PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Fujisawa T, Vogler AP, Barraclough TG. 2015 Ecology has contrasting effects on genetic variation within species versus rates of molecular evolution across species in water beetles. *Proc. R. Soc. B* **282**: 20142476.
http://dx.doi.org/10.1098/rspb.2014.2476

Received: 9 October 2014
Accepted: 19 November 2014

Subject Areas:
evolution, genetics

Keywords:
cytochrome oxidase 1 (CO1), genetic variation, rate of evolution, phylogenetic generalized least-squares, water beetles

Author for correspondence:
Tomochika Fujisawa
e-mail: t.fujisawa05@gmail.com

†Present address: Department of Zoology, Kyoto University, Sakyo, Kyoto, 606–8502, Japan.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2014.2476 or via http://rspb.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

Ecology has contrasting effects on genetic variation within species versus rates of molecular evolution across species in water beetles

Tomochika Fujisawa[1,2,†], Alfried P. Vogler[1,2] and Timothy G. Barraclough[1]

[1]Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, UK
[2]Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK

Comparative analysis is a potentially powerful approach to study the effects of ecological traits on genetic variation and rate of evolution across species. However, the lack of suitable datasets means that comparative studies of correlates of genetic traits across an entire clade have been rare. Here, we use a large DNA-barcode dataset (5062 sequences) of water beetles to test the effects of species ecology and geographical distribution on genetic variation within species and rates of molecular evolution across species. We investigated species traits predicted to influence their genetic characteristics, such as surrogate measures of species population size, latitudinal distribution and habitat types, taking phylogeny into account. Genetic variation of cytochrome oxidase I in water beetles was positively correlated with occupancy (numbers of sites of species presence) and negatively with latitude, whereas substitution rates across species depended mainly on habitat types, and running water specialists had the highest rate. These results are consistent with theoretical predictions from nearly-neutral theories of evolution, and suggest that the comparative analysis using large databases can give insights into correlates of genetic variation and molecular evolution.

1. Introduction

Genetic variation is a key parameter determining how populations evolve. Consequently, a central goal of population genetic studies is to understand the factors that control genetic variation in populations [1]. The neutral theory of molecular evolution predicts genetic variation based on mutation and drift alone. Under the assumption of selective neutrality of alleles, genetic variation of a population is proportional to the product of effective population size and mutation rate [2]. Theories have been developed to incorporate the other main factors that can affect neutral genetic variation, such as population structure [3] and demographic history [4]. These parameters are widely used to interpret population genetic data. However, there have been relatively few comparative tests of which factors correlate with genetic variation across multiple species [5]. A comparative study can reveal which parameters identified by theory are most important in explaining levels of genetic variation across species, but this requires detailed data across a large enough genetic sample for numerous closely related species.

Neutral theories of molecular variation have also been applied to the study of DNA substitution rates across species. Phylogenetic dating methods using molecular clocks have found considerable rate variation among taxa, which led to interest in the causes of that variation [6,7]. Many studies have investigated the importance of generation times, metabolic rates, environmental energy inputs and other potential correlates of neutral or nearly-neutral substitution rates [8–10]. However, population genetic parameters such as population size have been hard to incorporate because of the general lack of sequence data across a sufficiently broad sample of species. Estimates of effective population sizes from

relatively few well-known model organisms have been used for comparison with molecular variation [11], but these have limited potential to tease apart the effects of multiple factors. Matched pairs of island and mainland populations have demonstrated an elevated rate of non-synonymous substitution (interpreted as fixation of mildly deleterious mutations) in island populations with small effective population size [12]. The matched species pair design is elegant for investigating a single factor, but cannot easily allow the investigation of multiple factors simultaneously if each pair is distantly related from each other pair.

The increasing availability of within-species DNA sequence of putatively neutral markers for large clades, associated with the rise of DNA barcoding, offers the potential for comparative analysis of population genetic variation and substitution rates in concert. For example, if mutation rates alone vary systematically among different species, then genetic variation within species should correlate with substitution rates over longer evolutionary timescales. However, if population size variation is more important, then we predict different associations: genetic variation should correlate positively with measures of population size [13], whereas substitution rates should be independent of population size under strict neutrality or increase with decreasing population size if variants are nearly neutral (i.e. mildly deleterious [14]). Alternatively, we might find no evidence of predicted correlations if the parameters affecting present-day genetic variation have been inconsistent over the longer evolutionary timescales relevant for substitution rate differences.

Dispersal is another important factor to determine species' genetics by controlling connectivity between local populations and the chance of colonization into newly opened habitats. The effects of species' dispersal on genetic variation have been widely studied, and higher dispersal ability is commonly associated with less structured populations [15–17]. For example, Riginos *et al*. [17] reported that more dispersive 'pelagic spawner' fish have less structured populations than benthic guarders. The effect of dispersal on the rate of evolution is less widely studied. Faster evolution in less dispersive species is observed in darkling beetles (family Tenebrionidae) [16,18], but this has not been formally tested in a comparative analysis. The broad database of barcoding sequences enables us to test the relative importance of these multiple factors.

Here, we use a large sequence dataset of water beetle mitochondrial DNA (mtDNA) [19] to test the effect of species ecology on their genetic traits, taking other confounding factors into account. Mitochondrial genes have been most widely sampled for comparative studies of molecular evolution because of their favourable characteristics: supposed neutrality, simple maternal inheritance without recombination and easy amplification. Although the use of mtDNA for a population genetic marker has been questioned [20–22] because of introgression and the potential for selective sweeps that violate assumption of neutrality, we know of no nuclear DNA dataset of equivalent breadth and depth within a single clade.

The dataset includes not only genetic information of species, but also geographical distribution and habitat descriptions of samples, which can be used to measure other factors influencing their genetic traits. For example, as a surrogate estimate of the population size of each species, which is a fundamental parameter affecting genetic variation but hard to measure in natural populations, we use species occupancy and distribution estimated from geographical collection records because species occupancy typically correlates well with abundance [23–25]. Latitudinal distribution of species, which is reported to affect the rate of evolution, is obtained from sampling records. The genetic information also allows DNA-based species delimitation and circumvents the lack of reliable taxonomic identification of evolutionary units, which often hinders a large-scale sequence survey. Moreover, most of the sampled species in the dataset belong to the same family of water beetles (87% of samples belong to Dytiscidae) and it is possible to reconstruct their species phylogeny without major gaps in species sampling for the study region. Phylogeny can therefore be taken into account when judging the correlations among traits [26].

Habitat type is an ecological trait commonly used as a surrogate of dispersal ability in aquatic insects [27]. Water beetles provide a useful system for investigating the effects of habitat types on genetic traits. Water beetles inhabit two distinctive types of water bodies, standing and running water, which are expected to have contrasting effects on species traits [28]. Lentic (standing water) habitat is more ephemeral in evolutionary time scale than lotic (running water) habitat, hence lentic species are expected to have greater ability to colonize new habitat. This difference in dispersal propensity associated with habitat type has various effects, ranging from species range size to degree of gene flow and evolutionary species turnover [28]. Lentic species have greater colonization ability and larger ranges as a result of their higher flight ability relative to lotic species [29,30]. For instance, it was reported that lentic species have a range size that is, on average, 3.4 times larger than lotic species in *Hydroporus* diving beetles [29]. We predict that lentic species should therefore hold greater genetic variation than less dispersive lotic species. Small population sizes of lotic species might lead to a relative excess of non-neutral changes, however, because of the lower efficacy of purifying selection in smaller populations [31]. Alternatively, the greater variation in ephemeral habitats may be counteracted by frequent extinctions and loss of abundance on a local level. We test this hypothesis using the water beetle data, taking other potential factors into account.

2. Material and methods

(a) Species delimitation and phylogenetic inference

The water beetle dataset consists of 5062 CO1 sequences of a maximum of 700 bp (GenBank accession numbers JN840019–JN845080; see [19] for detailed descriptions of sample collection and sequencing). Quality filtering by checking in-frame stop codons and removal of identical haplotypes resulted in 2106 unique haplotypes. These sequences were used for DNA-based species delimitation. The gene tree of CO1 was reconstructed using RAxML v. 7.0.3 [32] and made ultrametric with Pathd8 [33]. Putative species groups were delineated with the generalized mixed Yule-coalescent method (GMYC [34]) implemented in the 'splits' package [35]. The congruence between field identification of diving beetle taxonomic species and GMYC-delimited species was measured by counting the exact matches of memberships between the two groupings. For convenience, we refer to GMYC-delimited groups as 'species' throughout.

Following species delimitation, a species tree of GMYC units was reconstructed to obtain a working phylogeny for comparative analyses. One CO1 sequence was randomly taken from samples of each species, and additional loci were retrieved from GenBank to improve the reliability of phylogenetic reconstruction. Sequences

of mitochondrial 16S and nuclear ribosomal RNA 18S, 28S and protein-coding *wingless* genes were searched using the taxonomic names of the CO1 samples. The downloaded sequences were separately aligned using MUSCLE v. 3.8 [36], and the resulting alignments were concatenated into a single data matrix. When two or more GMYC-delimited species had the same taxonomic names, only one GMYC species was randomly chosen, and the downloaded sequences were attached to it. Incorrect concatenations owing to misidentification were checked by comparing a resulting phylogeny and taxonomic literature (listed in the electronic supplementary material, table S1). The tree search described in the following paragraph was run twice, and a final species tree was built from the corrected matrix.

Markov chain Monte Carlo (MCMC) sampling of tree topology and substitution parameters were conducted with MRBAYES v. 3.2.2 [37]. Four independent MCMC chains of eight million generations were run with the GTR + I + G model, selected by jModelTest2 [38], with three partitions, and the first four million generations were discarded as burn-in. The effective sample sizes (ESSs) of the parameters were examined with TRACER v. 1.5 [39]. The ESSs of all parameters reached more than 100 within the eight million generations allowed for the searches. A maximum clade credibility (MCC) tree was taken from the MCMC samples by TREEANNOTATOR v. 1.61. Because no time calibration points were available for this diving beetle group, relative divergence time was estimated using BEAST [40] with the uncorrelated lognormal relaxed clock. The GTR + I + G model with three partitions was again used as the model of sequence evolution, and the mean substitution rate for the CO1 partition was set to 1.0. An MCMC of 50 million generations was run, and convergence of parameters was checked by TRACER.

To assess incongruence between the genealogy of mitochondrial and nuclear loci, a test of tree topology was conducted. Three maximum-likelihood topologies were separately estimated for mtDNA, 18S and whole concatenated alignments using RAxML v. 7.0.3 with GTR + I + G model. Then, likelihood values of the three topologies were calculated with re-optimization of parameters for each alignment. The significance between likelihoods of the three topologies was tested with the Shimodaira–Hasegawa test (SH test) with 10 000 bootstrap replicates [41]. Significant reduction of likelihood between topologies indicates incongruence of the loci. Taxa used for the tree search were reduced, so that all taxa have complete alignments without missing characters.

(b) Data extraction

(i) Habitat types, occupancy and geographical distribution

Species' habitat types were assigned by mining sampling records. First, the habitat types of samples were determined according to descriptions of habitats in Ribera & Vogler [42]: for example, lotic habitat for 'river' or 'creek' and lentic habitat for 'pond' or 'lake'. Then, frequencies of samples from lotic and lentic habitats were counted, and species with more than 90% of either lotic or lentic samples were marked as 'running' or 'standing' species, respectively. Species composed of mixed samples were assigned to the 'both' category.

The sequences were sampled at about 190 unique sites across Europe, and 4999 of 5062 sequences have GPS records of sampled localities. The sampling sites were clustered into 23 broad regions of about 50 km in diameter according to their geographical positions in Baselga *et al.* [19]. Geographical positions of regions and proportion of lotic or lentic samples collected from the regions are summarized in electronic supplementary material, figure S1 and table S2. Species occupancy was measured by counting the number of the geographical regions where the specimens from this study were sampled. The latitudinal ranges of species, which we call range size, were obtained by subtracting the minimum from the maximum latitude of species sampling records. The median and minimum latitudes of species samples were used as measures of their latitudinal distributions, which has previously been used as a variable in studies of the effects of environmental energy on molecular rates [8]. Geographical locations of samples lacking GPS information were approximated with the mean values of the regions within which the samples were collected. The number of samples was recorded for each species to incorporate the effect of sampling, because species with large sample size may have larger genetic variation than rarely sampled species. Habitat type and geographical records of samples are summarized in the electronic supplementary material, data A (available at Dryad: doi:10.5061/dryad.926pq).

(ii) Genetic variation and rate of evolution of mtDNA

Genetic variation measured as nucleotide diversity (π) of CO1 for each species was estimated from intraspecific genetic distances. Identical haplotypes that were removed for the construction of GMYC groups were assigned back to each group. Pairwise genetic distances under the Kimura's two-parameter model were then calculated for the CO1 sequences within each species represented by three or more sequences, and their mean values used as a measure of nucleotide diversity within the species. Genetic distances were calculated for the whole CO1 alignment, including all codon positions (π_{overall}) and third codon positions (π_{third}). The mean genetic distances were arcsine square-root transformed because their distribution was positively skewed. Species with fewer than three sequences were excluded, and the remaining 191 species were used for downstream analyses.

The branch lengths (d) of a CO1 gene tree were used as a measure of rate of molecular evolution. Branch lengths of the gene tree were recalculated using RAxML v. 7.0.3 with the GMYC species tree as a binary constraint. The branch lengths on the gene tree were summed up through the path from tip to root for each sample, and their median for each species was taken as a representative measure of substitution rate of the group. The branch lengths were measured on the trees from the entire alignment (d_{overall}) and the third codon positions (d_{third}). To account for the effect of node density on branch length estimation (node density effect, [43]), the number of nodes on the path from tip to root was recorded for each species and included in regression analyses described below.

(c) Phylogenetic regression and model averaging

To test the predicted correlation of the habitat types with species ecology and distribution, univariate regressions with each parameter as a response variable in turn and habitat types as an explanatory variable were conducted first. Then, the multivariate regression models with genetic measures (genetic variation and substitution rate) as response variables and all species traits including habitat types as explanatory variables were constructed to analyse the collective effects of the variables on the genetic properties. Species measures are summarized in the electronic supplementary material, data B (Dryad: doi:10.5061/dryad.926pq).

Phylogenetic generalized least-squares (PGLS) described in Freckleton *et al.* [26] were used to incorporate the effect of phylogenetic dependence on the correlations of the variables. Pagel's λ [44] was estimated for all explanatory and response variables, and the degree of phylogenetic dependency was examined before the correlation between each variable and habitat was tested. The *pgls* function in the 'caper' package [45] was used for the PGLS analysis. For the multivariate phylogenetic regressions including all variables, model averaging following Burnham & Anderson [46] was used to assess the effects of multiple variables. The 95% confidence sets were constructed from all possible models without interaction terms, and the sum of

Akaike weights, weighted average of estimates and standard errors were obtained for each parameter from the models within the confidence set. The explanatory variables in the maximal model included terms for habitat type, species occupancy, latitude, sample size, the number of nodes and genetic variation or rate of evolution. All statistical analyses were conducted using R [47] with the aid of packages 'ape' [48] and 'phangorn' [49] for phylogenetic analyses and 'pegas' [50] for population genetic analyses.

3. Results

(a) Delimitation and species phylogeny

The GMYC analysis delimited 274 putative species groups, of which 45% matched exactly with taxonomic identifications, i.e. all members of a taxonomic species were included in a single GMYC group that did not include any other taxonomic species. Taxonomic species were over-split in 28% of cases and lumped with other taxonomic species in 32% of cases. The numbers of sequences retrieved from GenBank that matched with taxonomic names of the studied species were 100 and 51 for 16S and 18S, respectively (accession numbers of downloaded sequences are in the electronic supplementary material, data C, available at Dryad: doi:10.5061/dryad.926pq). These sequences were concatenated into a single alignment with 1854 characters, in which 650 sites were parsimony informative (344/700, 233/546 and 73/608 for CO1, 16S and 18S, respectively) and 47% of nucleotides were missing. Other nuclear markers, 28S and *wingless*, had only two and 11 matched sequences, respectively, and were excluded from the phylogenetic analysis.

The Bayesian tree search resulted in an MCC tree in which 113 of 272 internal nodes had more than 95% of support (electronic supplementary material, figure S1; a tree file is available at Dryad: doi:10.5061/dryad.926pq). Basal relationships among water beetle genera were less resolved than intrageneric relationships (median posterior probability 0.77 and 0.88, respectively). Three of the main families of water beetles (Noteridae, Haliplidae and Gyrinidae) formed monophyletic clades with more than 95% support, and the largest family, Dytiscidae, was monophyletic with 66% of posterior probability. The SH test with the reduced data showed significant differences on the likelihoods between the mitochondrial and nuclear species trees (electronic supplementary material, table S3), indicating possible incongruence between mitochondrial and nuclear genealogies. Because the ML tree of mtDNA and the concatenated data were not significantly different, however, the effect of incongruence on the concatenation is small, and the concatenated tree was used for further analyses.

(b) Species ecology, distribution and genetic variables

Mean pairwise genetic distances and median branch lengths on the gene tree were obtained for 191 of the 274 GMYC species that contained more than three samples (sample sizes per GMYC species ranged from 3 to 356, median 13; nine species had three samples only, six species had >100 samples). The mean genetic distances ranged from 0.0 to 0.019 (median 0.004) for the entire alignment ($\pi_{overall}$), and from 0.0 to 0.055 (median 0.0085) for third codon positions (π_{third}). Branch lengths ranged from 0.31 to 0.73 (median 0.45) and from 0.62 to 1.90 (median 1.01), respectively ($d_{overall}$ and d_{third}). The number of nodes from tip to root ranged between 4 and 20 (median 11). Among the 191 GMYC species, there were 38 species categorized as 'running' water species and 86 as 'standing'. The remaining 63 groups were found in both lotic and lentic habitats and marked as 'both', and four groups did not have any available habitat descriptions. The most widespread species, mainly consisting of samples identified as *Agabus bipustulatus*, was sampled from 20 of 23 geographical regions, whereas 38 groups were found in only one geographical region.

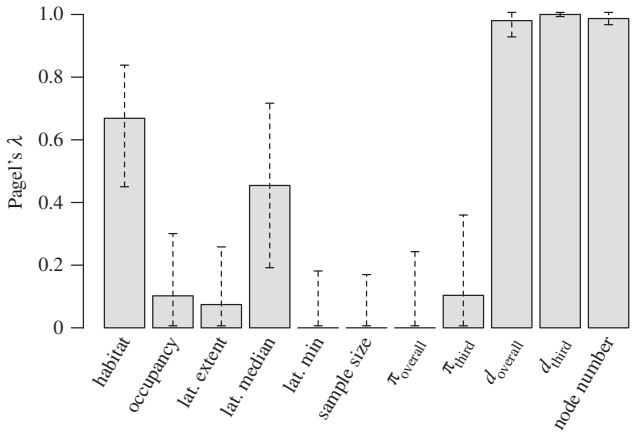

Figure 1. Phylogenetic dependency of traits measured as maximum-likelihood estimates of Pagel's λ. The 95% CIs of the estimates are shown by error bars.

The estimates of Pagel's λ for habitat type and median latitude were significantly different from both 0 and 1 ($\lambda = 0.67$ and 0.45, respectively), indicating partial phylogenetic dependence of these variables (figure 1). Estimated λ of branch length and node numbers was 1.0 and 0.99, respectively, as they were measured on paths from root to tip along the phylogeny. There was no phylogenetic dependence of the other variables, including occupancy, latitudinal range size, minimum latitude, sample size and genetic variations.

(c) Phylogenetic regressions

(i) Univariate phylogenetic regression

There were significant differences in species latitudinal extent, and median and minimum latitude between lotic and lentic groups (table 1). Lotic species had significantly smaller range sizes than the other groups, including the 'both'. Lentic species were distributed further north than the other groups regarding their southern limits (min. latitude) and distributional centres (median latitude). The median latitude of lotic species was significantly lower than in the other two groups. Neither genetic variation nor branch length had significant correlation with habitat types, except that median branch length of the tree from the full dataset ($d_{overall}$) showed a significant difference between lotic and other groups. The sample sizes were not correlated with habitat type, which indicates that there was no sampling bias from different habitat types.

(ii) Multivariate phylogenetic regression with model averaging

When genetic variation was a response variable, species occupancy had the largest effect with sum of Akaike weight 1.0 (table 2), which means that it was included in all models within the 95% confidence set. Of 63 candidate models, 23 were included within the confidence set. Model-averaged Pagel's λ was 0, indicating no phylogenetic dependency on genetic variation. Among the explanatory variables, the model-averaged estimate of species occupancy was significantly greater than 0, indicating positive effects of species

Table 1. Univariate PGLS regression analysis of effect of habitat types on species traits (only estimates of lotic and lentic species are shown).

		lotic	lentic	*p*-value	R^2	Pagel's λ
occupancy (log)		1.00	1.07	0.65	0.045	0.11
latitude	extent	5.97	10.24	0.01*	0.060	0.01
	median	42.71	52.36	0.00***	0.195	0.23
	minimum	40.38	47.33	0.00***	0.169	0.00
sample size (log)		2.55	2.65	0.61	0.008	0.00
$\pi_{overall}$		0.059	0.066	0.17	0.006	0.00
π_{third}		0.089	0.101	0.25	0.001	0.06
$d_{overall}$		0.46	0.43	0.01*	0.026	0.99
d_{third}		1.15	1.10	0.17	0.00	1.00

Table 2. Relative importance and model-averaged estimators of species traits from PGLS models of genetic variation ($\pi_{overall}$ and π_{third}). The parameter estimates that are significantly different from zero with 0.05 level are indicated by bold numbers, and the highest relative importance is shown by bold italic. Σw_i: sum of Akaike weights of variables. $\beta \pm$ s.e.: model-averaged estimates and standard errors.

		overall			third		
		Σw_i	$\beta \pm$ s.e.		Σw_i	$\beta \pm$ s.e.	
occupancy		***1.00***	**0.013**	**0.0054**	***1.00***	**0.023**	**0.0087**
med. latitude		**0.93**	**0.0007**	**0.0003**	**0.99**	**0.0014**	**0.0005**
med. *d*		0.26	0.0028	0.031	0.60	−0.023	0.014
habitat	lotic	0.50	0.0098	0.0065	0.55	0.019	0.012
	lentic	0.50	0.0081	0.0050	0.55	0.013	0.0081
sample size		0.58	−0.0059	0.0037	0.55	−0.0093	0.0061
node number		0.30	0.0003	0.0006	0.27	−0.0001	0.0009

Table 3. Relative importance and model-averaged estimators of species traits from PGLS models of branch length ($d_{overall}$ and d_{third}). The parameter estimates that are significantly different from zero with 0.05 level are indicated by bold numbers, and the highest relative importance is shown by bold italic.

		overall			third		
		Σw_i	$\beta \pm$ s.e.		Σw_i	$\beta \pm$ s.e.	
occupancy		0.38	−0.0061	0.0067	0.24	−0.0009	0.012
med. latitude		0.58	−0.0007	0.0005	0.27	−0.0007	0.0014
mean π		0.28	0.055	0.094	0.29	−0.084	0.17
habitat	lotic	**0.84**	**0.025**	**0.010**	0.24	0.045	0.032
	lentic	0.84	−0.0011	0.0062	0.24	0.0013	0.018
sample size		0.45	0.0048	0.0043	0.25	0.0014	0.0080
node number		***1.00***	**0.013**	**0.0018**	***1.00***	**0.025**	**0.0059**

occupancy on genetic variation. The second largest effect was median latitude, with Akaike weight 0.93. The alternate use of minimum or median of latitude and occupancy or latitudinal range did not affect the trend of model averaging (electronic supplementary material, table S4). The measures of latitudinal distribution were the second largest factors in both cases, and both occupancy and latitudinal range had significant positive effects on the genetic variation. Sample size and habitat had the third and fourth largest effects, but their estimates were not significantly different from zero. The relative importance of the explanatory variables using the third codon position resulted in patterns similar to the full dataset. Species occupancy and minimum latitude exhibited the highest, significant correlations and the relative importance of latitude was 0.99, which is higher than in the analysis with the total dataset (table 2). Model-averaged Pagel's λ was 0.0004.

For the model with branch lengths as a response, the number of nodes had the largest weight, 1.0, which suggests a strong correlation between branch lengths and the number of nodes from tip to root (table 3). Of 63 candidate models, 21

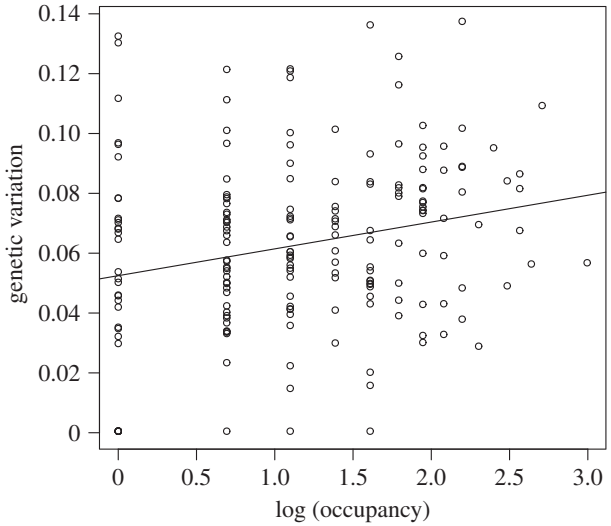

Figure 2. The effect of occupancy on genetic variation within species. The solid line is a regression line estimated by PGLS ($y = 0.009 \times x + 0.052$, $p = 0.0013$, $R^2 = 0.05$, Pagel's $\lambda = 0.0$).

were included in the 95% confidence set. Model-averaged Pagel's λ was 0.99. Apart from the effect of node density, habitat type was the second largest weight, 0.84, and the rate of evolution for lotic species was significantly higher than lentic and 'both'. Latitude was the third largest weight, but its effect was not significant. With the third codon positions, no variables except for the node number had significant effect on the branch lengths. The relative importance values for other variables were all smaller than 0.3, showing low explanatory power of the candidate models. Habitat type dropped the strong effects observed in the entire alignment.

4. Discussion

Habitat type had significant effects on both species range size and latitudinal distribution. Lotic species had significantly smaller range size (measured by latitudinal extent) and more southern distribution than lentic species. These patterns were consistent with well-supported effects of habitat type on dispersal propensity and distribution [42,51]. Lentic habitats are expected to be more ephemeral than lotic habitats, hence populations of lentic species only persist if they disperse more frequently to reach new suitable habitats, which predicts broader ranges and faster colonization of northern latitudes after the last ice age.

Occupancy and latitudinal range, in turn, displayed the strongest positive effect on genetic variation within species (table 2 and figure 2). Species with larger ranges and higher occupancy, which are associated with lentic habitats, contained more genetic variation than species with smaller ranges. There are well-established reports that the species occupancy and range size are tightly correlated with its abundance [23–25]. Therefore, the effect of occupancy on the genetic variation of CO1 is interpreted under the theory of neutral molecular evolution that predicts neutral genetic variation to be proportional to effective population size. The association of population size to mitochondrial genetic variation has been questioned [5] based on a comparison of levels of genetic variation in vertebrates and invertebrates (assumed to have low and high effective population size, respectively). Rate heterogeneity or frequent selective sweeps are proposed as possible alternative determinants of genetic variation of mtDNA [52,53]. Frequent fixation of adaptive mutation can reduce genetic variation and decouple it from population size [54]. However, the strong effect of abundance together with the weak effect of substitution rate (the product of mutation rate and fixation rate) on π observed in this study supported the traditional view of positive association between population size and genetic variation. An alternative explanation might be that species with larger ranges are older and have had time to accumulate more genetic variation, if speciation tended to result in new species with small ranges and low variation. However, there was no significant correlation between genetic variation and age of species measured on species tree (electronic supplementary material, figure S2).

The second predictor of genetic variation was latitude, yet contrary to the expected pattern, if high environmental energy promoted higher mutation rates [8], we find that minimal (low) latitude was correlated with lower π. It might also be expected for high latitude populations to show low genetic variation owing to recent colonization and expansion after the ice ages [55], but we found no such relationship with latitude either. One possible explanation is that the residual variation explained by latitude could be a surrogate of additional range size extending outside the sample area. It has been reported that high latitude species of water beetles have far wider range size than southern species, often encompassing the Holarctic [29], and the genetic variation in the northern species could be higher than expected by observed range size alone. Another possibility is that species at higher latitudes might be starting to diversify between local populations, yet too recently to have diverged into reciprocally monophyletic genetic clusters detectable by GMYC, whereas southern populations have diversified, and therefore are categorized as different species with small ranges and consequently small π. Weir & Schluter [56] found evidence for faster speciation rates at higher latitudes in birds, perhaps explained by new opportunities for speciation in empty areas recently recolonized since glacial retreat.

Next, we considered the factors correlated with substitution rates across species. Habitat type had the strongest effect on the rate of evolution, but unlike intraspecific genetic variation that was highest in lentic species with larger population size, the highest substitution rate was found in lotic species. The higher rate of evolution of lotic species could be explained by more frequent fixation of non-neutral mutations in small populations. If lotic species have smaller and more structured populations because of their low dispersal ability, fixation of deleterious alleles within local demes can increase because of stronger genetic drift [12]. This scenario is supported by the fact that the effect of habitat type was not observed for the rate of third codon positions, which represent more neutral substitutions. Therefore, the effect of habitat types on whole alignment should be attributable to substitutions on the first and second codon positions, and a separate analysis confirmed it (electronic supplementary material, table S5). Unexpectedly under this scenario, we found no direct negative correlation of substitution rate with occupancy. This may be explained by the fact that population size (measured by occupancy) is an evolutionarily labile trait ($\lambda = 0.10$) not consistent across a long time scale, whereas habitat type, as a more phylogenetically conserved trait ($\lambda = 0.67$), is a better predictor of long-term rate variation.

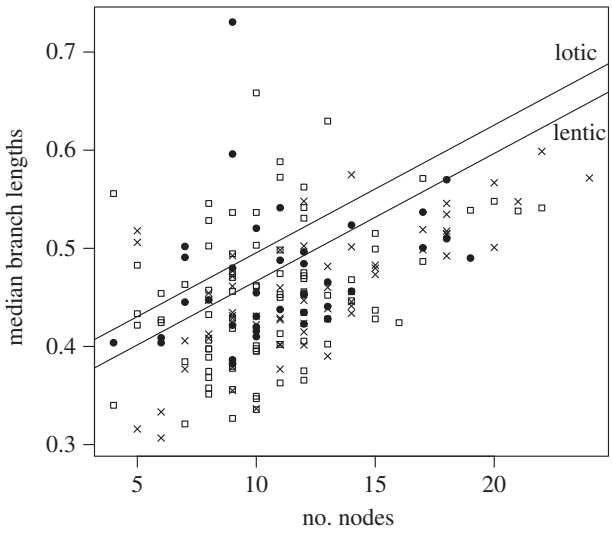

Figure 3. The effect of number of nodes on branch lengths from tip to root. Habitat type of species is indicated by three types of symbols (black dot: lotic, open square: lentic and black cross: 'both'). The solid lines are regression lines estimated by PGLS for lotic and lentic species ($y = 0.013 \times x +$ (lotic: 0.34, lentic: 0.33), $p \ll 0.001$, $R^2 = 0.24$, $\lambda = 1.0$). The interaction term between habitat type and node number was not significant.

The significant positive correlation between the rate of evolution and the number of nodes on the path from tip to root indicated the node density effect. The model-averaging analysis without controlling for node density resulted in different patterns of relative importance of explanatory variables (electronic supplementary material, table S6). Relative importance of habitat type was smaller than the analysis with node number included. In extreme cases, the effect of habitat type could possibly be overshadowed by biases introduced by node density effects owing to uneven sampling of species. This result is consistent with studies showing that using total path lengths as a measure of rate of evolution can be compromised by the node density artefact [43]. Methods to filter out the node density effect have been debated and remain controversial [7,43,57]. In this study, a clear linear relationship between branch lengths and the number of nodes was detected (figure 3). Thus, the inclusion of the node number as an explanatory variable into the regression modelling should moderate the node density effect.

A latitudinal gradient of the rate of molecular evolution is a commonly observed pattern of molecular evolution [8,58,59]. However, although latitude still had a negative effect on the rate of evolution, it was not significant in this study. The cause of the latitudinal gradient of molecular evolution has been widely debated, often associated with the cause of the latitudinal gradient of species diversity (reviewed in [60]). Hypotheses include the high energy input that results in higher metabolic rates and more mutagenic events at lower latitudes or more frequent fixation of deleterious alleles in small tropical populations. The lotic species in this study were distributed at lower latitudes on average and had higher rates of evolution. The gradient of rates might be a consequence of different life-history traits distributed along latitude.

Lastly, we do not observe direct correlations between genetic variation and rate of evolution, which would be expected if mutation rate were the only dominant factor controlling both genetic variation and substitution. Rather, the patterns are explained by occupancy and habitat types, parameters associated with effective population size. In addition, the rate at the third codon positions, which is a closer approximation of mutation rate, was not correlated with any variables considered in this study. This evidence supports the scenario that mutation rates are relatively constant or haphazard across species and both present-day genetic variation and long-term substitution rates are controlled by demographic factors, which are, in turn, controlled by species ecology.

The working phylogeny used for the comparative analysis has nodes with low support values (posterior probabilities $<$ 0.9), and this may have introduced the errors in the regression analysis. Nevertheless, because genetic variation did not show strong phylogenetic dependency, the effect of a suboptimal tree is probably minimal for the regression model of genetic variation. For the rate of evolution, the inaccuracy may have introduced biases as the branch lengths were estimated along the species phylogeny. However, the different rates between lotic and lentic groups observed in this study are unlikely to be an artefact. There were no differences in the levels of node support on the trees between lotic and lentic groups (electronic supplementary material, figure S3), therefore there should not be systematic biases affecting branch lengths of each type.

Our analyses were conducted with species solely delimited with DNA sequences, which allowed us to compare patterns of genetic variation between comparable units with positive evidence of independent evolution. Complete identification of specimens to taxonomic species would have been useful for comparison, but was not feasible here with available resources. DNA-based species delimitation is particularly useful when taxonomic information is not easily accessible, for example in a rapid survey of unstudied biota [61] or in taxonomically difficult groups such as fungi or meiofauna [62,63]. Because of the recent advances in sequencing technology, such as 'metabarcoding' [64], data without complete taxonomic identification are becoming more common in biodiversity surveys. The correlations between species' ecological and genetic parameters detected in this study were largely consistent with predictions from conventional studies, which supports the use of DNA-based delimitation as a tool to study data without clear taxonomic information.

We conclude that different correlates explain variation within species versus rates of molecular evolution between species. Neutral theory predicts higher genetic variation in larger populations, but that substitution rates of neutral mutations are independent of population size and substitution rates of nearly-neutral mutations are faster in smaller populations. Our results broadly support these alternative predictions: genetic variation was greatest in larger populations, whereas substitution rates were fastest in lotic species (which have smaller populations on average than lentic species). While we cannot identify mechanisms from correlations, and it is possible that other factors correlated with these variables influenced the patterns, our results show how comparative studies of DNA barcoding-type data can provide insights into correlates of molecular evolution.

Data accessibility. The CO1 sequences are available at GenBank under accession numbers JN840019–JN845080. Supplementary data were deposited in Dryad (doi:10.5061/dryad.926pq).

Acknowledgements. We thank Michael Monaghan and Anna Papadopoulou for discussions on this manuscript, Johannes Bergsten and other members of the Natural History Museum for collecting and sequencing the water beetles, and the anonymous reviewers for useful comments.

Funding statement. This work is partially supported by the NERC, UK (Grant No: NE/C510908/1).

References

1. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M. 2012 Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10**, e1001388. (doi:10.1371/journal.pbio.1001388)
2. Kimura M. 1984 *The neutral theory of molecular evolution*. Cambridge, UK: Cambridge University Press.
3. Charlesworth B, Charlesworth D, Barton NH. 2003 The effects of genetic and geographic structure on neutral variation. *Annu. Rev. Ecol. Evol. Syst.* **34**, 99–125. (doi:10.1146/annurev.ecolsys.34.011802.132359)
4. Knowles LL. 2006 Statistical phylogeography. *Annu. Rev. Ecol. Evol. Syst.* **40**, 593–612. (doi:10.1146/annurev.ecolsys.38.091206.095702)
5. Bazin E, Glémin S, Galtier N. 2006 Population size does not influence mitochondrial genetic diversity in animals. *Science* **312**, 570–572. (doi:10.1126/science.1122033)
6. Bromham L, Penny D. 2003 The modern molecular clock. *Nat. Rev. Genet.* **4**, 216–224. (doi:10.1038/nrg1020)
7. Lanfear R, Welch JJ, Bromham L. 2010 Watching the clock: studying variation in rates of molecular evolution between species. *Trends Ecol. Evol.* **25**, 495–503. (doi:10.1016/j.tree.2010.06.007)
8. Davies TJ, Savolainen V, Chase MW, Moat J, Barraclough TG. 2004 Environmental energy and evolutionary rates in flowering plants. *Proc. R. Soc. Lond. B* **271**, 2195–2200. (doi:10.1098/rspb.2004.2849)
9. Welch JJ, Bininda-Emonds ORP, Bromham L. 2008 Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol. Biol.* **8**, 53. (doi:10.1186/1471-2148-8-53)
10. Santos JC. 2012 Fast molecular evolution associated with high active metabolic rates in poison frogs. *Mol. Biol. Evol.* **29**, 2001–2018. (doi:10.1093/molbev/mss069)
11. Gossmann TI, Keightley PD, Eyre-Walker A. 2012 The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol. Evol.* **4**, 658–667. (doi:10.1093/gbe/evs027)
12. Woolfit M, Bromham L. 2005 Population size and molecular evolution on islands. *Proc. R. Soc. B* **272**, 2277–2282. (doi:10.1098/rspb.2005.3217)
13. Manier MK, Arnold SJ. 2006 Ecological correlates of population genetic structure: a comparative approach using a vertebrate metacommunity. *Proc. R. Soc. B* **273**, 3001–3009. (doi:10.1098/rspb.2006.3678)
14. Ohta T. 1992 The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**, 263–286. (doi:10.2307/2097289)
15. Burney CW, Brumfield RT. 2009 Ecology predicts levels of genetic differentiation in neotropical birds. *Am. Nat.* **174**, 358–368. (doi:10.1086/603613)
16. Papadopoulou A, Anastasiou I, Spagopoulou F, Stalimerou M, Terzopoulou S, Legakis A, Vogler AP. 2011 Testing the species–genetic diversity correlation in the Aegean archipelago: toward a haplotype-based macroecology? *Am. Nat.* **178**, 241–255. (doi:10.1086/660828)
17. Riginos C, Buckley YM, Blomberg SP, Treml EA. 2014 Dispersal capacity predicts both population genetic structure and species richness in reef fishes. *Am. Nat.* **184**, 52–64. (doi:10.1086/676505)
18. Papadopoulou A, Anastasiou I, Keskin B, Vogler AP. 2009 Comparative phylogeography of tenebrionid beetles in the Aegean archipelago: the effect of dispersal ability and habitat preference. *Mol. Ecol.* **18**, 2503–2517. (doi:10.1111/j.1365-294X.2009.04207.x)
19. Baselga A, Fujisawa T, Crampton-Platt A, Bergsten J, Foster PG, Monaghan MT, Vogler AP. 2013 Whole-community DNA barcoding reveals a spatio-temporal continuum of biodiversity at species and genetic levels. *Nat. Commun.* **4**, 1892. (doi:10.1038/ncomms2881)
20. Ballard JWO, Whitlock MC. 2004 The incomplete natural history of mitochondria. *Mol. Ecol.* **13**, 729–744. (doi:10.1046/j.1365-294X.2003.02063.x)
21. Meiklejohn CD, Montooth KL, Rand DM. 2007 Positive and negative selection on the mitochondrial genome. *Trends Genet.* **23**, 259–263. (doi:10.1016/j.tig.2007.03.008)
22. Galtier N, Nabholz B, Glémin S, Hurst GDD. 2009 Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol. Ecol.* **18**, 4541–4550. (doi:10.1111/j.1365-294X.2009.04380.x)
23. Gaston KJ, Blackburn TM, Greenwood JJD, Gregory RD, Quinn RM, Lawton JH. 2000 Abundance–occupancy relationships. *J. Appl. Ecol.* **37**, 39–59. (doi:10.1046/j.1365-2664.2000.00485.x)
24. Blackburn TM, Cassey P, Gaston KJ. 2006 Variations on a theme: sources of heterogeneity in the form of the interspecific relationship between abundance and distribution. *J. Anim. Ecol.* **75**, 1426–1439. (doi:10.1111/j.1365-2656.2006.01167.x)
25. Gaston KJ, He F. 2011 Species occurrence and occupancy. In *Biological diversity: frontiers in measurement and assessment* (eds AE Magurran, BJ McGill), pp. 141–151. Oxford, UK: Oxford University Press.
26. Freckleton RP, Harvey PH, Pagel M. 2002 Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* **160**, 712–726. (doi:10.1086/343873)
27. Marten A, Brändle 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)
28. Ribera I. 2008 Habitat constraints and the generation of diversity in freshwater macroinvertebrates. In *Aquatic insect: challenges to populations* (eds J Lancaster, RA Briers), pp. 289–311. Wallingford, UK: CAB International.
29. Abellán P, Ribera I. 2011 Geographic location and phylogeny are the main determinants of the size of the geographical range in aquatic beetles. *BMC Evol. Biol.* **11**, 344. (doi:10.1186/1471-2148-11-344)
30. Arribas P, Velasco J, Abellán P, Sánchez-Fernández D, Andújar C, Calosi P, Millán A, Ribera I, Bilton DT. 2012 Dispersal ability rather than ecological tolerance drives differences in range size between lentic and lotic water beetles (Coleoptera: Hydrophilidae). *J. Biogeogr.* **39**, 984–994. (doi:10.1111/j.1365-2699.2011.02641.x)
31. Subramanian S. 2013 Significance of population size on the fixation of nonsynonymous mutations in genes under varying levels of selection pressure. *Genetics* **193**, 995–1002. (doi:10.1534/genetics.112.147900)
32. Stamatakis A. 2006 RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690. (doi:10.1093/bioinformatics/btl446)
33. Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K. 2007 Estimating divergence times in large phylogenetic trees. *Syst. Biol.* **56**, 741–752. (doi:10.1080/10635150701613783)
34. Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP. 2006 Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol.* **55**, 595–609. (doi:10.1080/10635150600852011)
35. Ezard T, Fujisawa T, Barraclough TG. 2009 SPLITS: species' limits by threshold statistics. (https://r-forge.r-project.org/projects/splits/)
36. Edgar RC. 2004 MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113. (doi:10.1186/1471-2105-5-113)
37. Ronquist F *et al*. 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542. (doi:10.1093/sysbio/sys029)
38. Darriba D, Taboada GL, Doallo R, Posada D. 2012 jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772. (doi:10.1038/nmeth.2109)
39. Rambaut A, Drummond AJ. 2007 Tracer v1.4. (http://beast.bio.ed.ac.uk/Tracer%20)
40. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88. (doi:10.1371/journal.pbio.0040088)
41. Shimodaira H, Hasegawa M. 1999 Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116. (doi:10.1093/oxfordjournals.molbev.a026201)
42. Ribera I, Vogler AP. 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:10.1111/j.1095-8312.2000.tb01240.x)

43. Hugall AF, Lee MSY. 2007 The likelihood node density effect and consequences for evolutionary studies of molecular rates. *Evolution* **61**, 2293–2307. (doi:10.1111/j.1558-5646.2007.00188.x)
44. Pagel M. 1999 Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884. (doi:10.1038/44766)
45. Orme CDL. 2012 Caper: comparative analyses of phylogenetics and evolution in R. (http://cran.r-project.org/web/packages/caper/)
46. Burnham KP, Anderson DR. 2002 *Model selection and multimodel inference: a practical information-theoritic approach*, 2nd edn. New York, NY: Springer Science+Business Media, Inc.
47. Team RC. 2013 R: a language and environment for statistical computing.
48. Paradis E, Claude J, Strimmer K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)
49. Schliep KP. 2011 phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593. (doi:10.1093/bioinformatics/btq706)
50. Paradis E. 2010 pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419–420. (doi:10.1093/bioinformatics/btp696)
51. Hof C, Brändle 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)
52. Nabholz B, Mauffrey J-F, Bazin E, Galtier N, Glemin S. 2008 Determination of mitochondrial genetic diversity in mammals. *Genetics* **178**, 351–361. (doi:10.1534/genetics.107.073346)
53. Nabholz B, Glémin S, Galtier N. 2009 The erratic mitochondrial clock: variations of mutation rate, not population size, affect mtDNA diversity across birds and mammals. *BMC Evol. Biol.* **9**, 54. (doi:10.1186/1471-2148-9-54)
54. Gillespie JH. 2000 Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* **155**, 909–919.
55. Hewitt G. 2000 The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913. (doi:10.1038/35016000)
56. Weir JT, Schluter D. 2007 The latitudinal gradient in recent speciation and extinction rates of birds and mammals. *Science* **315**, 1574–1576. (doi:10.1126/science.1135590)
57. Webster AJ, Payne RJH, Pagel M. 2003 Molecular phylogenies link rates of evolution and speciation. *Science* **301**, 478. (doi:10.1126/science.1083202)
58. Allen AP, Gillooly JF, Savage VM, Brown JH. 2006 Kinetic effects of temperature on rates of genetic divergence and speciation. *Proc. Natl Acad. Sci. USA* **103**, 9130–9135. (doi:10.1073/pnas.0603587103)
59. Gillman LN, Keeling DJ, Ross HA, Wright SD. 2009 Latitude, elevation and the tempo of molecular evolution in mammals. *Proc. R. Soc. B* **276**, 3353–3359. (doi:10.1098/rspb.2009.0674)
60. Mittelbach GG *et al*. 2007 Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecol. Lett.* **10**, 315–331. (doi:10.1111/j.1461-0248.2007.01020.x)
61. Monaghan MT, Wild R, Elliot M, Fujisawa T, Balke M, Inward DJG, Lees DC, Ranaivosolo R, Eggleton P, Barraclough TG, Vogler AP. 2009 Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Syst. Biol.* **58**, 298–311. (doi:10.1093/sysbio/syp027)
62. Tang CQ, Leasi F, Obertegger U, Kieneke A, Barraclough TG, Fontaneto D. 2012 The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proc. Natl Acad. Sci. USA* **109**, 16 208–16 212. (doi:10.1073/pnas.1209160109)
63. Millanes AM, Truong C, Westberg M, Diederich P, Wedin M. 2014 Host switching promotes diversity in host-specialized mycoparasitic fungi: uncoupled evolution in the Biatoropsis–Usnea system. *Evolution* (*NY*) **68**, 1576–1593. (doi:10.1111/evo.12374)
64. Ji Y *et al*. 2013 Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett*. **16**, 1245–1257. (doi:10.1111/ele.12162)