Museum occurrence data predict genetic diversity in a species-rich clade of Australian lizards Supplementary Online Material

Sonal Singhal, Huateng Huang, Pascal O. Title,
Stephen C. Donnellan, Iris Holmes, Daniel L. Rabosky

March 9, 2017

Contents

1 Materials and Methods **2**
1.1 Sampling . . . 2
1.2 Library Preparation and Sequencing . . . 2
1.3 Testing Methods for ddRAD data assembly . . . 2
1.4 Species Delimitation . . . 3
1.5 Measures of Genetic Diversity . . . 4
1.5.1 Generating Pseudo-reference Genomes . . . 4
1.5.2 Within-population π . . . 4
1.5.3 Species-wide π . . . 5
1.5.4 mtDNA π . . . 5
1.5.5 Calculating diversity . . . 5
1.6 Demographic Analyses . . . 5
1.6.1 Running ADMIXTURE . . . 5
1.6.2 Running ANGSD . . . 5
1.6.3 Running LAMARC . . . 6
1.7 Species Tree . . . 6
1.8 Collecting data on possible drivers of genetic diversity . . . 7
1.8.1 Proxies for census population size . . . 7
1.8.2 Environmental hetereogeneity . . . 9
1.8.3 Historical demography . . . 9
1.8.4 Possible confounders . . . 9
1.9 Model-Testing . . . 10

2 Figures and Tables **10**
2.1 Tables . . . 10
2.2 Figures . . . 13

1 Materials and Methods

1.1 Sampling

This study takes advantage of the numerous tissue samples accessioned in natural history museums across the United States and Australia. In this study, we sampled tissues from 8 museums: Australian Museum, Cornell University Museum of Vertebrates, Australian Biological Tissue Collection, Northern Territory Museum, Queensland Museum, South Australian Museum, University of Michigan Museum of Zoology, and Western Australian Museum. Species boundaries in the genus *Ctenotus* have been subject to sufficient revision (1), and, like many squamate species, many *Ctenotus* species contain multiple, cryptic species. As such, our sampling strategy was to sample as broadly and densely across species ranges. As for many major geographic regions, sampling across Australia is non-random in space. For example, there are large chunks of the Australian arid zone that have been minimally sampled. Given these limitations, we tried to sample randomly throughout the range, maximizing our geographic coverage. That said, we acknowledge that biases certainly exist in such a dataset.

1.2 Library Preparation and Sequencing

We extracted DNA from liver and tail tissues using a Qiagen DNeasy kit and then measured DNA quantity and quality both using a ThermoFisher QuBit dsDNA HS Assay Kit (catalog number: Q32854) and a ThermoScientific NanoDrop spectrophotometer 2000. Anywhere from 200 ng to 1000 ng of DNA was used for preparation of ddRAD libraries. *In silico* experiments using the *Anolis carolinensis* genome as a reference showed that the enzymes EcoRI and MspI gave about 20K fragments in the 150 - 250 bp range. For the purposes of this experiment, these enzymes and this size range resulted in the right balance between the number of loci across the genome and projected coverage at any given locus. Thus, we digested DNA with the restriction enzymes EcoRI and MspI. Following digestion, we again quantified DNA using the QuBit. We then barcoded each library uniquely, pooled across 24 individuals, and size-selected (150 - 250 bp) these pools using a PippinPrep (2). Each pool was then amplified using Phusion HotStart Taq. Each amplification was done across 12 separate PCR reactions to reduce the amount of PCR bias that was introduced in any single reaction. We then pooled across 4 of the 24-individual pools to create a final pool of 96-individuals. The Center for Applied Genomics at the Hospital for Sick Children in Toronto sequenced each pool across a single 100 PE run of an Illumina HiSeq2500, for a total of six lanes.

1.3 Testing Methods for ddRAD data assembly

At the time of analysis, we were unaware of any papers that compared across ddRAD assembly methods to test which ddRAD assembly method performs best. As such, we tested three different methods for assembling ddRAD data to identify what method would best enable population genetic analyses. We tried (1) clustering the reverse read from read pairs, (2) Rainbow, and (3) pyRAD (3). For a set of individuals from a random selection of eight species (N=55), we measured assembly success by using five different metrics:

- the percent of reads used in the original assembly that could be mapped to the assembly, indicated as 'percentMapped',
- the mapping quality of these reads, abbreviated as 'MAQ',
- the percent of read-pairs that mapped as paired reads, indicated as 'percentPaired',
- the percent of read-pairs that only mapped to one location in the assembly, indicated as 'percentUnique',
- for every given locus in the assembly, the number of individuals in that species sequenced at that locus

Across all metrics, higher values reflect better assemblies, and our results suggest Rainbow generates the best assemblies (Fig. S16). Based on these results, we used Rainbow to assemble all individuals in our data set.

Scripts used in this analysis are:

- `https://github.com/singhal/ct_gen_div/blob/master/R2_assemblies.py`
- `https://github.com/singhal/ct_gen_div/blob/master/pyrad_assembly.py`
- `https://github.com/singhal/ct_gen_div/blob/master/rainbow_assemblies.py`
- `https://github.com/singhal/ct_gen_div/blob/master/evaluate_assembly.py`

1.4 Species Delimitation

Because species boundaries in *Ctenotus* have been subject to revisions and because many species contain multiple, morphologically-cryptic lineages, our first step was to delimit putative species-level lineages – or, operational taxonomic units (OTUs). We used a coalescent-based approach (GMYC) that requires an ultrametric phylogenetic tree of all the individuals that are to be delimited.

To do so, we first trimmed the ddRAD raw reads for both low-quality sequence and adapter sequencing using Trimmomatic v0.33 (4). The script used was: `https://github.com/singhal/ct_gen_div/blob/master/trimmomatic_reads.py`. Here, we trimmed the ends of reads once the average base quality (as measured in 4bp sliding windows) dropped below a phred33 score of 25. We then discarded any reads that were less than 36bp long. With these trimmed reads, we generated one assembly per individual using the program Rainbow using default parameters (5). Then, we identified homologous ddRAD loci across all individuals. We used vsearch v1.11.1 (6), requiring all putative homologs to share at least 80% similarity. The script used here was: `https://github.com/singhal/ct_gen_div/blob/master/homology_across_species.py`. We then restricted our analyses to only those loci found in at least 60% of individuals. We aligned each locus individually using MUSCLE v3.8.31 (7) and generated a concatenated alignment across loci. To test if the method used to infer the phylogeny impacted species delimitation, we used two methods to infer a phylogeny: RAxML v8.2.0 and FastTree v2.1.8 (8; 9).

To generate these trees, we used the following commands:

```
/Volumes/heloderma4/sonal/bin/RAxML/raxmlHPC -x 123 -# 100 -p 543 -m GTRGAMMA -f a -n
    Ctenotus_bootstrap -s Ctenotus_nuclear_percent0.6.aln.phy
/Volumes/heloderma4/sonal/bin/FastTree -gtr -nt Ctenotus_nuclear_percent0.6.aln.fa >
    Ctenotus_nuclear_percent0.6.FastTree
```

We then used midpoint rooting to root each tree as implemented by the function midpoint in the R package phangorn (10). Then, to generate an ultrametric tree as required by GMYC, we used TreePL. We constrained the root age for the phylogeny to be between 15 and 20 million years based on data from other phylogenetic analyses of *Ctenotus* (11). We ran TreePL for a range of λ values (0.001, 0.01, 0.1, 1), finding that λ=0.1 was the best fit for both trees based on coefficient of variation (CV) results. Then, to define OTUs within these species, we used a Python-based implementation of GMYC on both the FastTree and RAxML inferred trees. This implementation of GMYC runs the single-threshold model, which assumes that all speciation events precede coalescent events. This model tends to outperform the multiple-threshold model (12). The nodes inferred as the most recent common ancestor for a given OTU were well-supported in this RAxML tree (Fig. S3A), and trees inferred using FastTree and RAxML resulted in similar OTU delimitations (Fig. S3B). Based on these results, we used the GMYC-inferred OTUs from RAxML for all subsequent analysis.

1.5 Measures of Genetic Diversity

The goal of this study is to understand the drivers of genetic diversity. To do so, we calculated genetic diversity as average pairwise distances (π) three ways:

1. within-population π: we treated each individual as a population. Because we sampled an average of 44K loci per individual, estimating π per individual should yield an accurate estimate of population-level π. We called SNPs for each individual separately. We then calculated π for each individual and took the mean π across all individuals in an OTU as our within-population estimate of π.

2. species-wide π: we called SNPs for all individuals in an OTU together. We then calculated π across all individuals in an OTU using an approach that allows for different levels of missingness across loci (13).

3. mtDNA π: using a previously published mitochondrial sequence alignment from the locus cytochrome B (1; 11), we inferred π across all mitochondrial sequences sampled for a given OTU

Scripts used in this analysis are:

- `https://github.com/singhal/ct_gen_div/blob/master/align_reads1.py`
- `https://github.com/singhal/ct_gen_div/blob/master/align_reads2.py`
- `https://github.com/singhal/ct_gen_div/blob/master/get_vcf_depth_by_ind.py`
- `https://github.com/singhal/ct_gen_div/blob/master/get_vcf_depth.py`

1.5.1 Generating Pseudo-reference Genomes

For each OTU, we created a pseudo-reference genome based on the Rainbow assemblies for each individual in that OTU. To do so, we used vsearch to cluster contigs across individuals, requiring clustered sequences to have 97% similarity. We only retained those locus clusters that were well-represented across individuals to avoid retaining spurious loci; for OTUs sampling two or fewer individuals, we kept all clusters, for OTUs sampling three to five individuals, we kept all clusters for which representative sequences were found in at least two individuals, and for all remaining OTUs, we kept all clusters for which representative sequences were found in at least 30% individuals. This approach was implemented in `https://github.com/singhal/ct_gen_div/blob/master/simpler_homology.py`.

1.5.2 Within-population π

These analyses were done at the individual-level. We aligned reads from each individual to its respective pseudo-reference genome to identify variants. To align reads, we used bwa v0.7.12 (14), fixed mate-pairs using samtools v1.2 (15), and identified indels and realigned around indels using GATK v3.4 (16). We then called raw SNPs using samtools and bcftools v1.2, filtering out SNPs that had quality score lower than 20, and then using the resulting variant set to recalibrate base quality scores in the alignment files with GATK. Using the recalibrated alignment files, we once again called genotypes, retaining only those invariable and variable sites that had $\geq 10\times$ coverage to ensure we could accurately call heterozygous sites. Additionally, to avoid including collapsed paralogs, we removed sites with $\geq 3\times$ the median coverage for the individual. After this step, we dropped four individuals because less than 1×10^5 of their genomic sites had sufficient coverage to call SNPs. We then calculated π (see 1.5.5) for each individual and found the mean π across individuals within an OTU.

1.5.3 Species-wide π

These analyses were done across all individuals in an OTU at once. We aligned reads from each individual to the pseudo-reference genome to identify variants. To align reads, we used bwa, fixed mate-pairs using samtools, and identified indels and realigned around indels using GATK. We then called raw SNPs across all individuals in that OTU using samtools and bcftools, filtering out SNPs that had quality score lower than 20. We used the resulting variant set to recalibrate base quality scores in the alignment files with GATK. Using the recalibrated alignment files, we called genotypes across all individuals in an OTU simultaneously. In the resultant variant file, we set to 'missing' any site in any individual that had $<10\times$ coverage. We then calculated π (see 1.5.5) across all individuals at once to get species-wide π.

1.5.4 mtDNA π

From the previously published mtDNA alignment, we only used those individuals for which we had a revised OTU designation – *i.e.*, those individuals that were successfully sequenced for ddRAD data. We were only able to analyze those OTU for which we had at least two individuals sampled; this reduced our sampling to 60 OTU. We then calculated π (see 1.5.5) across all individuals at once to get mtDNA π.

1.5.5 Calculating diversity

We calculated three indices of nucleotide diversity: within-population π, species-wide π (13), and mtDNA π. For these metrics, we used a raw, uncorrected measure for diversity. For nuclear estimates of π, π was calculated per SNP and then averaged.

Scripts used in this analysis are:

- `https://github.com/singhal/ct_gen_div/blob/master/calculate_pi_per_ind_by_ind.py`
- `https://github.com/singhal/ct_gen_div/blob/master/calculate_pi_per_species.py`

1.6 Demographic Analyses

1.6.1 Running ADMIXTURE

For the 42 OTUs for which we sequenced $\geq$3 individuals, we used the program ADMIXTURE v1.23 to determine if there was genetic clustering of individuals within OTUs. We used the variant set called across all individuals, restricting our analyses to only those SNPs that had $<$33% missing data and then selecting only one SNP per contig as ADMIXTURE cannot explicitly model linkage disequilibrium across SNPs. We then ran ADMIXTURE for a range of cluster values (k=1 to n), where n is the number of individuals sampled for that OTU. We picked the most likely number of clusters by using coefficient of variation (CV) values. We further visualized our results because occasionally ADMIXTURE will infer a cluster to which no individual is assigned. In such cases, we adjusted the total number of clusters to reflect only those clusters to which individuals were assigned.

Scripts used in this analysis are:

- `https://github.com/singhal/ct_gen_div/blob/master/vcf_to_admixture.py`
- `https://github.com/singhal/ct_gen_div/blob/master/run_admixture.py`
- `https://github.com/singhal/ct_gen_div/blob/master/run_admixture_chooseK.py`

1.6.2 Running ANGSD

To infer Tajima's D for these OTUs, we needed to employ a method that could account for heterogeneous and low coverage across individuals and sites, thus we used ANGSD v0.910. Further, without complete genome sequences, inferring demography with just one or two individuals is difficult. Thus, we ran ANGSD

on only the 42 OTUs for which we sequenced $\geq$3 individuals. We randomly chose these three individuals from the most common genetic cluster as identified using ADMIXTURE. We ran ANGSD on subsamples of the total sample set so that our comparisons across species would be based on matched sample sizes. Because initial tests suggested that the number of individuals used during SNP calling positively biased the number of SNPs found (Fig. S2), we repeated our SNP discovery pipeline (see *Calculating species-wide* π) on these subsampled data. This approach ensured that our results were not biased across OTUs due to uneven sampling. Exemplar commands for ANGSD were the following:

```
angsd -bam list_of_bam_files -doSaf 1 -anc lineage_PRG.fa -GL 2 -P 12 -out lineage_out
    -fold 1
misc/realSFS lineage_out.saf.idx -P 12 > lineage_out.sfs
angsd -bam bam.list_of_bam_files -out outFold -doThetas 1 -doSaf 1 -pest lineage_out.sfs
    -anc lineage_PRG.fa -GL 2 -fold 1
misc/thetaStat make_bed lineage_out.thetas.gz
misc/thetaStat do_stat lineage_out.thetas.gz -nChr num_of_chr_in_lineage
```

1.6.3 Running LAMARC

To infer population growth for these OTUs, we used LAMARC on SNP data from any OTUs with three or more individuals. Again, we subsampled the data to only consider three individuals per OTUs, using the same sampling used in the ANGSD analyses. We used the revised SNP calls that were done with just 3 individuals per OTUs.

We sampled all SNPs that were more than 70% complete and ran LAMARC in the Bayesian mode using the F84 mutational model. We adjusted the prior for growth rate (g) to range from [-1000, 10000]; all other priors were kept as default. We ran 10 chains with 10,000 burn-in iterations followed by 2 chains of 5e6 iterations. Convergence was checked by visually inspecting the output in Tracer (17). The script used in this analysis is: `https://github.com/singhal/ct_gen_div/blob/master/vcf_to_lamarc.py`.

1.7 Species Tree

To account for phylogenetic structure in the relationship between predictor variables and genetic diversity, we first constructed a species tree for these OTUs. To do so, we did the following:

1. We identified homologous loci across the OTU pseudo-reference genomes using VSEARCH, restricting matches to those with $\geq$ 80% shared identity or higher. This was implemented in the script `https://github.com/singhal/ct_gen_div/blob/master/homology_across_species.py`. This resulted in 18,816 loci.

2. We then aligned homologous loci across OTUs using MUSCLE v3.8.31 (7). We used the script `https://github.com/singhal/ct_gen_div/blob/master/make_nuclear_alignment.py`

    - Some alignments consisted of too few individuals or were quite poor (i.e., as much as 70% of the alignment consisted of gaps). We retained only those loci that consisted of $\geq$30% of the individuals sampled and consisted of $\leq$30% gaps. In total, we retained 14,187 (75%) of the original loci.

3. We then inferred gene trees for each locus using RAXML and this script: `https://github.com/singhal/ct_gen_div/blob/master/gene_trees.py`. The trees were run unpartitioned using the GTRCAT model of molecular evolution.

    - Poorly resolved nodes were converted to polytomies. ASTRID assumes gene trees have been inferred accurately and that all nodes are bifurcating. So that we did not give undue confidence to unresolved or poorly resolved nodes, we changed short bifurcating nodes to polytomies using

the script: https://github.com/singhal/ct_gen_div/blob/master/gene_trees_polytomize.py. This used the function di2multi in the R package ape.

4. We then used ASTRID to infer a species tree based on our inferred gene trees. This used the script: https://github.com/singhal/ct_gen_div/blob/master/make_astrid_astral.py. We used the 'bionj' method, allowing for missing data. We ran ASTRID for 100 bootstraps using a gene-tree bootstrapping approach, in which each bootstrap consisted of a set of gene trees drawn from the bootstraps of the gene tree inference. Trees were summarized using sumtrees from DendroPy v4.1.0 (18).

```
sumtrees.py BOOTTREES --unrooted -t TARGETTREE -o OUT --to-newick
    --no-summary-metadata
```

5. We generated a concatenated alignment for the 14,187 loci for which we inferred gene trees. With these alignments, we used RAxML to infer branch lengths for the best ASTRID tree along with all bootstraps. We then summarized the patterns across all bootstraps, both in terms of support and median branch length, to the best ASTRID tree using sumtrees in DendroPy.

```
raxmlHPC-PTHREADS -f e -t ASTRID.tre -m GTRGAMMAI -s concatenated_alignment.fa -n
    Ctenotus
sumtrees.py --edges median-length -t ASTRID_best.tre -F newick -l support --unrooted
    --suppress-annotations -o ASTRID_best_summary.tre ASTRID_boot_brlens.trees
```

6. To get an ultrametric and rooted tree, we used the script: https://github.com/singhal/ct_gen_div/blob/master/date_and_plot_species_trees.R. We rooted the tree using data from other multilocus phylogenies of *Ctenotus* and related genera (11). We could do so because the deep splits in the these trees' topologies were similar. We were unable to root this tree using outgroups because there were too few homologous ddRAD loci between *Ctenotus* and its nearest outgroup, *Lerista*.

   For this rooted tree, we used penalized likelihood to infer a chronogram for the ASTRID topology as implemented in the function chronos in the R package ape (19). To determine the appropriate λ, we tried a series of values: 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1. λ was chosen by comparing log-likelihood across different values. The best lambda was 1e-5.

1.8 Collecting data on possible drivers of genetic diversity

We have three hypothese that are not mutually-exclusive that might explain variation in genetic diversity across species. The first is that genetic diversity reflects census population size, or N_c. The second is that environmental heterogeneity increases genetic diversity. The final is that historical demographic shifts impact genetic diversity.

To test these hypotheses, we collated data on 11 independent ecological, geographical, and morphological variables. These are:

1.8.1 Proxies for census population size

1.8.1.1 Morphology

Morphology, particularly body size, is often used as a proxy for census population size, because bigger animals are believed to be less abundant (20). We used previously published data on eight different morphological traits to characterize patterns of morphology (11). First, we took the average morphological measurement for a species. On average, 5.87 individuals were measured per nominal species. Then, we log-transformed these measurements to get normal distributions across species. We then summarized patterns across these eight traits by doing a scaled principal component analysis using the R function prcomp. We

retained the first two axes; together, these axes explain 98% of the variation. The first axis largely describes size, and the second axis describes shape variation.

1.8.1.2 Range size

Range size is often used as a proxy for census population size; species with bigger ranges are assumed to have larger populations (21). To infer geographic ranges for each OTU, we used an approach that combines data from biodiversity databases with ecological niche modeling. This approach is particularly useful for species for which ranges are not well-delimited in field guides, such as many Australian lizards. We implemented this approach in a script that will be included in a manuscript in preparation (Title *et al.*, in prep.). As such, the script will be embargoed on DataDryad until the manuscript is published.

Occurrence data were accessed from Atlas of Living Australia, VertNET and GBIF data portals on March 2015, and then data were filtered to remove duplicates and as described previously (22). For OTUs that were synonymous with a nominal species, we followed a multiple-step procedure to construct ranges. First, occurrence points for that species were used to construct an alpha-hull polygon. Note that all the species used in this study had at least 3 or more occurrence points. Then, we defined an ecological niche model (ENM) using MaxEnt v3.3.3k (23). We first built a model that used 22 environmental variables, which included all 19 bioclimatic variables and elevation from WorldClim v1.4 at 5 arc-minute resolution, actual evapotranspiration, and aridity index. The actual evapotransportation index was from (24); the aridity index was from (25). We only used occurrence data to build these models; we did not include observed absences. To avoid over-fitting, we then thinned this model to include only those environmental variables that were significant above 5% after permutation tests. We then identified the minimum presence threshold as the value that would allow us to retain the top 99% of occurrences as present. Using this minimum presence threshold, we then projected the ENM as presence-absence by applying a minimum presence threshold. Finally, we defined the range as the intersection between the alpha-hull polygon and the ENM-defined habitat. If necessary, we further clipped this range to the coastal limits of Australia. These analyses used the R packages `alphahull`, `raster`, and `rgeos` (26; 27; 28).

For OTUs that combined across nominal species, we combined the nominal species ranges to define the range. For OTUs that were split from nominal species, we followed the same procedure as above to create the nominal species range. Then, we used the genetic sampling points for each cryptic OTU to construct individual ENMs. We then subdivided the original nominal range by assigning to each cryptic OTU the portion of the range in which it had the highest probability for occurrence based on its individual ENM. This approach assumes that ranges for cryptic species complexes do not overlap. We believe this is a fair assumption for three reasons. First, we sample densely throughout many of these ranges and across presumed boundaries between cryptic lineages. Through this sampling, we have yet to identify any evidence for syntopy among lineages in a cryptic complex. Second, we recover evidence for limited mtDNA-nDNA discordance at presumed boundaries between cryptic lineages in Ctenotus, suggestive of introgression at parapatric borders ([(1), Singhal et al, unpublished). Third, data from other lizard systems on cryptic diversity also typically recover narrowly parapatric boundaries between cryptic lineages (29; 30; 31). For OTUs that reflected more complex relationships to nominal species, we followed the same approach, instead defining ranges using our genetic sampling points as our occurrence points.

1.8.1.3 Number of museum occurrences

Number of museum occurrences could be a proxy for census population sizes, though some have suggested that this relationship is likely to be weak given that sampling is generally biased (32). We test this here. To count the number of database occurrences for each OTU, we used the script in `https://github.com/singhal/ct_gen_div/blob/master/species_counts.R`.

1.8.2 Environmental hetereogeneity

1.8.2.1 Elevational range

Habitat heterogeneity might lead to different alleles being favored in different habitats, and even with low migration, this can lead to increased genetic diversity (33). One measure of habitat heterogeneity is elevational range. To calculate the elevational range encompassed by a range, we extracted elevational data for a range using the R package raster and the altitude data available from Bioclim (27; 34). We took the range of elevation seen across the range as our metric. This is implemented in the script: `https://github.com/singhal/ct_gen_div/blob/master/heterogeneity.R`.

1.8.2.2 Climate heterogeneity

As with elevational range, climate heterogeneity can be a proxy for habitat heterogeneity. We summarized climate heterogeneity using the script `https://github.com/singhal/ct_gen_div/blob/master/heterogeneity.R`. We used the range in values for the first three principal component axes (PC1, PC2, PC3) describing climatic heterogeneity in our original analyses. Together, these three axes explained 79% of the variation. The first axis heavily loads on measures of precipitation, the second on measures of temperature, and the third on measures of elevation and slope. For our final analyses, we dropped PC2 and PC3 because they were highly correlated with elevational range. This approach used the R package raster (27).

1.8.3 Historical demography

1.8.3.1 Average historical stability

History is expected to impact genetic diversity; in particular, demographic shifts through time can lead to effective and census population sizes becoming uncorrelated (35). To infer the effects of changing demography, we inferred stability of each OTU's geographic range through time.

We first prepared the paleomodels for use in modeling, using the script: `https://github.com/singhal/ct_gen_div/blob/master/prep_paleomodel_layers.R`. This script clips the global layers to Australia to increase the speed of analyses. It further aggregates the last interglacial data from 30 arc-seconds to 2.5 minutes resolution. We avoid using this narrower resolution because it is artificially precise given the challenges of inferring historical climates. We then used the script `https://github.com/singhal/ct_gen_div/blob/master/habitat_stability.R` to infer average historical stability through time.

1.8.4 Possible confounders

1.8.4.1 Time in tree

We included the branch length for a OTU in the species tree as a possible confounding variable. To do so, we calculated the branch length from each OTU tip to its most-recent common ancestor using the R package ape (19). The tree we used here was our species tree of all 83 OTUs (Fig. S15).

1.8.4.2 Latitudinal midpoint

We included the latitudinal midpoint at which a OTU is found as a possible confounding variable. Using the inferred geographic ranges, we found the latitudinal midpoint using the function gCentroid implemented in the R package rgeos.

1.8.4.3 Biomes

For some analyses, we included the biome(s) in which an OTU was found as a confounding variable. To get these data, we accessed biome data from The Nature Conservancy's Terrestrial Ecoregions of the World, available at `http://maps.tnc.org/gis_data.html`, on 14 December 2015. We then determined which

biomes a range overlaps by using the script: `https://github.com/singhal/ct_gen_div/blob/master/biomes_pi.R`

1.9 Model-Testing

Organization of the final data set and model testing is implemented in `https://github.com/singhal/ct_gen_div/blob/master/final.R`. We took the natural log of two variables – range size and number of museum occurrences – because profiling the data suggested these two variables had non-normal distributions. Further, because we were testing a full array of additive models, in which every factor was included in every combination, we needed to ensure that the number of data points included across models was constant. As such, we dropped any rows with missing data; most OTUs were dropped due to missing morphology data. This lead to a final data set of 74 OTUs.

To determine which factors significantly predicted genetic diversity, we used an approach that calculates the relative importance of a factor in predicting genetic diversity across all possible additive models. For a given set of factors, all possible additive models are tested in their ability to predict genetic diversity. Each additive model was run as a phylogenetic linear model using the R package phylolm (36). For each additive model, Akaike weights are calculated. To calculate Akaike weights, we followed the process outlined (37). First, we calculated corrected AIC (AICc) scores for each model. Then, we identified the best scoring model (*best*) and calculated the raw Akaike weight by using the formula $e^{\frac{best-model}{2}}$. The final Akaike weights were then calculated by dividing the raw weights by the sum of all the raw weights. The relative importance for a given factor is then calculated by summing the relative Akaike weights for all the additive models in which it appears.

2 Figures and Tables

2.1 Tables

Table S1: Data on the individual samples used in this study, including the sample names, the OTU to which they were assigned, the nominal species to which they were identified, their latitude and longitude, the number of sites that had $>10\times$ coverage (denominator), the number of reads sequenced (original reads), the number of reads aligned to the pseudo-reference genome (aligned reads), and the median coverage at sites for which genotypes were called (i.e., those with $>10\times$ coverage). Many of these tissues are accessioned: AMSR (Australian Museum), CUMV (Cornell University Museum of Vertebrates), ABTC (Australian Biological Tissue Collection), NTMR (Northern Territory Museum), QM (Queensland Museum), SAM (South Australian Museum), UMMZ (University of Michigan Museum of Zoology), and WAM (Western Australian Museum). These data are also available at `https://github.com/singhal/ct_gen_div/blob/master/TableS1_individual_data.csv`.

variable	lambda	p-value
within-population π	6.78e-05	1
mtDNA π	6.78e-05	1
species-wide π	6.78e-05	1
range size	6.78e-05	1
elev. range	0.253	1
PC1 range, climate	4.62e-05	1
avg. hist. stability	6.78e-05	1
number of occurrences	0.511	1
PC1, morphology	1.03	1.52e-15
PC2, morphology	0.896	0.000376
lat. midpoint	0.894	1.04e-08
time in tree	1.03	1.27e-36

Table S2: Dependent and independent variables used in the model and their phylogenetic signal (λ) and its significance. None of the dependent variables measuring genetic diversity showed signal.

formula	λ	AIC	Akaike weight	adj. r^2
within-population $\pi \sim$ ln(number of occurrences)	1.89E-08	-868.87	0.0338	0.15
within-population $\pi \sim$ ln(range size) + ln(number of occurrences)	1.96E-08	-869.07	0.0374	0.17
within-population $\pi \sim$ ln(number of occurrences) + PC1, morphology	2.07E-08	-867.83	0.0201	0.175
within-population $\pi \sim$ ln(number of occurrences) + PC2, morphology	1.96E-08	-867.46	0.0167	0.17
within-population $\pi \sim$ ln(range size) + lat. midpoint + ln(number of occurrences)	2.06E-08	-867.78	0.0196	0.166
within-population $\pi \sim$ ln(range size) + ln(number of occurrences) + PC1, morphology	2.27E-08	-868.83	0.033	0.201

Table S3: The models in the top 1% of AIC weights of all models fit, showing model's formula, the phylogenetic signal (λ) inferred for the model's covariance, the AIC, and the AIC model weight of the model. Because all models showed low levels of phylogenetic signal, also shown are the adjusted r^2 terms for a non-phylogenetic linear model fit with the same formula. All top-scoring models include the factor with the highest relative importance – number of museum occurrences. This factor is the only significant factor identified across the entire model selection process.

nominal species	clutch size	citation
Ctenotus ariadnae	4	(38)
Ctenotus atlas	1.5	(38)
Ctenotus brooksi	2.2	(39)
Ctenotus calurus	2.7	(38)
Ctenotus colletti	2	(38)
Ctenotus essingtoni	2.9	(40)
Ctenotus gagadju	2	(40)
Ctenotus helenae	4.3	(39)
Ctenotus helenae	3.5	(38)
Ctenotus labillardieri	4.1	(41)
Ctenotus lancelini	3	(42)
Ctenotus leae	3.7	(38)
Ctenotus leonhardii	3.3	(39)
Ctenotus leonhardii	5.8	(38)
Ctenotus leonhardii	2	(43)
Ctenotus pantherinus	6.6	(39)
Ctenotus pantherinus	6.1	(38)
Ctenotus pantherinus	6	(44)
Ctenotus piankai	2.8	(39)
Ctenotus quattuordecimlineatus	3.1	(39)
Ctenotus regius	2.04	(43)
Ctenotus rhomboidalis	2	(45)
Ctenotus robustus	5.4	(40)
Ctenotus robustus	5.53	(40)
Ctenotus robustus	2	(45)
Ctenotus saxatilis	4	(40)
Ctenotus schomburgkii	2.1	(39)
Ctenotus schomburgkii	3	(38)
Ctenotus schomburgkii	1.96	(43)
Ctenotus strauchii	2.2	(43)
Ctenotus taeniolatus	4	(40)
Ctenotus taeniolatus	4.6	(40)
Ctenotus taeniolatus	3.7	(45)
Ctenotus uber	2	(43)

Table S4: Clutch sizes for nominal species of lizards, as reported in the literature. Some species have multiple measurements, reflecting values estimated from different localities and / or different studies. Clutch size, which is one measure of life history variation, shows a trend toward being correlated with genetic diversity (Fig. S10).

Table S5: Survey of studies that have compared genetic diversity across multiple species. These studies were found opportunistically throughout this study. Additional studies were found by reviewing the citations for each study and citing papers. We only included those studies that measured genetic diversity across 5 or more species and explicitly identified a significant factor explaining variation in genetic diversity. These studies span a wide variety of taxa and use many types of genetic markers (i.e., allozymes, organellar genomes, microsatellites).

For each study, we have included the number of species studied, the crown age of the species studied (as inferred from timetree.org), a rough sense of the phylogenetic scale, which major hypothesis the study is testing, the factors identified to be significant predictors of genetic diversity, and how much variation was explained. We also include details on what type of genetic markers were used and what measure of genetic diversity was included. For studies reporting a correlation coefficient (r) between the predictor varaiable and genetic diversity, proportion of variance is shown as r^2. These results show that most studies are at a broad phylogenetic scale, and the average study explains a relatively small (<40%) of the variation in genetic diversity. A visual summary of these data is available at Fig. 5. These data are also available at `https://github.com/singhal/ct_gen_div/blob/master/TableS5_XXX.csv`.

2.2 Figures

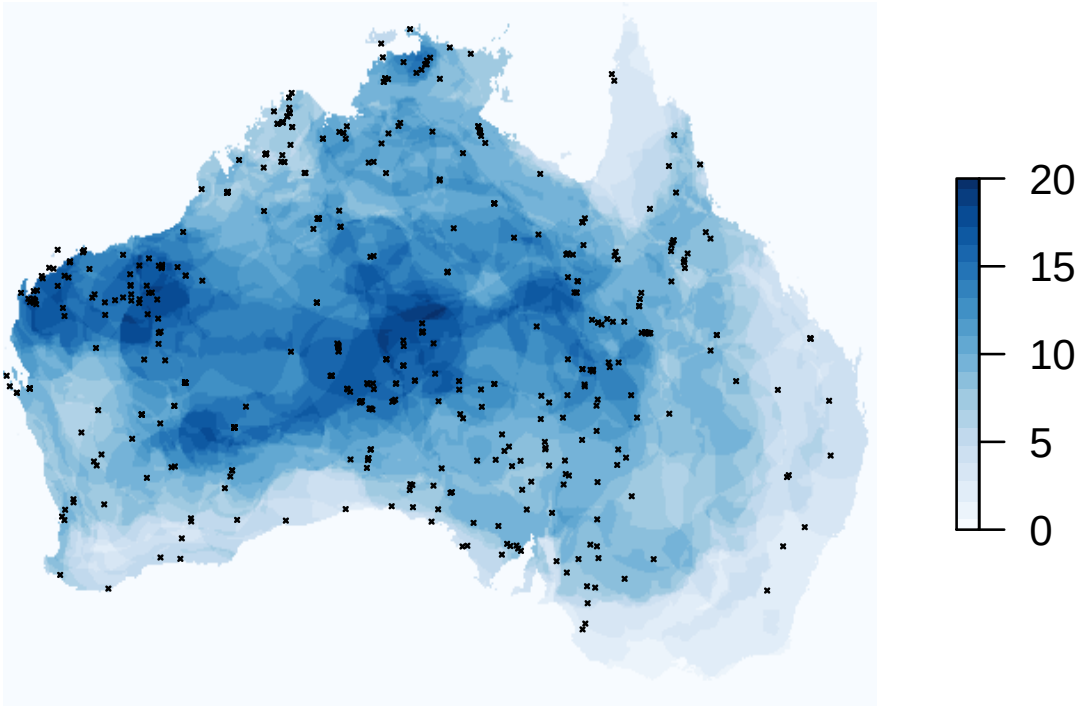

Figure S1: Map of samples (N=575) used in this study. The map reflects the alpha diversity of nominal *Ctenotus* species across Australia.

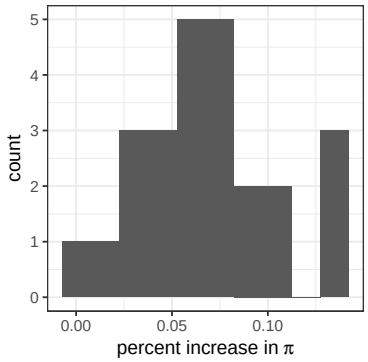

Figure S2: Changes in nucleotide diversity (π) due to technical artifacts of SNP calling. For five individuals from each of five randomly-selected OTUs, we called SNPs for each individual separately and for all five individuals at once. We then estimated π per individual for each SNP call set. Shown is the percent increase in π for a given individual for the five-individual SNP call set over the one-individual call set. On average, π increased by 6.8% between the two sets. These results suggest that calling SNPs for unbalanced numbers of individuals can introduce technical artifacts.

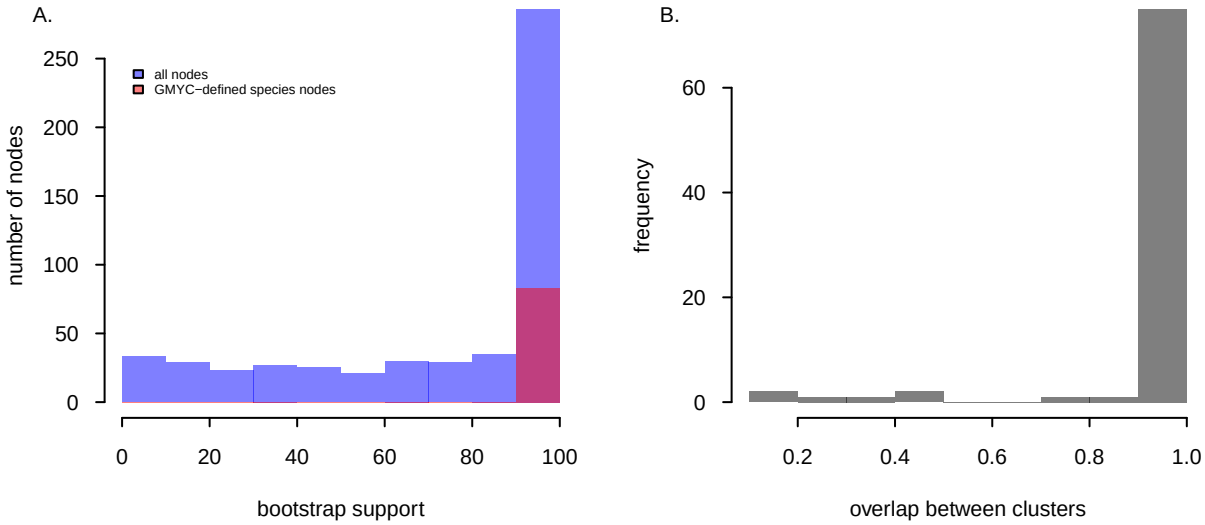

Figure S3: (A) Bootstrap support, as inferred by RAxML, for all nodes in the phylogeny used to delimit OTUs and for GMYC-defined species nodes. The nodes that define OTUs tend to be well-supported. (B) Overlap between OTU identities as determined by GMYC when applied to phylogenies inferred by two methods, FastTree and RAxML. In general, species delimitation across the two phylogenies is similar.

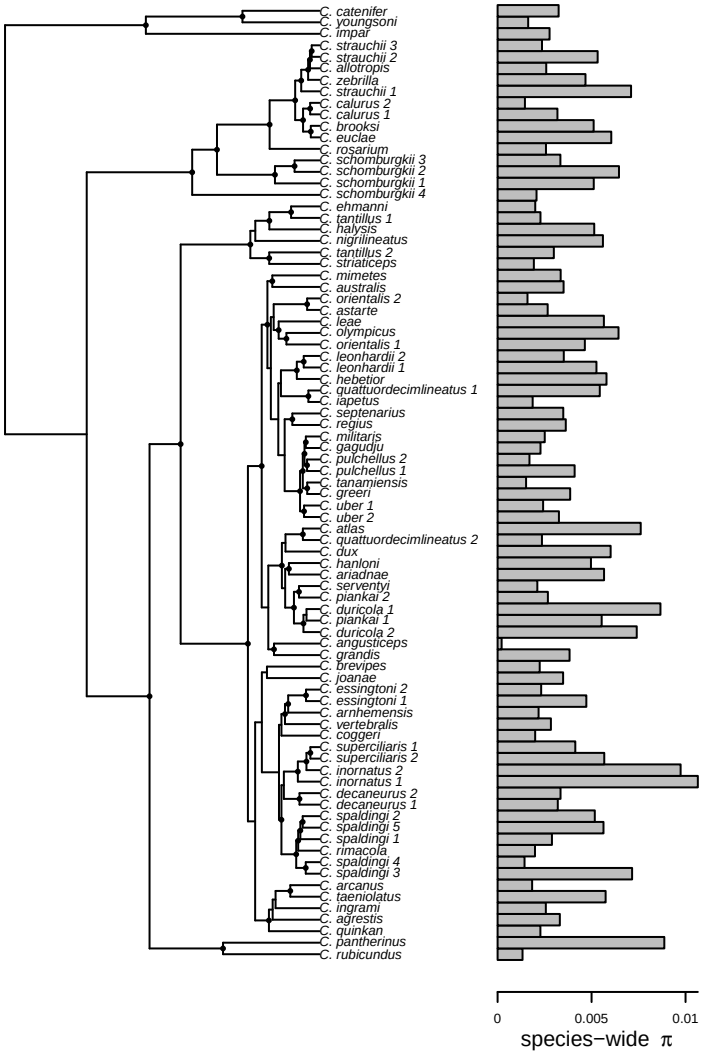

Figure S4: The "species tree" for *Ctenotus*, as inferred by ASTRID, shown with values of species-wide estimates of nucleotide diversity (π). Closely-related OTUs show significant variation in levels of genetic diversity. Nodes labeled with circles have bootstrap support >95%. Fig. S15 depicts the species tree with full bootstrap support.

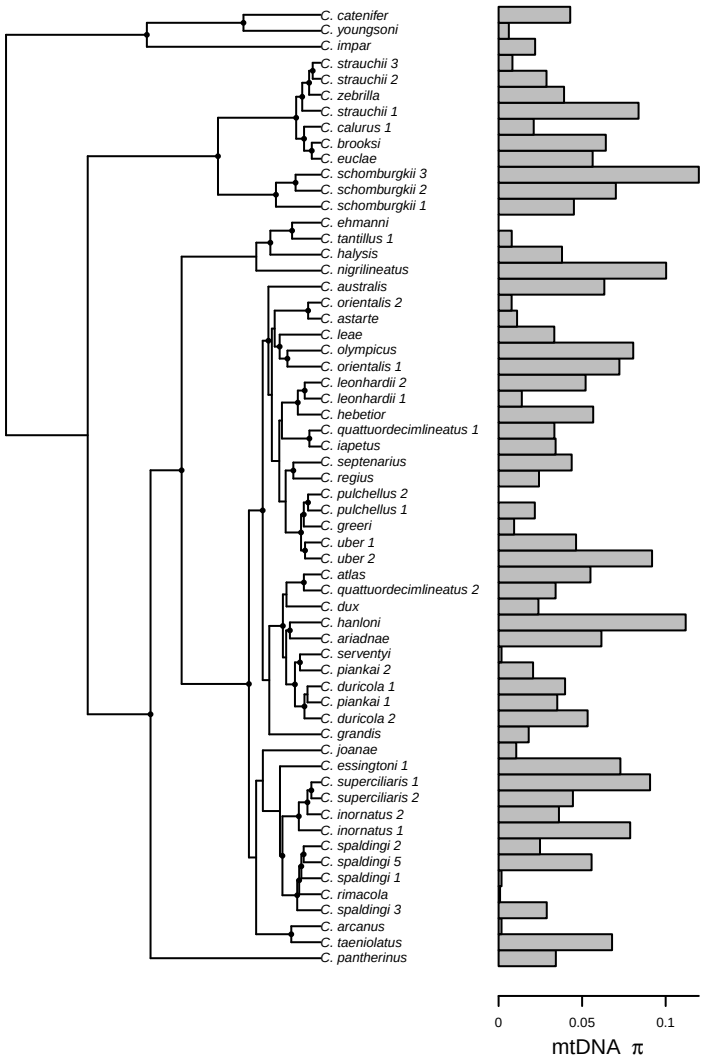

Figure S5: The "species tree" for *Ctenotus*, as inferred by ASTRID, shown with values of species-wide estimates of mitochondrial DNA nucleotide diversity (π). Closely-related OTUs show significant variation in levels of genetic diversity. Nodes labeled with circles have bootstrap support >95%. Fig. S15 depicts the species tree with full bootstrap support. Note fewer OTUs are shown than depicted in Fig. S15. We could only infer mtDNA for 60 OTUs; all other OTUs sampled two or more individuals for mtDNA.

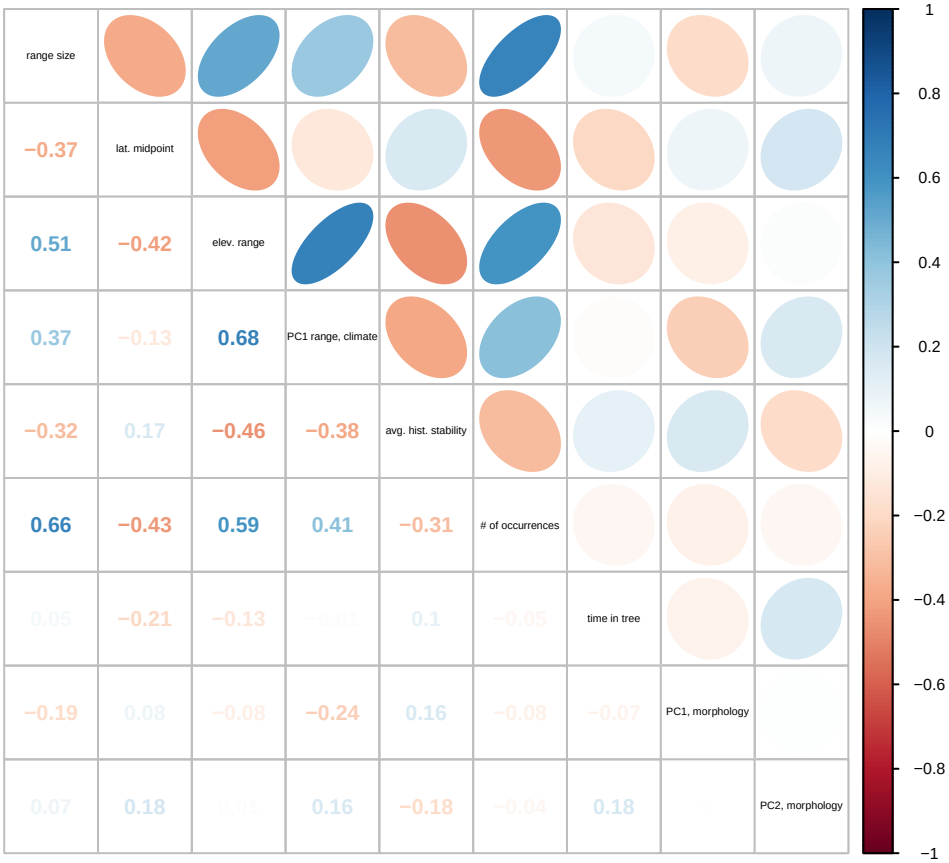

Figure S6: Spearman correlations between the nine independent variables (range size, latitudinal midpoint of the range, elevational span of the range, PC1 span in climate across the range, average historical stability of the range, number of museum occurrences, terminal branch length, and PC1 and PC2 in morphology) retained in the final model-fitting procedure. All variables were measured for for each OTU. These nine variables test the three non-exclusive hypotheses that might predict genetic diversity.

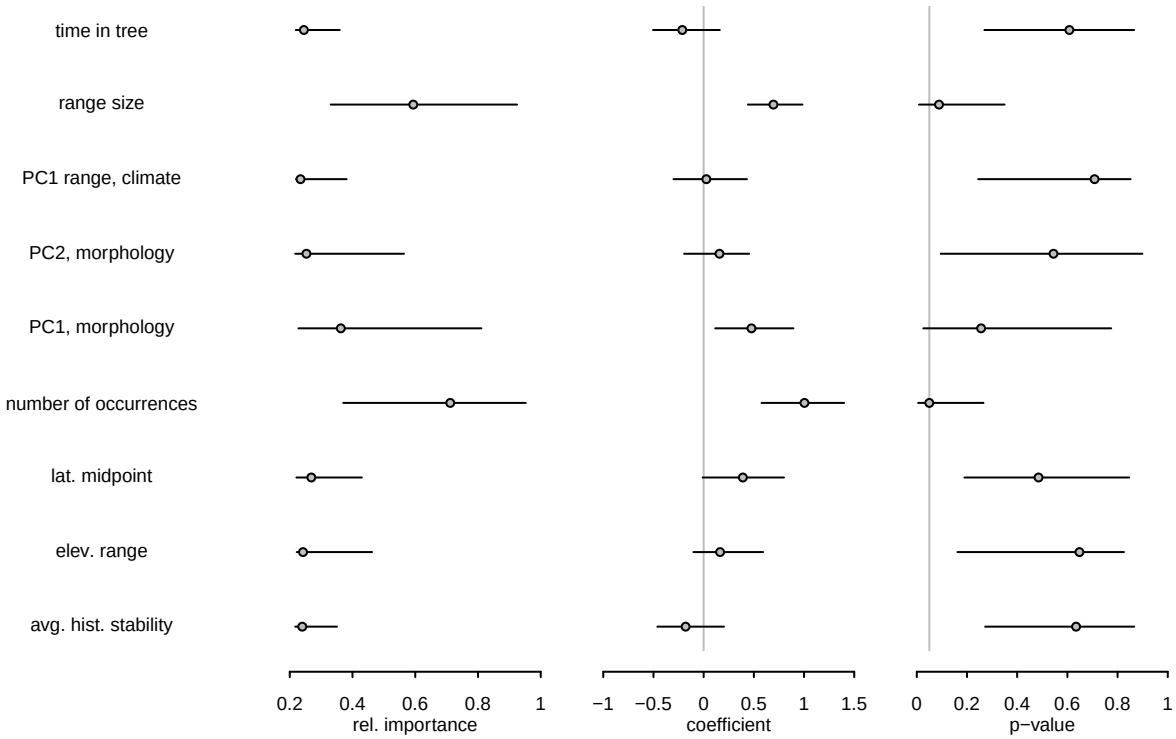

Figure S7: Coefficient of variation results for models fitting nine independent variables to within-population π based on 100 bootstraps of the full data set. Shown are the 95% range and median values for relative importance, coefficient, and p-value for each independent variable. The vertical gray line in the coefficient plot is at 0 and at 0.05 in the p-value plot.
These results show broad variance around most parameter estimates for most independent variables, suggesting that more sampling would be beneficial and indicates the presence of outliers. These results further suggest that range size is likely an important factor predicting genetic variation.

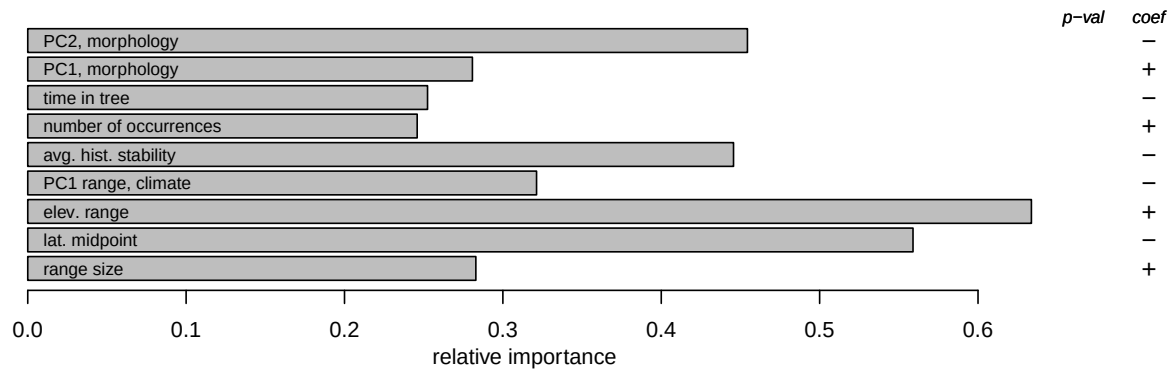

Figure S8: Model fitting for species-wide estimates of mitochondrial DNA nucleotide diversity (π) by nine independent variables that test the three hypotheses that might explain variation in genetic diversity. Shown are the relative importance, p-value significance, and directionality of coefficient for each variable as summarized across all additive models, weighted by relative AIC weights. None of these factors significantly predicted genetic diversity.

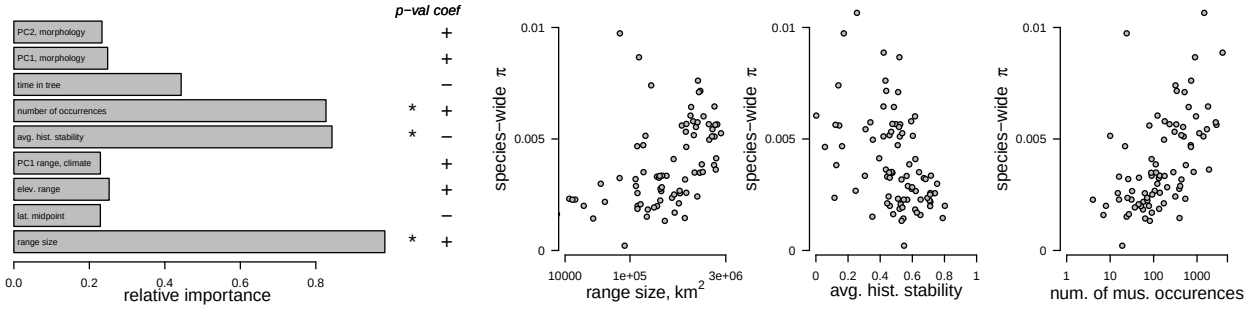

Figure S9: Model fitting for species-wide estimates of nucleotide diversity (π) by nine independent variables that test the three hypotheses that might explain variation in genetic diversity. (A) Shown are the relative importance, p-value significance, and directionality of coefficient for each variable as summarized across all additive models, weighted by relative AIC weights. (B) The relationship between species-wide π and one of the model's significant variables, range size. (C) The relationship between species-wide π and one of the model's significant variables, average historical stability. (D) The relationship between species-wide π and one of the model's significant variables, number of museum occurrences. The significance of range size and number of museum occurrences lend support the hypothesis that genetic diversity is a function of census population size, and the significance of historical stability supports the hypothesis that history impacts genetic diversity.

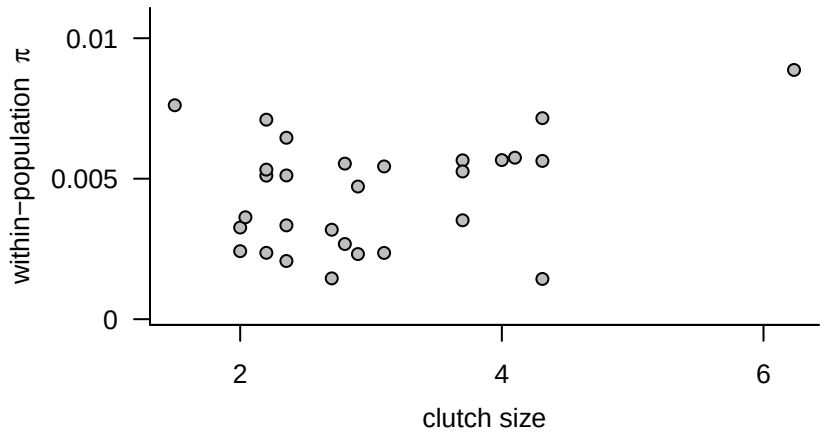

Figure S10: Correlation between clutch size (as detailed in Table S2) and within-population nucleotide diversity, π. The trend between the two variables is positive and non-significant (phylogenetic generalized least-squares; p-value=0.31). Overall variation in this life-history trait is low – but for one outlier, all species have clutch sizes between 2 and 4. Raw data used in this figure are given in Table S4.

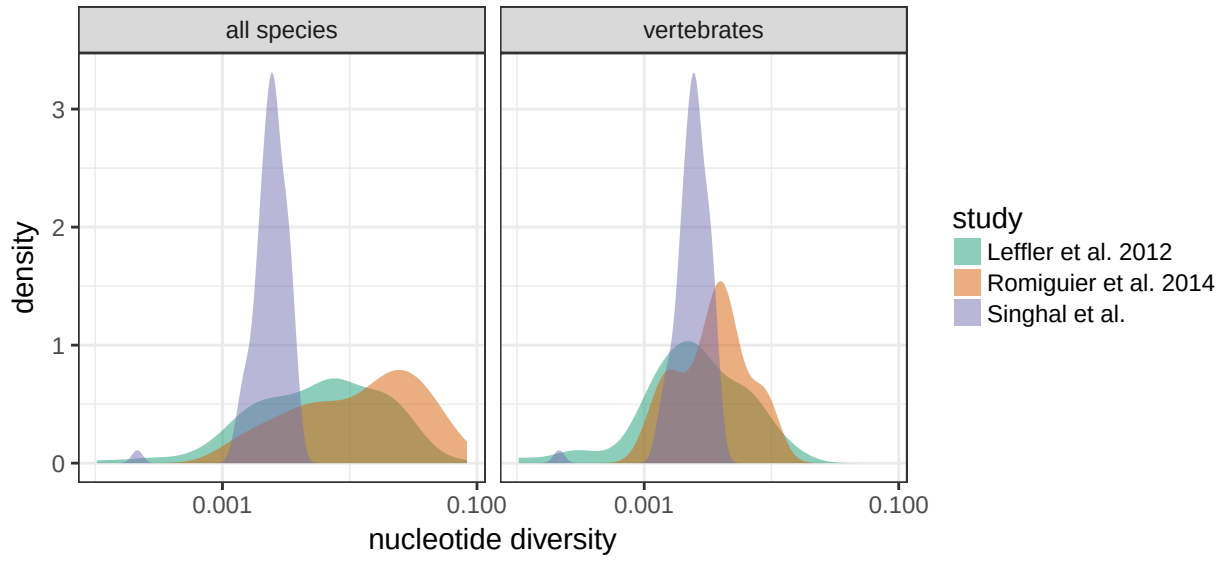

Figure S11: Comparisons of nucleotide diversity estimates across three comparative studies of genetic diversity: Leffler *et al.*, 2012 (72), Romiguier *et al.* 2014 (87), and the present study. Leffler *et al.* estimates are based on mean values of nucleotide diversity (measured as π_S) across populations within a species for a varied set of nuclear loci, and these data summarize patterns in eukaryotes, including plants, paramecium, and chordates. Romiguier *et al.* estimates are π_S measured across individuals within a species for a set of highly-expressed coding sequences, and these data summarize patterns in animals, including sea squirts, termites, and penguins. The left panel shows all species; the right panel only shows vertebrate species. These three studies estimated diversity across non-homologous sets of loci that likely have different selection histories and local recombination environments. The data should be interpreted accordingly. In general, these results show that the OTUs in *Ctenotus* exhibit a narrower range of genetic diversity levels compared to that seen in Leffler *et al.* and Romiguier *et al.*. However, this range is broader than might be expected given the ecological and phylogenetic similarity of *Ctenotus*.

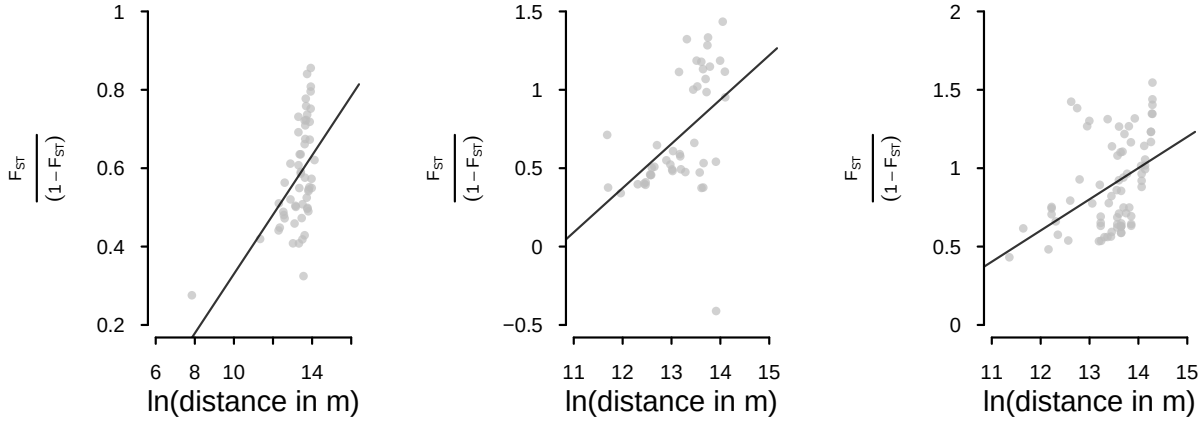

Figure S12: Isolation-by-distance relationships for three randomly selected OTUs from the present study. Linear regressions are fit to those OTUs which show significant isolation-by-distance (IBD) relationships across their range. This figure indicates that many of the OTUs in this study show significant IBD patterns. As such, calculating genetic diversity across individuals will summarize both local patterns of diversity and range-wide patterns of differentiation. Thus, we instead focus on local patterns of diversity (as summarized by within-population π), although we report species-wide patterns elsewhere (Fig. S4, S9).

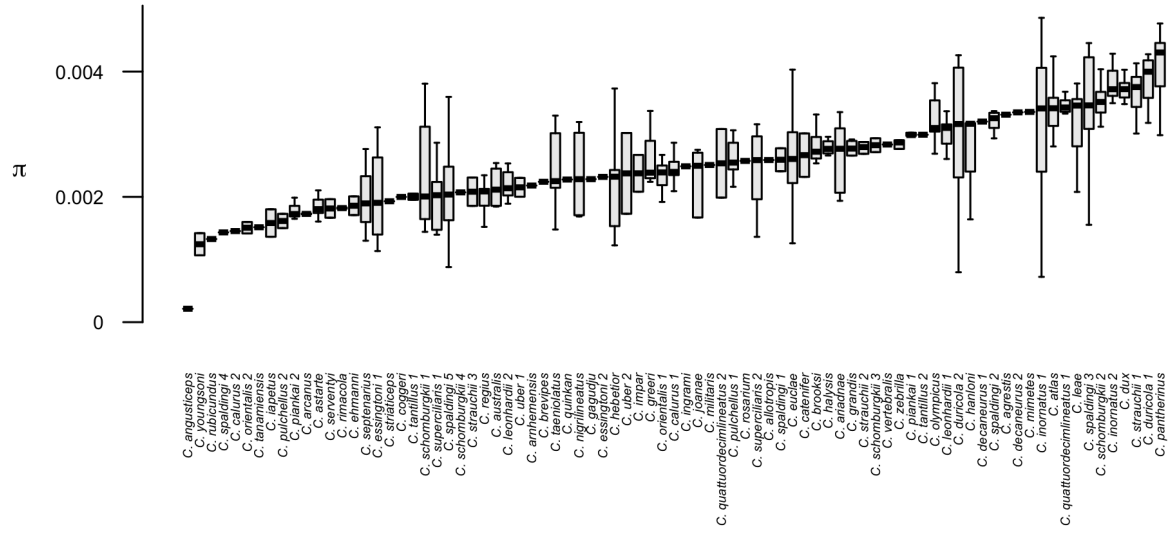

Figure S13: Range of within-population (π) for each OTU. These data suggest genetic diversity varies across an OTU's range. Thus, this work focuses on mean within-population genetic diversity.

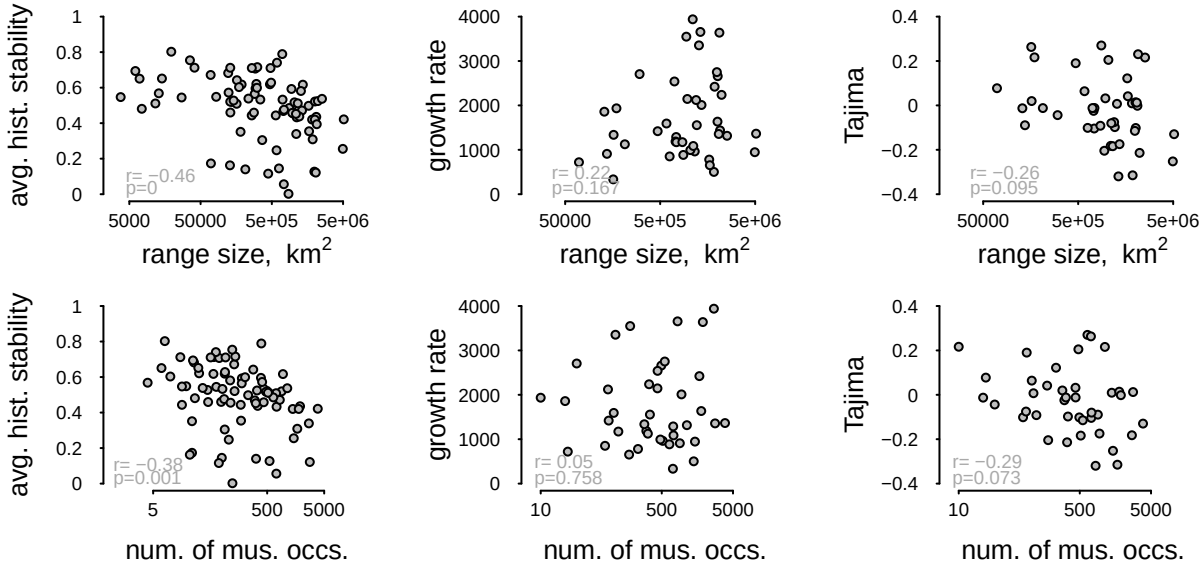

Figure S14: Three measures of changes in historical demography (average historical stability, growth rate as inferred by LAMARC, and Tajima's D as inferred by ANGSD) as they relate to two proxies for census population size – range size in km^2 and number of occurrences in museum databases. Shown are Spearman correlations. Growth rate and Tajimas D were only calculated for OTUs with $\geq$3 individuals (N=42). These data suggest that more abundant species have experienced less stable histories than less abundant species.

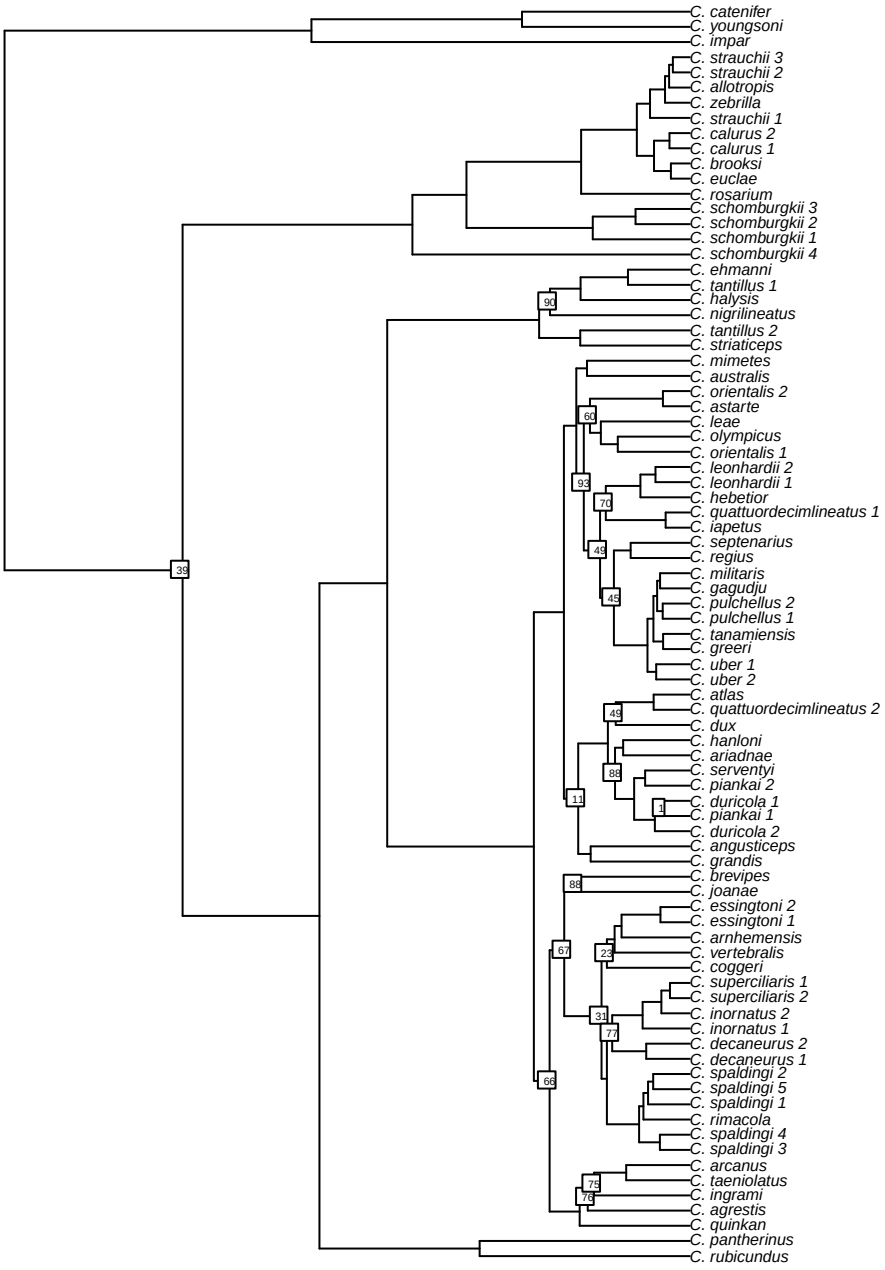

Figure S15: The "species tree" for species-level lineages in *Ctenotus*, as inferred by ASTRID. Nodes with bootstrap support <95% are labeled.

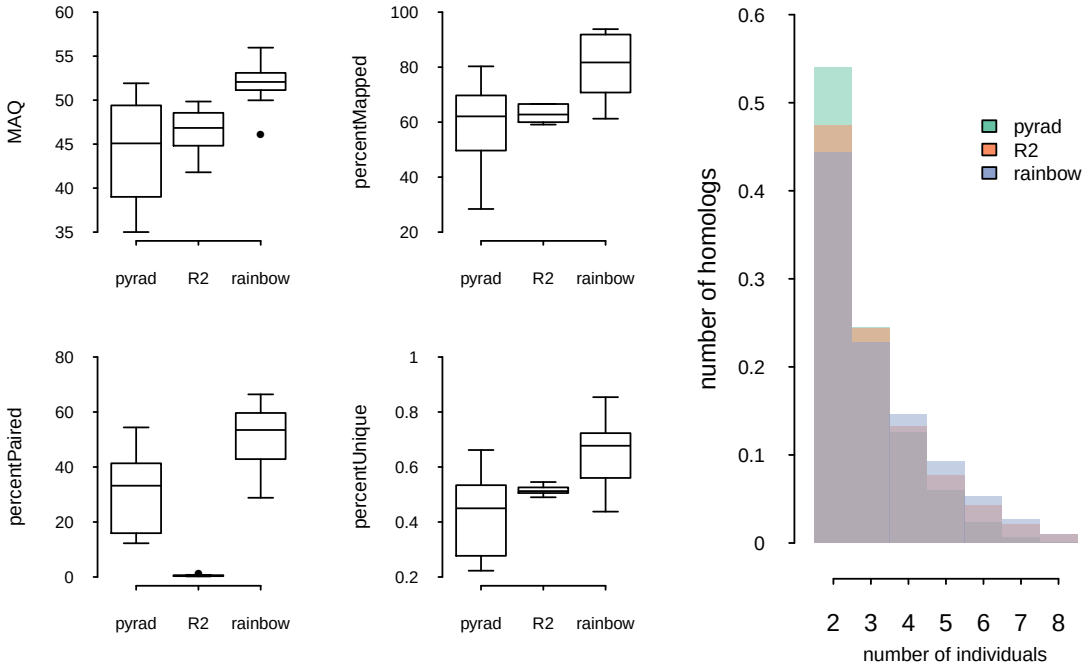

Figure S16: Quality metrics for three different ddRAD assembly approaches: PyRAD, R2: a method that clusters the reverse read of each read-pair, and Rainbow. Metrics measured are: mapping alignment quality (MAQ), the percent of reads used in the assembly that can align back to the assembly (percentMapped), the percent of reads that map as pairs (percentPaired), and the percent of reads that align uniquely to the assembly (percentUnique), and for every given locus in the assembly, the number of individuals in that species with a homolog. These metrics were quantified for eight randomly selected OTUs. Across all metrics, higher values indicate higher assembly quality. Thus, all metrics suggest that Rainbow resulted in the best assemblies.

References

[1] Rabosky, D. L., Hutchinson, M. N., Donnellan, S. C., Talaba, A. L. & Lovette, I. J., 2014 Phylogenetic disassembly of species boundaries in a widespread group of australian skinks (scincidae: Ctenotus). *Molecular phylogenetics and evolution* **77**, 71–82.

[2] Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E., 2012 Double digest radseq: an inexpensive method for de novo snp discovery and genotyping in model and non-model species. *PloS one* **7**, e37135.

[3] Eaton, D. A., 2014 PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* p. btu121.

[4] Bolger, A. M., Lohse, M. & Usadel, B., 2014 Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* p. btu170.

[5] Chong, Z., Ruan, J. & Wu, C.-I., 2012 Rainbow: an integrated tool for efficient clustering and assembling rad-seq reads. *Bioinformatics* **28**, 2732–2737.

[6] Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F., 2016 Vsearch: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584.

[7] Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792–1797.

[8] Stamatakis, A., 2014 Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

[9] Price, M. N., Dehal, P. S. & Arkin, A. P., 2010 Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490.

[10] Schliep, K. P., 2011 phangorn: phylogenetic analysis in r. *Bioinformatics* **27**, 592–593.

[11] Rabosky, D. L., Donnellan, S. C., Grundler, M. & Lovette, I. J., 2014 Analysis and visualization of complex macroevolutionary dynamics: an example from australian scincid lizards. *Systematic biology* **63**, 610–627.

[12] Fujisawa, T. & Barraclough, T. G., 2013 Delimiting species using single-locus data and the generalized mixed yule coalescent (gmyc) approach: a revised method and evaluation on simulated datasets. *Systematic biology* p. syt033.

[13] Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N. *et al.*, 2007 Population genomics: whole-genome analysis of polymorphism and divergence in drosophila simulans. *PLoS Biol* **5**, e310.

[14] Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997* .

[15] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. *et al.*, 2009 The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079.

[16] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.*, 2010 The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research* **20**, 1297–1303.

[17] Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A., 2012 Bayesian phylogenetics with beauti and the beast 1.7. *Molecular biology and evolution* **29**, 1969–1973.

[18] Sukumaran, J. & Holder, M. T., 2010 DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571.

[19] Paradis, E., Claude, J. & Strimmer, K., 2004 Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics* **20**, 289–290.

[20] Nee, S., Read, A. F., Greenwood, J. J. & Harvey, P. H., 1991 The relationship between abundance and body size in british birds. *Nature* **351**, 312–313.

[21] Corbett-Detig, R. B., Hartl, D. L. & Sackton, T. B., 2015 Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol* **13**, e1002112.

[22] Rabosky, A. R. D., Cox, C. L., Rabosky, D. L., Title, P. O., Holmes, I. A., Feldman, A., McGuire, J. A. *et al.*, 2016 Coral snakes predict the evolution of mimicry across new world snakes. *Nature communications* **7**.

[23] Phillips, S. J. & Dudík, M., 2008 Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161–175.

[24] Zomer, R., Trabucco, A., van Straaten, O. & Bossio, D., 2006 *Carbon, land and water: A global analysis of the hydrologic dimensions of climate change mitigation through afforestation/reforestation*, volume 101. IWMI.

[25] Zomer, R. J., Trabucco, A., Bossio, D. A. & Verchot, L. V., 2008 Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture, ecosystems & environment* **126**, 67–80.

[26] Pateiro-López, B. & Rodrıguez-Casal, A., 2010 Generalizing the convex hull of a sample: the r package alphahull. *Journal of Statistical Software* **34**, 1–28.

[27] Hijmans, R. J. & van Etten, J., 2014 raster: Geographic data analysis and modeling. *R package version* **2**, 15.

[28] Bivand, R. & Rundel, C., 2013 rgeos: interface to geometry engine-open source (geos). *R package version 0.3-2* .

[29] Moritz, C., Fujita, M., Rosauer, D., Agudo, R., Bourke, G., Doughty, P., Palmer, R., Pepper, M., Potter, S., Pratt, R. *et al.*, 2015 Multilocus phylogeography reveals nested endemism in a gecko across the monsoonal tropics of australia. *Molecular ecology* .

[30] Oliver, P. M., Adams, M., Lee, M. S., Hutchinson, M. N. & Doughty, P., 2009 Cryptic diversity in vertebrates: molecular data double estimates of species diversity in a radiation of australian lizards (diplodactylus, gekkota). *Proceedings of the Royal Society of London B: Biological Sciences* **276**, 2001–2007.

[31] Potter, S., Bragg, J. G., Peter, B. M., Bi, K. & Moritz, C., 2016 Phylogenomics at the tips: inferring lineages and their demographic history in a tropical lizard, carlia amax. *Molecular ecology* .

[32] Boakes, E. H., McGowan, P. J., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K. & Mace, G. M., 2010 Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol* **8**, e1000385.

[33] Levene, H., 1953 Genetic equilibrium when more than one ecological niche is available. *American Naturalist* **87**, 331–333.

[34] Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A., 2005 Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology* **25**, 1965–1978.

[35] Charlesworth, B., 2009 Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10**, 195–205.

[36] Ho, L. S. T., Ane, C., Lachlan, R., Tarpinian, K., Feldman, R. & Ho, M. L. S. T., 2016 Package phylolm .

[37] Burnham, K. & Anderson, D., 1998 Model selection and inference: a practical information-theoretic approach springer-verlag. *New York* .

[38] Pianka, E. R., 1969 Sympatry of desert lizards (Ctenotus) in Western Australia. *Ecology* pp. 1012–1030.

[39] James, C. D., 1991 Annual variation in reproductive cycles of scincid lizards (Ctenotus) in central Australia. *Copeia* pp. 744–760.

[40] James, C. & Shine, R., 1988 Life-history strategies of Australian lizards: a comparison between the tropics and the temperate zone. *Oecologia* **75**, 307–316.

[41] Shine, R. & Greer, A. E., 1991 Why are clutch sizes more variable in some species than in others? *Evolution* pp. 1696–1706.

[42] Pearson, D. J., Jones, B., Australia, W. & Team, L. I. S. R., 2000 *Lancelin Island skink recovery plan*. Citeseer.

[43] Read, J., 1998 The ecology of sympatric scincid lizards (Ctenotus) in arid South Australia. *Australian Journal of Zoology* **46**, 617–629.

[44] Vitt, L. J. & Congdon, J. D., 1978 Body shape, reproductive effort, and relative clutch mass in lizards: resolution of a paradox. *American Naturalist* pp. 595–608.

[45] Taylor, J. A., 1985 Reproductive biology of the Australian lizard Ctenotus taeniolatus. *Herpetologica* pp. 408–418.

[46] Avise, J. C., 1992 Molecular population structure and the biogeographic history of a regional fauna: a case history with lessons for conservation biology. *Oikos* pp. 62–76.

[47] Bazin, E., Glémin, S. & Galtier, N., 2006 Population size does not influence mitochondrial genetic diversity in animals. *Science* **312**, 570–572.

[48] Berkelhamer, R. C., 1983 Intraspecific genetic variation and haplodiploidy, eusociality, and polygyny in the hymenoptera. *Evolution* pp. 540–545.

[49] Burney, C. W. & Brumfield, R. T., 2009 Ecology predicts levels of genetic differentiation in neotropical birds. *The American Naturalist* **174**, 358–368.

[50] Chen, J., Glemin, S. & Lascoux, M., 2017 Genetic diversity and the effiacy of purifying selection across plant and animal species. *Molecular Biology and Evolution* **msx088**.

[51] Cole, C. T., 2003 Genetic variation in rare and common plants. *Annual Review of Ecology, Evolution, and Systematics* **34**, 213–237.

[52] Dalongeville, A., Andrello, M., Mouillot, D., Albouy, C. & Manel, S., 2015 Ecological traits shape genetic diversity patterns across the mediterranean sea: a quantitative review on fishes. *Journal of Biogeography* .

[53] Delrieu-Trottin, E., Maynard, J. & Planes, S., 2014 Endemic and widespread coral reef fishes have similar mitochondrial genetic diversity. *Proceedings of the Royal Society of London B: Biological Sciences* **281**, 20141068.

[54] DeWoody, J. & Avise, J., 2000 Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. *Journal of fish biology* **56**, 461–473.

[55] Doyle, J. M., Hacking, C. C., Willoughby, J. R., Sundaram, M. & DeWoody, J. A., 2015 Mammalian genetic diversity as a function of habitat, body size, trophic class, and conservation status. *Journal of Mammalogy* **96**, 564–572.

[56] Duminil, J., Fineschi, S., Hampe, A., Jordano, P., Salvini, D., Vendramin, G. G. & Petit, R. J., 2007 Can population genetic structure be predicted from life-history traits? *The American Naturalist* **169**, 662–672.

[57] Eo, S., Doyle, J. & DeWoody, J., 2011 Genetic diversity in birds is associated with body mass and habitat type. *Journal of Zoology* **283**, 220–226.

[58] Evans, S. R. & Sheldon, B. C., 2008 Interspecific patterns of genetic diversity in birds: correlations with extinction risk. *Conservation Biology* **22**, 1016–1025.

[59] Frankham, R., 1997 Do island populations have less genetic variation than mainland populations? *Heredity* **78**.

[60] Fujisawa, T., Vogler, A. P. & Barraclough, T. G., 2015 Ecology has contrasting effects on genetic variation within species versus rates of molecular evolution across species in water beetles. In *Proc. R. Soc. B*, volume 282, p. 20142476. The Royal Society.

[61] Garner, A., Rachlow, J. L. & Hicks, J. F., 2005 Patterns of genetic diversity and its loss in mammalian populations. *Conservation Biology* **19**, 1215–1221.

[62] Glémin, S., Bazin, E. & Charlesworth, D., 2006 Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proceedings of the Royal Society of London B: Biological Sciences* **273**, 3011–3019.

[63] Graur, D., 1985 Gene diversity in hymenoptera. *Evolution* **39**, 190–199.

[64] Gyllensten, U., 1985 The genetic structure of fish: differences in the intraspecific distribution of biochemical genetic variation between marine, anadromous, and freshwater species. *Journal of Fish Biology* **26**, 691–699.

[65] Hamrick, J. L. & Godt, M., 1990 Allozyme diversity in plant species. *Plant population genetics, breeding, and genetic resources.* pp. 43–63.

[66] Hamrick, J. L., Godt, M. J. W. & Sherman-Broyles, S. L., 1992 Factors influencing levels of genetic diversity in woody plant species. In *Population genetics of forest trees*, pp. 95–124. Springer.

[67] Hamrick, J. L. & Godt, M., 1996 Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society B: Biological Sciences* **351**, 1291–1298.

[68] Harvey, M., Aleixo, A., Ribas, C. C. & Brumfield, R. T., 2016 Habitat preference predicts genetic diversity and population divergence in amazonian birds. *bioRxiv* p. 085126.

[69] Hedrick and, P. W. & Parker, J. D., 1997 Evolutionary genetics and genetic variation of haplodiploids and x-linked genes. *Annual Review of Ecology and Systematics* **28**, 55–83.

[70] James, J. E., Lanfear, R. & Eyre-Walker, A., 2016 Molecular evolutionary consequences of island colonization. *Genome Biology and Evolution* **8**, 1876–1888.

[71] Karron, J. D., 1987 A comparison of levels of genetic polymorphism and self-compatibility in geographically restricted and widespread plant congeners. *Evolutionary Ecology* **1**, 47–58.

[72] Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Segurel, L., Venkat, A., Andolfatto, P. & Przeworski, M., 2012 Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol* **10**, e1001388.

[73] Loveless, M., 1992 Isozyme variation in tropical trees: patterns of genetic organization. *New forests* **6**, 67–94.

[74] McCusker, M. R. & Bentzen, P., 2010 Positive relationships between genetic diversity and abundance in fishes. *Molecular Ecology* **19**, 4852–4862.

[75] Miller, M. J., Bermingham, E., Klicka, J., Escalante, P. & Winker, K., 2010 Neotropical birds show a humped distribution of within-population genetic diversity along a latitudinal transect. *Ecology Letters* **13**, 576–586.

[76] Mitton, J. B. & Lewis Jr, W. M., 1989 Relationships between genetic variability and life-history features of bony fishes. *Evolution* pp. 1712–1723.

[77] Nabholz, B., Mauffrey, J.-F., Bazin, E., Galtier, N. & Glemin, S., 2008 Determination of mitochondrial genetic diversity in mammals. *Genetics* **178**, 351–361.

[78] Nei, M. & Graur, D., 1984 Extent of protein polymosphism and the neutral mutation theory. *Evolutionary Biology* **17**, 73–118.

[79] Nevo, E., 1988 Genetic diversity in nature. In *Evolutionary biology*, pp. 217–246. Springer.

[80] Nybom, H. & Bartish, I. V., 2000 Effects of life history traits and sampling strategies on genetic diversity estimates obtained with rapd markers in plants. *Perspectives in plant ecology, evolution and systematics* **3**, 93–114.

[81] Nybom, H., 2004 Comparison of different nuclear dna markers for estimating intraspecific genetic diversity in plants. *Molecular ecology* **13**, 1143–1155.

[82] Packer, L., Zayed, A., Grixti, J. C., Ruz, L., Owen, R. E., Vivallo, F. & Toro, H., 2005 Conservation genetics of potentially endangered mutualisms: reduced levels of genetic variation in specialist versus generalist bees. *Conservation Biology* **19**, 195–202.

[83] Papadopoulou, A., Anastasiou, I., Spagopoulou, F., Stalimerou, M., Terzopoulou, S., Legakis, A. & Vogler, A. P., 2011 Testing the species–genetic diversity correlation in the aegean archipelago: toward a haplotype-based macroecology? *The American Naturalist* **178**, 241–255.

[84] Perry, G. H., Melsted, P., Marioni, J. C., Wang, Y., Bainer, R., Pickrell, J. K., Michelini, K., Zehr, S., Yoder, A. D., Stephens, M. *et al.*, 2012 Comparative rna sequencing reveals substantial genetic variation in endangered primates. *Genome research* **22**, 602–610.

[85] Pinsky, M. L. & Palumbi, S. R., 2014 Meta-analysis reveals lower genetic diversity in overfished populations. *Molecular Ecology* **23**, 29–39.

[86] Roberts, D. R. & Hamann, A., 2015 Glacial refugia and modern genetic diversity of 22 western north american tree species. *Proceedings of the Royal Society of London B: Biological Sciences* **282**, 20142903.

[87] Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N. *et al.*, 2014 Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**, 261–263.

[88] Roselius, K., Stephan, W. & Städler, T., 2005 The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics* **171**, 753–763.

[89] Rossetto, M., Crayn, D., Ford, A., Mellick, R. & Sommerville, K., 2009 The influence of environment and life-history traits on the distribution of genes and individuals: a comparative study of 11 rainforest trees. *Molecular Ecology* **18**, 1422–1438.

[90] Shaffer, H. B. & Breden, F., 1989 The relationship between allozyme variation and life history: Non-transforming salamanders are less variable. *Copeia* pp. 1016–1023.

[91] Soltis, P. S. & Gitzendanner, M. A., 1999 Molecular systematics and the conservation of rare species. *Conservation Biology* **13**, 471–483.

[92] Soulé, M., 1976 Allozyme variation: its determinants in space and time. *Molecular evolution* pp. 60–77.

[93] Taberlet, P., Zimmermann, N. E., Englisch, T., Tribsch, A., Holderegger, R., Alvarez, N., Niklfeld, H., Coldea, G., Mirek, Z., Moilanen, A. *et al.*, 2012 Genetic diversity in widespread species is not congruent with species richness in alpine plant communities. *Ecology Letters* **15**, 1439–1448.

[94] Thiel-Egenter, C., Gugerli, F., Alvarez, N., Brodbeck, S., Cieślak, E., Colli, L., Englisch, T., Gaudeul, M., Gielly, L., Korbecka, G. *et al.*, 2009 Effects of species traits on the genetic diversity of high-mountain plants: a multi-species study across the alps and the carpathians. *Global Ecology and Biogeography* **18**, 78–87.

[95] Ward, R., Woodwark, M. & Skibinski, D., 1994 A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. *Journal of fish biology* **44**, 213–232.

[96] Wooten, M. C. & Smith, M. H., 1985 Large mammals are genetically less variable? *Evolution* **39**, 210–212.