

## Research



**Cite this article:** Singhal S, Huang H, Title PO, Donnellan SC, Holmes I, Rabosky DL. 2017 Genetic diversity is largely unpredictable but scales with museum occurrences in a species-rich clade of Australian lizards. *Proc. R. Soc. B* **284**: 20162588. <http://dx.doi.org/10.1098/rspb.2016.2588>

Received: 22 November 2016

Accepted: 3 April 2017

**Subject Category:**

Evolution

**Subject Areas:**

evolution, genetics

**Keywords:**

population genetics, genetic diversity, Lewontin's paradox, *Ctenotus* skinks, biodiversity databases

**Author for correspondence:**

Sonal Singhal

e-mail: [sonal.singhal1@gmail.com](mailto:sonal.singhal1@gmail.com)

Electronic supplementary material is available online at <https://dx.doi.org/10.6084/m9.figshare.c.3745469>.

# Genetic diversity is largely unpredictable but scales with museum occurrences in a species-rich clade of Australian lizards

Sonal Singhal<sup>1</sup>, Huateng Huang<sup>1</sup>, Pascal O. Title<sup>1</sup>, Stephen C. Donnellan<sup>2,3</sup>, Iris Holmes<sup>1</sup> and Daniel L. Rabosky<sup>1</sup>

<sup>1</sup>Museum of Zoology and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

<sup>2</sup>South Australian Museum, North Terrace, Adelaide 5000, Australia

<sup>3</sup>Australian Centre for Evolutionary Biology and Biodiversity, University of Adelaide, Adelaide 5005, Australia

SS, 0000-0001-5407-5567

Genetic diversity is a fundamental characteristic of species and is affected by many factors, including mutation rate, population size, life history and demography. To better understand the processes that influence levels of genetic diversity across taxa, we collected genome-wide restriction-associated DNA data from more than 500 individuals spanning 76 nominal species of Australian scincid lizards in the genus *Ctenotus*. To avoid potential biases associated with variation in taxonomic practice across the group, we used coalescent-based species delimitation to delineate 83 species-level lineages within the genus for downstream analyses. We then used these genetic data to infer levels of within-population genetic diversity. Using a phylogenetically informed approach, we tested whether variation in genetic diversity could be explained by population size, environmental heterogeneity or historical demography. We find that the strongest predictor of genetic diversity is a novel proxy for census population size: the number of vouchered occurrences in museum databases. However, museum occurrences only explain a limited proportion of the variance in genetic diversity, suggesting that genetic diversity might be difficult to predict at shallower phylogenetic scales.

## 1. Introduction

One of the fundamental characteristics of a species is the amount of genetic variation segregating in its populations [1], which can impact several aspects of a species's biology, including phenotypic variation and response to selection [2]. Evolutionary biologists have long sought to understand the factors that influence levels of genetic diversity in natural populations, from both theoretical and empirical perspectives [3–5]. Theory predicts that the amount of genetic diversity in a given population is straightforward if it conforms to a simple Wright–Fisher model (constant population size; panmixia; no selection); genetic diversity should scale positively and linearly as a function of census population size ( $N$  or  $N_c$ ) and mutation rate [6]. Indeed, many species appear to have levels of variation that correspond with ecological approximations of their census population sizes. For example, previous studies have documented greater genetic diversity in species with larger ranges than those with smaller ranges, in mainland species relative to island species and in species of high abundance relative to low abundance taxa [7–10].

Despite the simple prediction that genetic diversity should be positively correlated with  $N_c$ , many studies have found no correlation between species genetic diversity and aspects of species ecology and geography that are expected to be proxies for total species abundance [11–15]. Further, those studies that have reported positive correlations all show a puzzling pattern: species exhibit a much narrower range of genetic diversity (suggesting a narrower range of effective population sizes,  $N_e$ ) than one would expect given their range of  $N_c$  [5,9,16].

To explain the lack of correlation between  $N_c$  and genetic diversity in some datasets as well as the mismatch in how widely these values range in others, researchers have invoked organismal traits and non-neutral and non-equilibrium processes that can decouple  $N_c$  and  $N_e$ . These include changes in population size through time, variance in offspring number, inbreeding, and the combined effects of background selection and selective sweeps (i.e. linked selection) [17].

Previous studies investigating the drivers of genetic diversity have sampled broadly across the tree of life, focusing on sets of taxa that vary greatly in their life-history strategies, body sizes and habitat. These studies have suggested that traits such as propagule size [18], dispersal strategy [7] and mating system [19] can predict genetic diversity. However, we typically only see large variation in these traits when we sample across broad phylogenetic scales (e.g. animal phyla [18]) and it is unclear whether the results of such studies can be generalized to account for patterns of genetic diversity among closely related sets of species that share many ecological traits.

In this article, we study the determinants of genetic diversity across a single species-rich clade of scincid lizards from Australia in the genus *Ctenotus*. *Ctenotus* consists of 101 nominal species, most of which are restricted to Australia's arid or semi-arid biomes [20]. Across this set of ecologically similar and closely related species, we tested three, non-exclusive hypotheses that might explain interspecific variation in genetic diversity. First, we used proxies for the census population size to test the basic population genetic prediction that genetic diversity should be correlated with  $N_c$ , under the implicit assumption that  $N_c$  and  $N_e$  are correlated. Second, we tested the hypothesis that increased environmental heterogeneity leads to increased genetic diversity. Heterogeneity can favour different alleles in different environmental contexts, leading to increased genetic diversity even within populations [21]. As support for this hypothesis, we would expect proxies for environmental heterogeneity to positively correlate with diversity levels. Finally, we considered historical explanations by testing the role of demographic shifts through time on genetic diversity patterns [22]. Population extirpations and range expansions can have dramatic effects on genetic diversity [23], and these changes in census population size can decouple  $N_c$  and  $N_e$ , and thus  $N_c$  and genetic diversity. We predicted that populations characterized by more stable demographic histories should have greater genetic diversity than those with dynamic histories.

By collecting genome-wide data for 76 nominal species in *Ctenotus*, we delineated a comparable set of operational taxonomic units (OTUs) across the genus and inferred levels of genetic variation. We then collated a set of morphological, geographical and ecological attributes relating to the three hypotheses described above, and we tested which—if any—of these factors best explain the observed variation in genetic diversity across *Ctenotus*.

## 2. Methods

### (a) Sampling and data collection

We obtained tissue samples for 575 individuals from 76 of the 101 nominal *Ctenotus* species from Australia (electronic supplementary material, figure S1, table S1). *Ctenotus* has been subject to significant taxonomic revision, because many species defined on morphology have been rearranged with the inclusion of genetic

data [24]. As such, our strategy was to sample as broadly and densely across species ranges as possible based on specimen availability from eight museums and our own field collections. However, our sampling was necessarily patchy, reflecting heterogeneity in sample availability across much of Australia [25]. The number of individuals sampled and range size per nominal species were highly correlated ( $r = 0.80$ ;  $p < 2 \times 10^{-16}$ ).

Both to delimit major lineage OTUs and to estimate levels of genetic variation within species, we collected genome-wide data for each individual using double-digest restriction-associated DNA (ddRAD) sequencing [26]. Details on molecular work can be found in the electronic supplementary material.

### (b) Species delimitation

Many nominal species in *Ctenotus*, as in other squamate species [27,28], comprise multiple, cryptic lineages [24]. To account for this largely undescribed cryptic diversity and to generate a set of comparable OTUs for analysis, we used a quantitative species delimitation approach. Even new, more computationally efficient methods for species delimitation cannot handle a multi-locus dataset of this size [29]. Instead, we used the coalescent-based approach implemented in generalized mixed yule coalescent (GMYC) [30]. GMYC requires an ultrametric tree inferred with a single locus per individual. In most typical implementations, researchers use mtDNA; however, in *Ctenotus*, mtDNA introgression is pervasive across the phylogeny [31]. We instead inferred an ultrametric tree using RAxML v. 8.2.0 [32] and TREEPL [33] on a concatenated alignment of multi-locus ddRAD data ( $N = 1232$ , length = 320 037 bp). Our approach explicitly violated GMYC's assumption of a single-lineage history, but we believe that the method can still identify nodes in phylogenies where lineages transition from behaving under a population-level coalescent model to a species-level birth-death model. Crossreferencing the GMYC delimitations to geography and nominal species boundaries suggested that the model performed well under this application, largely recovering traditional species groupings (see 'Results').

### (c) Measuring genetic diversity

The genetic diversity of a species is a function of both levels of standing genetic variation within populations and differentiation across populations. We are primarily interested in within-population diversity because understanding its origins is a longstanding evolutionary question [4], and it is generally less sensitive to migration rates [17]. We also present estimates of species-wide diversity for both the nuclear and mitochondrial genomes. To calculate within-population diversity, we estimated the average pairwise difference ( $\pi$ ) per population and then averaged this across populations [34]. Because we sampled an average of 44 K loci per individual, summarizing patterns across all loci for a given individual allowed us to accurately estimate population-level parameters [35]. Thus, we treated each individual as a population. We also calculated species-wide  $\pi$  across all individuals assigned to a particular OTU, using an approach that extends the method introduced in [34] to allow for different levels of missingness across loci [36]. Preliminary analyses uncovered a technical artefact in which the number of individuals used in single nucleotide polymorphism (SNP) calling positively biased the number of SNPs recovered (electronic supplementary material, figure S2). Results for species-wide  $\pi$  should be accordingly treated with caution. Because our bioinformatics pipeline for estimating within-population  $\pi$  called SNPs per each individual, within-population  $\pi$  should not be subject to this technical artefact. Finally, we calculated mtDNA  $\pi$  per OTU based on previously published data [24].

To estimate within-population  $\pi$ , we first generated pseudo-reference genomes across all individuals for a given OTU using VSEARCH [37]. We then aligned reads and called variant and invariant sites using standard methods [38]. We filtered sites, retaining

sites with a quality score  $> 20$ , with  $\leq 3\times$  the median coverage for the individual to avoid collapsed paralogues, and with  $\geq 10\times$  coverage to ensure we could accurately call heterozygous sites. After this step, we dropped four individuals because more than  $1 \times 10^5$  of their genomic sites had sufficient coverage to call SNPs. We then used these filtered sites to estimate  $\pi$  per individual and took the mean  $\pi$  across individuals as our within-population estimate for  $\pi$ . See electronic supplementary material for further details on our bioinformatics pipeline.

For OTUs with three or more sampled individuals ( $N = 42$ ), we additionally summarized patterns of genetic variation across OTUs by using ANGSD v. 0.910 to infer Tajima's D [39] and LAMARC v. 2.1.10 to infer population growth rate [40]. For these analyses, we used three randomly chosen individuals from the most common genetic cluster as identified by ADMIXTURE v. 1.23 [41] because we found that calling SNPs across variable sample sizes introduces technical bias (electronic supplementary material, figure S2).

#### (d) Species tree

As a phylogenetic framework for the comparative analyses described below, we constructed a species tree across OTUs. We first identified homologous loci across the OTU pseudo-reference genomes using VSEARCH, resulting in 14 187 loci that were, on average, 44% complete across taxa. We then inferred gene trees for each locus using RAxML and constructed a species tree across these gene trees using ASTRID [42]. We inferred 100 bootstraps using a gene-tree bootstrapping approach, in which we took the bootstraps from the gene tree inference and used them to infer bootstrapped trees. ASTRID only provides a topology for the species tree; thus, to infer branch lengths on the tree, we ran RAxML with a concatenated alignment of the 14 187 loci and the fixed topology from ASTRID. We rooted the tree using data from a broader phylogeny of Australian skinks [43] and fitted a chronogram to the phylogeny using penalized likelihood.

#### (e) Geographical ranges

Many of our proposed factors (see 'Ecological and geographical drivers of genetic diversity') involved aspects of species' geography, but accurate ranges for Australian squamates are not available. Thus, we modified an approach that constructs occurrence record-based ranges, defined as the intersection of an alpha-hull polygon and an ecological niche model (ENM) [44]. We used MAXENT v. 3.3.3 k [45] with 22 environmental variables, which included the 19 WorldClim v. 1.4 variables, elevation, actual evapotranspiration and an aridity index [46]. This approach avoids over-predicting the species range into regions where the species has never been found and is particularly useful for climatically homogeneous landscapes like Australia. We then modified geographical ranges for nominal species to reflect revised OTU delimitations; further details on these modifications and on this approach are available in the electronic supplementary material.

#### (f) Ecological and geographical drivers of genetic diversity

We identified a series of factors that allowed us to test the three hypotheses for why genetic diversity varies across species, as well as a few possible confounding factors. We then collated previously published and publicly available datasets to characterize these factors for each OTU. Where necessary, we modified attributes of nominal species to match the OTUs.

##### (i) Factors relating to census population size

To test the basic prediction that genetic diversity should be correlated with  $N_c$ , we included three proxies for  $N_c$  including range

size [7], body size [47] and number of museum occurrence records [44,48]. We used the inferred geographical ranges to calculate range size. For body size, we used measurements from Rabosky *et al.* that captured eight standard aspects of lizard external morphology, including snout-vent length, head depth, head width and limb proportions [43]. An average of 5.9 individuals was measured per species, so we calculated species means and summarized them through a principal component analysis. The first axis explained 95% of the variation and loaded on body size; we also retained the second axis that explained an additional 3% of the variation. For cryptic OTUs within a nominal species, we ascribed to them the same morphological measurements as the nominal species.

To determine the number of museum occurrences associated with each OTU, we summarized counts from data downloaded from Atlas of Living Australia on 5 November 2015, using only museum specimens (i.e. points that were classified as 'Preserved-Specimen') and dropping any points that occurred more than one degree from the range as these probably reflected either mis-identification or database errors. For OTUs that collapsed multiple nominal species, we summed the counts across those nominal species; for OTUs split from nominal species, we counted only those points from the nominal species that occurred within the OTU range; for OTUs with more complicated relationships to their nominal species, we set counts to missing.

##### (ii) Factors relating to environmental heterogeneity

To explore the hypothesis that current environmental heterogeneity leads to increased genetic diversity, we included two proxies for environmental heterogeneity: the elevation range and climatic space spanned by each geographical range. We calculated the elevation span encompassed by each range using elevation layers from WorldClim. To infer climatic heterogeneity, we extracted environmental data for 50 000 random points across Australia using the same 22 variables as used in the ENM. We then used a principal component analysis to ordinate these data. We retained the first three axes, which explained 79% of the variation. For each range, we summarized variation across these three axes, allowing us to infer the breadth of environmental space for a given geographical range across standardized units.

##### (iii) Factors relating to demographic history

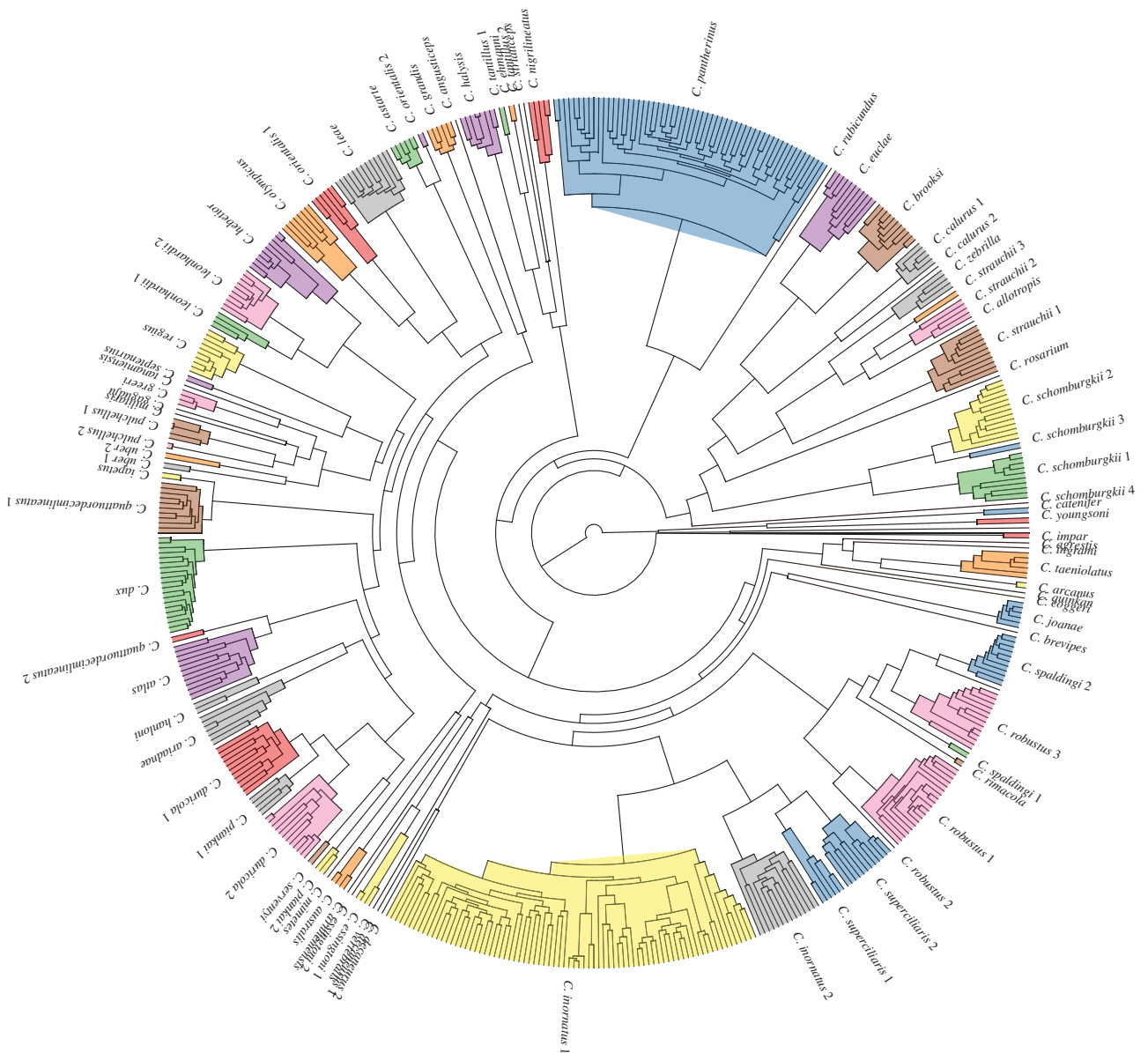
To test if historical demographic changes impacted genetic diversity [22], we measured the stability of these ranges through time. First, we inferred past habitat suitability by projecting our ENM for each OTU to past climates (mid-Holocene, Last Glacial Maximum and last interglacial) [49]. Then, for all museum occurrences in an OTU, we extracted the probability of occurrence at each historical time point, took the geometric mean of these values through time and averaged across all points to calculate average historical suitability.

##### (iv) Possible confounding variables

We attempted to account for possible confounders by including the latitudinal midpoint of taxon ranges and the terminal branch length for each OTU in the species tree as a crude estimate for how long the taxon as existed as an independent entity.

##### (g) Hypothesis testing

We first determined if our diversity metrics and predictor variables showed phylogenetic signal by estimating Pagel's  $\lambda$  [50]. Then, to determine which factors best predict levels of genetic variation, we implemented the phylogenetic multiple regression strategy outlined by Kisel & Barraclough [51]. This approach infers the relative importance of a series of factors in predicting a variable of interest by determining if their inclusion in linear models leads to



**Figure 1.** Phylogeny of all samples ( $N = 555$ ) used in this study, based on a concatenated alignment of ddRAD loci and inferred using RAxML. Colours demarcate the clades inferred to be putative species (operational taxonomic units; OTUs) by GMYC, and species names indicate possible names for these OTUs based on their relationship to nominal species. While the majority of OTUs are synonymous with nominal taxa, a number of nominal forms (e.g. *C. decaneurus*, *C. leonhardii* and *C. schomburgkii*) have been split into multiple distinct OTUs.

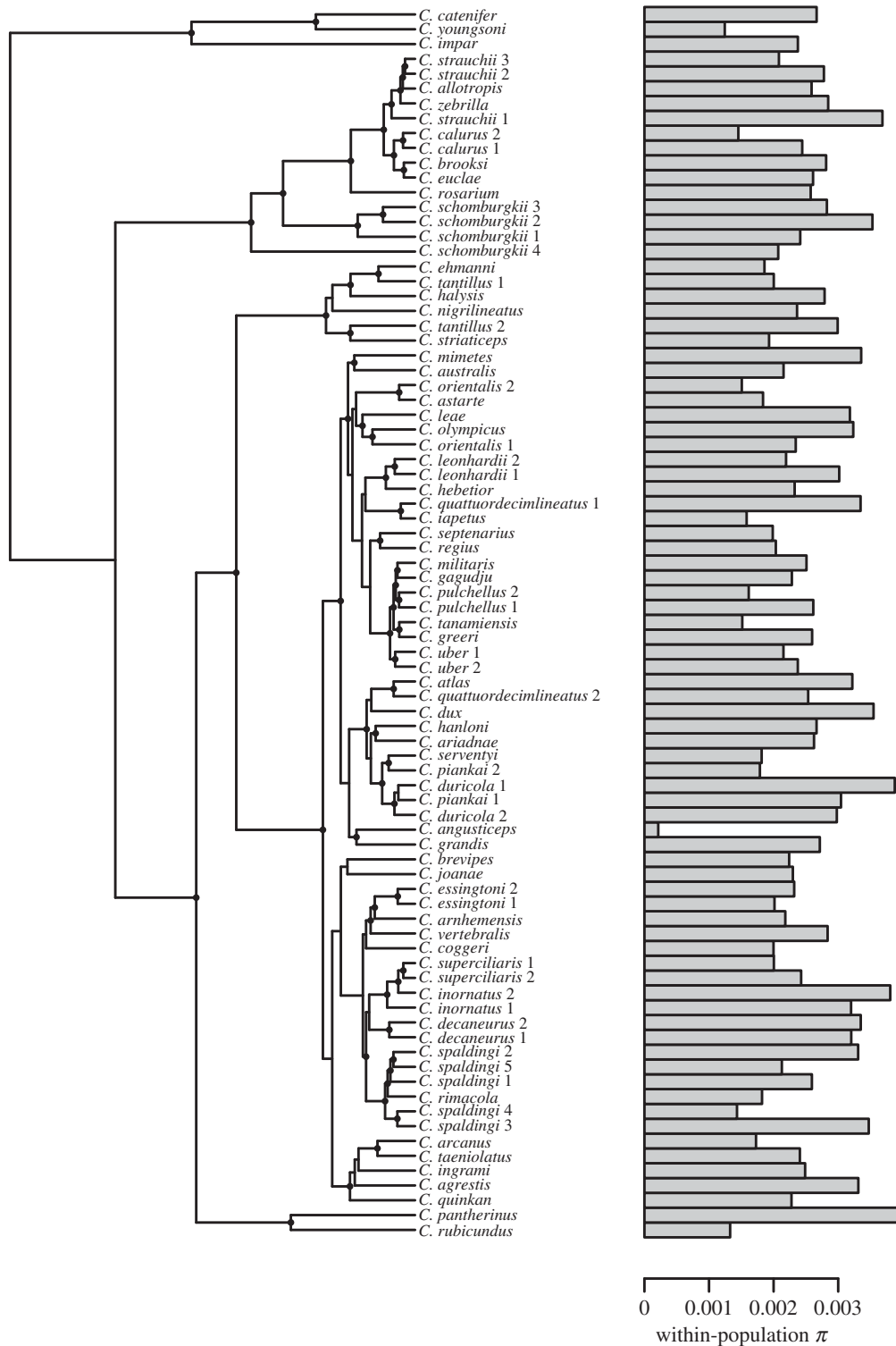
better-fitting models. We first dropped predictor variables that were highly correlated with each other ( $r > 0.7$ ). We then ran phylogenetic general linear models (PGLMs) for all possible additive models given our predictor variables. We calculated the Akaike weight for each model [52]. We then determined the relative importance of each factor, which indicates how much a given factor contributes to highly scoring models. Relative importance is calculated as the sum of the relative Akaike weights for the models in which that factor appears. The  $p$ -value and regression coefficient for each factor were calculated by weighting individual model estimates by the relative Akaike weight for that model. To cross-validate these results, we repeated this approach 100 times, randomly subsampling 80% of the complete dataset in each bootstrap.

### 3. Results

The 76 nominal species in our dataset were revised to include 83 putative species-level OTUs, which contained anywhere from 1 to 78 individuals, with a mean and median value of 6.7 and 3

individuals, respectively (figure 1). Furthermore, 52% of these OTUs were synonymous with nominal species, 39% split a nominal taxon, 7% combined taxa, and 2% were compound OTUs that both split and combined elements of existing taxa. The nodes delimiting OTUs are well-supported (electronic supplementary material, figure S3a), and alternative tree inference methods recover nearly identical groupings (electronic supplementary material, figure S3b). Most discrepancies between traditional taxonomy and our revised delimitation arise from discovery of deep, cryptic lineages within species, as has been commonly found in many squamate species [27,28].

Using an average of 2.6 million sites per individual (electronic supplementary material, table S1), we estimated a range of within-population  $\pi$  values from  $2.14 \times 10^{-4}$  to  $3.99 \times 10^{-3}$ , with a standard deviation of  $6.8 \times 10^{-4}$  (figure 2). Species-wide estimates of  $\pi$  show a greater variation across lineages than within-population  $\pi$ ; nuclear  $\pi$  ranges from  $2.14 \times 10^{-4}$  to  $1.06 \times 10^{-2}$ , with a standard deviation of  $2.0 \times 10^{-3}$  (electronic supplementary material, figure S4), and mtDNA  $\pi$

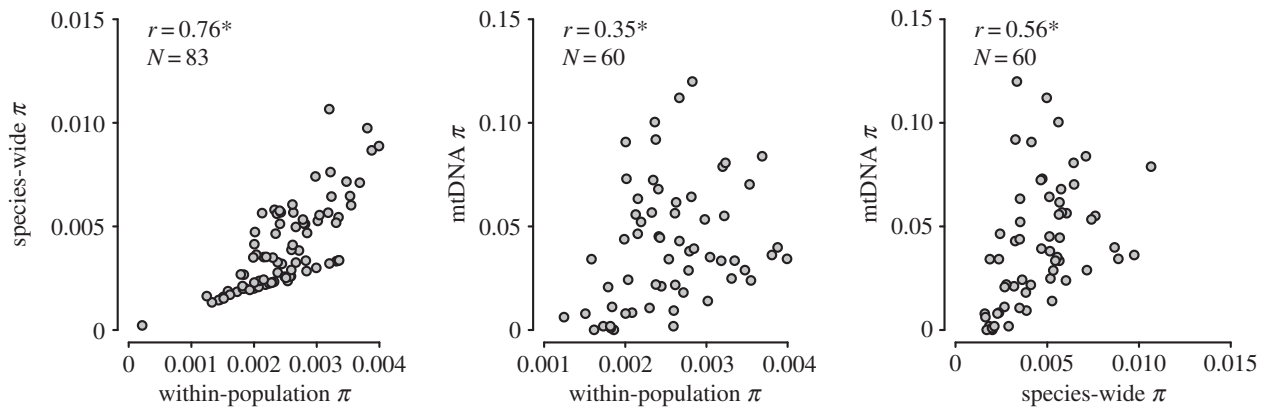


**Figure 2.** The ‘species tree’ for the 83 OTUs in *Ctenotus*, as inferred by ASTRID, shown with values of estimates of within-population nucleotide diversity ( $\pi$ ). Nodes labelled with circles have bootstrap support  $>95\%$ ; tree with full bootstrap support shown in electronic supplementary material, figure S14. We recover no evidence for phylogenetic signal in  $\pi$  across these OTUs ( $\lambda < 1 \times 10^{-4}$ ,  $p$ -value = 1; electronic supplementary material, table S2), suggesting that conserved traits do not explain interspecific variation in genetic diversity at this scale.

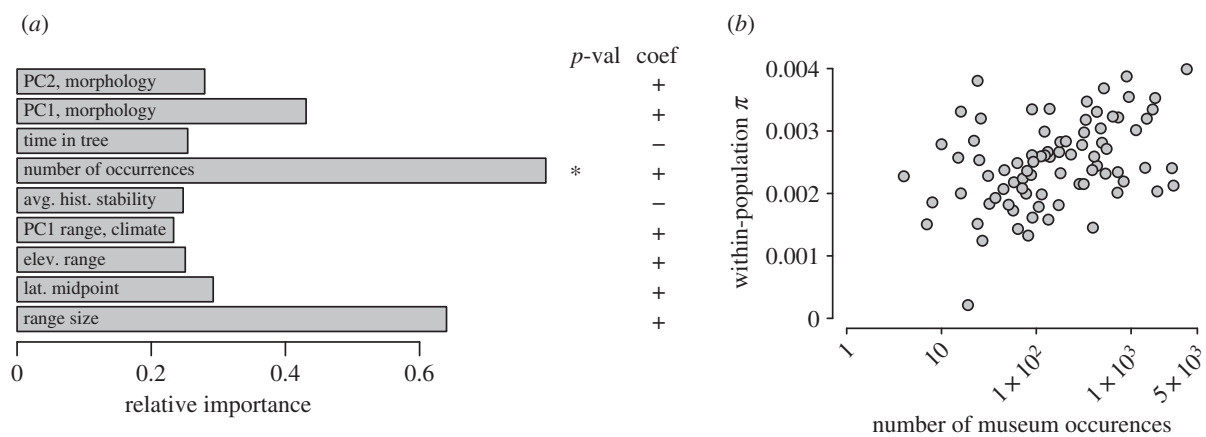
ranged from  $8.75 \times 10^{-4}$  to 0.120, with a standard deviation of  $3.0 \times 10^{-2}$  (electronic supplementary material, figure S5). All estimates of  $\pi$  are significantly correlated. However, the two measures of nuclear  $\pi$  (species-wide and within-population  $\pi$ ) and the two measures of species-wide  $\pi$  (nuclear and mtDNA  $\pi$ ) are much more strongly correlated (figure 3).

Our full model included 11 independent variables to explain variation in genetic diversity across OTUs. We dropped the variables describing heterogeneity in PC2 and PC3 climatic

space because they are highly correlated with elevation range ( $r > 0.7$ ). The remaining nine independent variables (number of museum occurrences, range size, PC1 and PC2 describing morphology, elevation range, heterogeneity in PC1 climatic space, average historical stability, range latitudinal midpoint and branch length for each OTU) are correlated below  $r < 0.7$  (electronic supplementary material, figure S6). None of our measures of genetic diversity show phylogenetic signal; some closely related OTUs exhibit very dissimilar levels of



**Figure 3.** Correlations between the three indices of nucleotide diversity used in this study: within-population nucleotide diversity ( $\pi$ ), species-wide  $\pi$  and species-wide mitochondrial DNA (mtDNA)  $\pi$ . Reported are Spearman correlations and the number of comparisons included in each correlation; asterisks reflect significant correlations. Both nuclear estimates of  $\pi$  and both species-wide estimates of  $\pi$  are strongly correlated, whereas within-population  $\pi$  and mtDNA  $\pi$  are more weakly correlated.



**Figure 4.** Relative importance of morphological, ecological and geographic factors in explaining within-population nucleotide diversity ( $\pi$ ) from phylogenetic multi-predictor models. These factors test three hypotheses for why genetic diversity varies across species. (a) Shown are the relative importance,  $p$ -value significance and directionality of coefficient for each variable as summarized across all additive models, weighted by relative AIC weights. (b) The relationship between  $\pi$  and the sole variable showing significance, the number of occurrences in museum databases. This weak but significant correlation suggests that the number of museum occurrences is a coarse proxy for census population size.

genetic variation (figure 2; electronic supplementary material, table S2, figures S4 and S5).

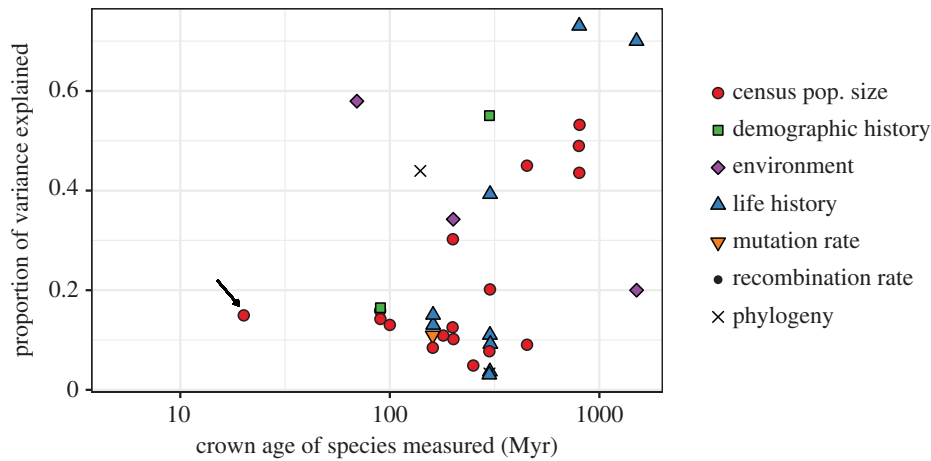
Of the factors included in a phylogenetic linear model for within-population  $\pi$ , the factor with the highest relative importance—and the only significant factor overall—is the number of museum occurrences for that OTU (figure 4; electronic supplementary material, table S3). Because the estimate of phylogenetic signal ( $\lambda$ ) for the highest-ranking models was negligible, we fitted non-phylogenetic linear models to the same data, finding that the model with the highest relative AIC weight has an adjusted  $r^2 = 0.17$  (electronic supplementary material, table S3). Coefficient of variation results confirm these findings but recover broad variances around estimates of relative importance and significance, suggesting that range size also probably predicts genetic diversity and that additional sampling would help strengthen these findings (electronic supplementary material, figure S7). No factors significantly predict species-wide mtDNA  $\pi$  (electronic supplementary material, figure S8), and range size, average historical stability and number of museum occurrences significantly predict species-wide nuclear  $\pi$  (electronic supplementary material, figure S9).

## 4. Discussion

Previous studies have been equivocal regarding whether census population sizes predict genetic diversity [11–15]. Here, the best predictor of genetic diversity is the number of occurrences in museum records, both confirming a basic prediction of population genetics that census population size should correlate with levels of segregating variation and indirectly suggesting that museum occurrences are a rough proxy for  $N_c$ . We find limited support for other proxies of census population size such as range size (figure 4; electronic supplementary material, figure S7), and we recover no support for hypotheses that suggest current environmental heterogeneity or historical stability drives variation in genetic diversity.

### (a) The unpredictability of genetic diversity

Although museum occurrences emerged as a significant predictor of genetic diversity, they only explain 15% of the variation in patterns of genetic diversity (electronic supplementary material, table S3). Our best model, which included both range size and museum occurrences, offers only marginally more



**Figure 5.** Summary of major factors predicting genetic diversity across our survey of 53 studies as a function of phylogenetic scale. We classified explanatory variables from previous studies (electronic supplementary material, table S5) into six general hypotheses for why genetic diversity varies: census population size (and its proxies), demographic history, environment and environmental variation, life-history traits, mutation rate variation and recombination rate variation. Also included as a factor is phylogeny, which suggests that there are unknown or unmeasured phylogenetically conserved traits that partially explain the variance. Twenty-six studies reported the proportion of variance explained; we have no data for recombination rate. The arrow indicates the current study, which is an outlier in investigating these patterns across a narrow phylogenetic scale. Studies explain an average of 31% of the variance, and the crown age of species in a study is positively correlated with proportion of variance explained (adj.  $r^2 = 0.22$ ;  $p < 0.005$ ). See electronic supplementary material, table S5, for details on these studies. (Online version in colour.)

explanatory power. We considered a broad set of predictor variables—including proxies for a range of ecological, historical and demographic traits—yet we observed little power to predict the variation in genetic diversity across the species-level lineages in our dataset.

Some of this low power is likely to be attributable to measurement error in both genetic diversity and its explanatory factors. It also points to the complexity of the biological factors impacting genetic diversity, some of which our study did not consider. At the genomic level, both variation in the rate of mutation and the strength of linked selection impact genetic diversity [16]. While mutation rate varies across the tree of life, data from substitution rates suggest that the mutation rate is likely to be conserved across closely related taxa with similar life histories [53]. As for the role of linked selection, our genomic data are anonymous, so we cannot infer the recombination rates for the regions that harbour these loci or reconstruct their selection history. However, although linked selection can depress diversity levels [17], it is unlikely to affect genome-wide variation [16,17]. Mutation rate variation and linked selection are therefore unlikely to be the culprits behind this unexplained variation.

Our study also did not fully account for variation in life history and mating system, both of which are known to impact genetic diversity [16]. Previous surveys have found levels of diversity vary between social versus solitary insects [54] and between selfing versus outcrossing plants [9,55]. More recent work suggests that the r–K continuum (i.e. the trade-off between high fecundity and low parental investment versus low fecundity and high parental investment) explains more than 70% of the variation in genetic diversity seen across animals [18]. Collating across the limited field-based studies of *Ctenotus* life history, we find variation in one life-history trait, clutch size (electronic supplementary material, table S4). Clutch size is a positive but non-significant predictor of genetic diversity (electronic supplementary material, figure S10), although it shows only modest variation among

species. Moreover, this life-history variation pales in comparison with ecological differences identified by other studies. For example, even after removing *C. angusticeps*, whose genetic diversity is an outlier in the genus, the range of genetic diversity seen in *Ctenotus* spans that of species along the r–K continuum from termites and ants to penguins and tortoises [18].

This points to a more general pattern that, despite being ecologically similar and closely related, *Ctenotus* exhibits levels of genetic diversity seen across much more ecologically and phylogenetically distinct taxa. Comparing estimates of genetic diversity in *Ctenotus* with other species, *Ctenotus* shows the greatest overlap with other vertebrates (electronic supplementary material, figure S11) [9,18]. However, these vertebrates span taxa as different as house mice, rattlesnakes and grey whales. Further, factors identified by previous surveys that explain a significant portion of the variation in genetic diversity (e.g. r versus K, breeding system, size of historical refugia) appear to either explain less of the variation in *Ctenotus* or be fairly conserved across the genus (figure 5; electronic supplementary material, table S5). This review of the greater literature underlines how the phylogenetic scale at which we query genetic diversity informs our understanding and ability to explain patterns (figure 5). Accordingly, we find no evidence for phylogenetic signal across our measures of genetic diversity (electronic supplementary material, table S2), which suggests that conserved traits (including traits not included in our study) have little effect on genetic variation at this scale. Studies that have sampled a wider breadth of organisms across the tree of life (i.e. all of land plants or all of animals) have seen phylogenetic signal in diversity patterns [12,13,18,19]. This scale-dependence suggests that levels of genetic variation might be controlled by traits at multiple hierarchical levels, some of which—like breeding and mating system—are phylogenetically conserved. Thus, our study raises a number of questions about this unexplained genetic variation, and how ecologically and closely related species maintain such differing levels of diversity.

### (b) Exploring Lewontin's paradox

Our proxies for  $N_c$  and genetic diversity are correlated, but they scale non-linearly (figure 4; electronic supplementary material, figure S9); this result has been observed previously and termed 'Lewontin's paradox' [5,9]. Our two significant proxies for  $N_c$  (number of occurrences and range size) both show a thousand-fold difference across *Ctenotus*, yet our two corresponding estimates for genetic diversity (within-population and species-wide  $\pi$ ) exhibit more than an order of magnitude less difference. Thus, our results confirm studies of numerous wild species that show  $N_c$  and  $N_e$  tend to be correlated but differ greatly in magnitude [56,57].

This discrepancy in scaling could simply be because our proxies for census population size do not scale linearly with census population size itself. Alternatively, both this discrepancy and the relatively low explanatory power of our model (electronic supplementary material, table S3) could be the result of demographic processes that influence the relationship between  $N_c$  and  $N_e$ , including population structure, non-random mating and temporal variation in population size [17]. Our findings of isolation-by-distance (electronic supplementary material, figure S12) and variation in diversity throughout the range (electronic supplementary material, figure S13)—realized as much greater variation in species-wide  $\pi$  than within-population  $\pi$ —are all evidence for demography's impacts on genetic diversity. Although our analyses found no evidence that historical demographic processes impact levels of genetic variation, demography can profoundly impact how  $N_e$  and  $N_c$  covary. If more numerous species are less stable through time, then their more dynamic demography would lead  $N_e$  (and thus genetic diversity) to be more depressed relative to  $N_c$  than in species with smaller population sizes [9,58]. Indeed, we find evidence that OTUs that are more numerous have less stable habitat through time compared with less numerous OTUs, even after controlling for the confounding factor of the biomes in which OTUs are found (electronic supplementary material, figure S14). Concordantly, more numerous OTUs show a trend towards having experienced higher growth rates and lower Tajima's  $D$  values compared with less numerous OTUs (electronic supplementary material, figure S14). While these results are preliminary given the difficulty in accurately modelling historical niches and inferring past demographies, they suggest that demography might help resolve Lewontin's paradox [17].

### (c) Comparisons across metrics of diversity

Species-wide genetic diversity captures both local levels of diversity and differentiation across the range, and, not surprisingly, species-wide estimates deviate from local estimates (figure 3). Correspondingly, species-wide  $\pi$  is predicted by factors that characterize the entire range, including its size and its stability through time (electronic supplementary material, figure S9). In contrast, none of the factors testing our three hypotheses significantly predict patterns of nucleotide diversity at the mitochondrial genome. This null result could be because we had reduced power—we were only able to infer mtDNA  $\pi$  for OTUs with multiple sampled individuals ( $N = 60$ )—and we find evidence for geographically restricted mtDNA introgression between *Ctenotus* OTUs, which would artificially inflate estimates of  $\pi$ . However, the unpredictability of mtDNA diversity also parallels other datasets [11–13], in which proxy estimates of census population size are uncorrelated with

mtDNA diversity. This might be because the mitochondrial genome, both because it functions as one giant-linked locus and because it contains so many genes with pivotal physiological roles [59], does not evolve neutrally, and therefore reflects different demographic and selective processes than the nuclear genome does as a whole. These results indicate that how we measure genetic diversity matters, and both the genomic marker used and the spatial scale considered can impact our understanding of both genetic diversity and its drivers.

### (d) Museum occurrences: a novel proxy for census size?

Given the current landscape of rapid global change, researchers are relying on museum databases to help document shifts in species abundances and distributions [48,60,61]. Historically, many have considered the use of museum databases as limited, partially because accounting for sampling bias and effort both across different species and throughout their ranges can be difficult or impossible [62]. In this work, we were also unable to account for sampling biases, so we might expect the number of museum occurrences to be a rather noisy measure of population sizes. Yet we still recovered a relationship between occurrence counts and genetic diversity, suggesting that museum records are a coarse proxy for census population size in this group. This correlation probably exists because *Ctenotus* are commonly sampled using general trapping methods like pitfall traps [63]. Although all field sampling methods for squamate reptiles show bias [64], we note that most *Ctenotus* taxa are ecologically and phenotypically similar in the broadest sense: for example, most are surface-active, heliothermic and trophically similar species. We predict that this similarity serves to reduce the effects of sampling bias on interspecific variation in occurrence counts in museum collections. In other groups—particularly those with greater ecological heterogeneity or those that span more complex geopolitical boundaries—occurrence data are potentially more biased and the relationship found here might be less likely to hold. We note that measuring census population size by standard ecological approaches (i.e. mark–recapture, aerial tracking) is time-consuming and expensive, and impossible for many rare or small taxa. In systems where sampling methods are less biased, and especially where data on species abundances are needed across many species [44,48], museum data might provide a rough proxy for relative population sizes.

## 5. Conclusion

Genetic diversity is a fundamental characteristic of species and the populations that comprise them. Although this work supports the basic population genetic prediction that census population size should positively correlate with genetic diversity, it is more notable that our analyses explained only a small fraction of the variation in diversity levels across this genus. Our results suggest that processes that explain variation in diversity across broad taxonomic scales tend to lack explanatory power at this narrow phylogenetic scale (figure 5), underlining the 'enduring riddle' that is genetic diversity [9].

**Data accessibility.** Code is available at [https://github.com/singhal/ct\\_gen\\_div](https://github.com/singhal/ct_gen_div). Code use, and a full version of these methods, are described in the electronic supplementary material. Raw sequence data are available at the NCBI BioProject: PRJNA382545. Species tree,



pseudo-reference genomes and variant data are available at doi:10.5061/dryad.kk73p.

**Authors' contributions.** S.S. conducted data analysis, designed the study and wrote the manuscript; H.H. carried out lab work and helped design the study; P.O.T. contributed analytical methods; S.C.D. contributed samples and laboratory support; I.H. carried out lab work; D.L.R. designed the study and wrote the manuscript. All authors gave final approval for publication.

**Competing interests.** We have no competing interests.

**Funding.** This work was supported by the University of Michigan, and by National Science Foundation (NSF) grant no. OSIE-0612855 and DEB-0814277. S.S. and I.H. are funded by NSF Postdoctoral Fellowship in Biology and Graduate Research Fellowship, respectively.

**Acknowledgements.** The authors thank the numerous museums and curators who provided access to tissues. They acknowledge useful feedback from the Rabosky Lab, the editor and three anonymous reviewers, and technical support from the staff of University of Michigan's Advanced Research Computing.

## References

- Hartl DL, Clark AG. 1997 *Principles of population genetics*. Sunderland, MA: Sinauer Associates.
- Hughes AR, Inouye BD, Johnson MT, Underwood N, Vellend M. 2008 Ecological consequences of genetic diversity. *Ecol. Lett.* **11**, 609–623. (doi:10.1111/j.1461-0248.2008.01179.x)
- Kimura M, Crow JF. 1964 The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.
- Lewontin RC, Hubby JL. 1966 A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54**, 595.
- Lewontin RC. 1974 *The genetic basis of evolutionary change*. New York, NY: Columbia University Press.
- Wright S. 1931 Evolution in Mendelian populations. *Genetics* **16**, 97–159.
- Hamrick JL, Godt MJW, Sherman-Broyles SL. 1992 Factors influencing levels of genetic diversity in woody plant species. In *Population genetics of forest trees* (eds WT Adams, SH Strauss, DL Copes, AR Griffin), pp. 95–124. Berlin, Germany: Springer.
- Frankham R. 1997 Do island populations have less genetic variation than mainland populations? *Heredity* **78**, 311–327. (doi:10.1038/hdy.1997.46)
- Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel 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)
- McCusker MR, Bentzen P. 2010 Positive relationships between genetic diversity and abundance in fishes. *Mol. Ecol.* **19**, 4852–4862. (doi:10.1111/j.1365-294X.2010.04822.x)
- 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)
- Nabholz B, Mauffrey J-F, Bazin E, Galtier N, Glémin S. 2008 Determination of mitochondrial genetic diversity in mammals. *Genetics* **178**, 351–361. (doi:10.1534/genetics.107.073346)
- Delrieu-Trottin E, Maynard J, Planes S. 2014 Endemic and widespread coral reef fishes have similar mitochondrial genetic diversity. *Proc. R. Soc. B* **281**, 20141068. (doi:10.1098/rspb.2014.1068)
- Perry GH *et al.* 2012 Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* **22**, 602–610. (doi:10.1101/gr.130468.111)
- James JE, Lanfear R, Eyre-Walker A. 2016 Molecular evolutionary consequences of island colonization. *Genome Biol. Evol.* **8**, 1876–1888. (doi:10.1093/gbe/evw120)
- Ellegren H, Galtier N. 2016 Determinants of genetic diversity. *Nat. Rev. Genet.* **17**, 422–433. (doi:10.1038/nrg.2016.58)
- Charlesworth B. 2009 Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205. (doi:10.1038/nrg2526)
- Romiguier J *et al.* 2014 Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**, 261–263. (doi:10.1038/nature13685)
- Duminil J, Fineschi S, Hampe A, Jordano P, Salvini D, Vendramin GG, Petit RJ. 2007 Can population genetic structure be predicted from life-history traits? *Am. Nat.* **169**, 662–672.
- Cogger H. 2014 *Reptiles and amphibians of Australia*. Clayton, Australia: CSIRO Publishing.
- Levene H. 1953 Genetic equilibrium when more than one ecological niche is available. *Am. Nat.* **87**, 331–333. (doi:10.1086/281792)
- Awise JC. 2012 *Molecular markers, natural history and evolution*. Berlin, Germany: Springer.
- Hewitt G. 2000 The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913. (doi:10.1038/35016000)
- Rabosky DL, Hutchinson MN, Donnellan SC, Talaba AL, Lovette IJ. 2014 Phylogenetic disassembly of species boundaries in a widespread group of Australian skinks (Scincidae: Ctenotus). *Mol. Phylogenet. Evol.* **77**, 71–82. (doi:10.1016/j.ympev.2014.03.026)
- How R, Cowan M. 2006 Collections in space and time: geographical patterning of native frogs, mammals and reptiles through a continental gradient. *Pacific Conserv. Biol.* **12**, 111–133. (doi:10.1071/PC060111)
- Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012 Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**, e37135. (doi:10.1371/journal.pone.0037135)
- Moritz C *et al.* 2016 Multilocus phylogeography reveals nested endemism in a gecko across the monsoonal tropics of Australia. *Mol. Ecol.* **25**, 1354–1366. (doi:10.1111/mec.13511)
- Oliver PM, Adams M, Lee MS, Hutchinson MN, Doughty P. 2009 Cryptic diversity in vertebrates: molecular data double estimates of species diversity in a radiation of Australian lizards (*Diplodactylus*, Gekkota). *Proc. R. Soc. B* **276**, 2001–2007. (doi:10.1098/rspb.2008.1881)
- Fujisawa T, Aswad A, Barracough TG. 2016 A rapid and scalable method for multilocus species delimitation using Bayesian model comparison and rooted triplets. *Syst. Biol.* **65**, 759–771. (doi:10.1093/sysbio/syw028)
- Fujisawa T, Barracough TG. 2013 Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent (GMYC) approach: a revised method and evaluation on simulated datasets. *Syst. Biol.* **62**, 707–724. (doi:10.1093/sysbio/syt033)
- Rabosky DL, Talaba AL, Donnellan SC, Lovette IJ. 2009 Molecular evidence for hybridization between two Australian desert skinks, *Ctenotus leonhardii* and *Ctenotus quattuordecimlineatus* (Scincidae: Squamata). *Mol. Phylogenet. Evol.* **53**, 368–377. (doi:10.1016/j.ympev.2009.06.020)
- 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)
- Smith SA, O'Meara BC. 2012 treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**, 2689–2690. (doi:10.1093/bioinformatics/bts492)
- Tajima F. 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.
- Wakeley J. 2009 *Coalescent theory: an introduction*. Englewood, CO: Roberts & Company.
- Begun DJ *et al.* 2007 Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5**, e310. (doi:10.1371/journal.pbio.0050310)
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016 VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584. (doi:10.7717/peerj.2584)
- Auwerwa GA *et al.* 2013 From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protocol Bioinformatics* **11**, 11.10.1–11.10.33. (doi:10.1002/0471250953.bi1110s43)
- Korneliusen TS, Albrechtsen A, Nielsen R. 2014 ANGSD: analysis of next generation sequencing

- data. *BMC Bioinformatics* **15**, 1. (doi:10.1186/s12859-014-0356-4)
40. Kuhner MK. 2006 LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**, 768–770. (doi:10.1093/bioinformatics/btk051)
  41. Alexander DH, Novembre J, Lange K. 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664. (doi:10.1101/gr.094052.109)
  42. Vachaspati P, Warnow T. 2015 ASTRID: accurate species trees from internode distances. *BMC Genomics* **16**, S3. (doi:10.1186/1471-2164-16-S10-S3)
  43. Rabosky DL, Donnellan SC, Grundler M, Lovette IJ. 2014 Analysis and visualization of complex macroevolutionary dynamics: an example from Australian scincid lizards. *Syst. Biol.* **63**, 610–627. (doi:10.1093/sysbio/syu025)
  44. Rabosky ARD, Cox CL, Rabosky DL, Title PO, Holmes IA, Feldman A, McGuire JA. 2016 Coral snakes predict the evolution of mimicry across New World snakes. *Nat. Commun.* **7**, 11484. (doi:10.1038/ncomms11484)
  45. Phillips SJ, Dudík M. 2008 Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. *Ecography* **31**, 161–175. (doi:10.1111/j.0906-7590.2008.5203.x)
  46. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005 Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978. (doi:10.1002/joc.1276)
  47. White EP, Ernest SM, Kerkhoff AJ, Enquist BJ. 2007 Relationships between body size and abundance in ecology. *Trends Ecol. Evol.* **22**, 323–330. (doi:10.1016/j.tree.2007.03.007)
  48. Boakes EH, McGowan PJ, Fuller RA, Chang-qing D, Clark NE, O'Connor K, Mace GM. 2010 Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol.* **8**, e1000385. (doi:10.1371/journal.pbio.1000385)
  49. Hijmans R, Cameron S, Parra J, Jones P, Jarvis A. 2005 *WorldClim, version 1.3*. Berkeley, CA: University of California.
  50. Pagel M. 1994 Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B* **255**, 37–45. (doi:10.1098/rspb.1994.0006)
  51. Kisel Y, Barraclough TG. 2010 Speciation has a spatial scale that depends on levels of gene flow. *Am. Nat.* **175**, 316–334. (doi:10.1086/650369)
  52. Burnham K, Anderson D. 1998 *Model selection and inference: a practical information-theoretic approach*. Berlin, Germany: Springer.
  53. Zhang G *et al.* 2014 Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320. (doi:10.1126/science.1251385)
  54. Graur D. 1985 Gene diversity in Hymenoptera. *Evolution* **39**, 190–199. (doi:10.2307/2408528)
  55. Roselius K, Stephan W, Städler T. 2005 The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics* **171**, 753–763. (doi:10.1534/genetics.105.043877)
  56. Frankham R. 2007 Effective population size/adult population size ratios in wildlife: a review. *Genet. Res.* **89**, 491–503. (doi:10.1017/S0016672308009695)
  57. Palstra FP, Fraser DJ. 2012 Effective/census population size ratio estimation: a compendium and appraisal. *Ecol. Evol.* **2**, 2357–2365. (doi:10.1002/ece3.329)
  58. Coop G. 2016 Does linked selection explain the narrow range of genetic diversity across species? *bioRxiv*, 042598.
  59. 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)
  60. Moritz C, Patton JL, Conroy CJ, Parra JL, White GC, Beissinger SR. 2008 Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. *Science* **322**, 261–264. (doi:10.1126/science.1163428)
  61. Pyke GH, Ehrlich PR. 2010 Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biol. Rev.* **85**, 247–266. (doi:10.1111/j.1469-185X.2009.00098.x)
  62. Guralnick R, Van Cleave J. 2005 Strengths and weaknesses of museum and national survey data sets for predicting regional species richness: comparative and combined approaches. *Divers. Distrib.* **11**, 349–359. (doi:10.1111/j.1366-9516.2005.00164.x)
  63. Read J. 1995 Subhabitat variability: a key to the high reptile diversity in chenopod shrublands. *Aust. J. Ecol.* **20**, 494–501. (doi:10.1111/j.1442-9993.1995.tb00568.x)
  64. McDiarmid RW, Foster MS, Guyer C, Gibbons JW, Chernoff N. 2012 *Reptile biodiversity: standard methods for inventory and monitoring*. Berkeley, CA: University of California Press.