.2007.1035)

Excoffier, L., Novembre, J. & Schneider, S. 2000 Computer note. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J. Hered.* **91**, 506–509. (doi:10.1093/jhered/91.6.506)

Excoffier, L., Laval, G. & Schneider, S. 2005 Arlequin (v. 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* **1**, 47–50.

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. (doi:10.1371/journal.pbio.0050087)

Funk, D. J. & Omland, K. E. 2003 Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Syst.* **34**, 397–423. (doi:10.1146/annurev.ecolsys.34.011802.132421)

Hebert, P. D. N. & Gregory, T. R. 2005 The promise of DNA barcoding for taxonomy. *Syst. Biol.* **54**, 852–859. (doi:10.1080/10635150500354886)

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. (doi:10.1098/rspb.2002.2218)

Hickerson, M. J., Meyer, C. P. & Moritz, C. 2006 DNA barcoding will often fail to discover new animal species over broad parameter space. *Syst. Biol.* **55**, 729–739. (doi:10.1080/10635150600969898)

Hof, C., Brandle, M. & Brandl, R. 2006 Lentic odonates have larger and more northern ranges than lotic species. *J. Biogeogr.* **33**, 63–70. (doi:10.1111/j.1365-2699.2005.01358.x)

Hudson, R. R. 1991 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–44.

Hudson, R. R. & Coyne, J. A. 2002 Mathematical consequences of the genealogical species concept. *Evolution* **56**, 1557–1565. (doi:10.1554/0014-3820(2002)056[1557:MCOTGS]2.0.CO;2)

Knowles, L. L. & Carstens, B. C. 2007 Delimiting species without monophyletic gene trees. *Syst. Biol.* **56**, 887–895. (doi:10.1080/10635150701701091)

Maddison, W. P. 1997 Gene trees in species trees. *Syst. Biol.* **46**, 523–536. (doi:10.2307/2413694)

Mallet, J. 2008 Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Phil. Trans. R. Soc. B* **363**, 2971–2986. (doi:10.1098/rstb.2008.0081)

Marten, A., Brandle, M. & Brandl, R. 2006 Habitat type predicts genetic population differentiation in freshwater invertebrates. *Mol. Ecol.* **15**, 2643–2651. (doi:10.1111/j.1365-294X.2006.02940.x)

Meier, R., Shiyang, K., Vaidya, G. & Ng, P. K. L. 2006 DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Syst. Biol.* **55**, 715–728. (doi:10.1080/10635150600969864)

Meyer, C. P. & Paulay, G. 2005 DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol.* **3**, e422. (doi:10.1371/journal.pbio.0030422)

Nee, S., Mooers, A. O. & Harvey, P. H. 1992 Tempo and mode of evolution revealed from molecular phylogenies. *Proc. Natl Acad. Sci. USA* **89**, 8322–8326. (doi:10.1073/pnas.89.17.8322)

Nei, M. & Li, W. H. 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. USA* **76**, 5269–5273. (doi:10.1073/pnas.76.10.5269)

Neigel, J. E. & Avise, J. C. 1986 Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In *Evolutionary processes and theory* (eds E. Nevo & S. Karlin). New York, NY: Academic Press.

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. (doi:10.1080/10635150600852011)

Ribera, I. & Vogler, A. P. 2000 Habitat type as a determinant of species range sizes: the example of lotic–lentic differences in aquatic Coleoptera. *Biol. J. Linn. Soc.* **71**, 33–52. (doi:10.1006/bijl.1999.0412)

Ribera, I., Foster, G. N. & Vogler, A. P. 2003 Does habitat use explain large scale species richness patterns of aquatic beetles in Europe? *Ecography* **26**, 145–152. (doi:10.1034/j.1600-0587.2003.03271.x)

Ronquist, F. & Huelsenbeck, J. P. 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574. (doi:10.1093/bioinformatics/btg180)

Rosenberg, N. A. 2003 The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution* **57**, 1465–1477. (doi:10.1554/03-012)

Sanderson, M. J. 2003 r8s: inferring absolute rates of molecular evolution, divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302. (doi:10.1093/bioinformatics/19.2.301)

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. & Flook, P. 1994 Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Ann. Entomol. Soc. Am.* **87**, 651–701.

Sites, J. W. & Marshall, J. C. 2004 Operational criteria for delimiting species. *Annu. Rev. Ecol. Evol. Syst.* **35**, 199–227. (doi:10.1146/annurev.ecolsys.35.112202.130128)

Slatkin, M. 1985 Gene flow in natural populations. *Annu. Rev. Ecol. Syst.* **16**, 393–430. (doi:10.1146/annurev.ecolsys.16.1.393)

Southwood, T. R. E. 1977 Habitat, the templet for ecological strategies. *J. Anim. Ecol.* **46**, 337–365.

Southwood, T. R. E. 1988 Tactics, strategies and templets. *Oikos* **52**, 3–18. (doi:10.2307/3565974)

Travis, J. M. J. & Dytham, C. 1999 Habitat persistence, habitat availability and the evolution of dispersal. *Proc. R. Soc. B* **266**, 723–728. (doi:10.1098/rspb.1999.0696)

Vogler, A. P. & Monaghan, M. T. 2007 Recent advances in DNA taxonomy. *J. Zool. Syst. Evol. Res.* **45**, 1–10. (doi:10.1111/j.1439-0469.2006.00384.x)

Wakeley, J. 2006 *Coalescent theory: an introduction*. Colorado, CO: Roberts and Co.

Wright, S. 1931 Evolution in Mendelian populations. *Genetics* **6**, 111–123.

Yule, G. U. 1924 A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, FRS. *Phil. Trans. R. Soc. B* **213**, 21–87. (doi:10.1098/rstb.1925.0002)