Syst. Biol. 0(0):1–21, 2021

DOI:10.1093/sysbio/syab034

Gene Flow Increases Phylogenetic Structure and Inflates Cryptic Species Estimations: A Case Study on Widespread Philippine Puddle Frogs (*Occidozyga laevis*)

KIN ONN CHAN[1,*], CARL R. HUTTER[2,3], PERRY L. WOOD JR.[4], YONG-CHAO SU[5], AND RAFE M. BROWN[2]

[1]*Lee Kong Chian National History Museum, Faculty of Science, National University of Singapore, 2 Conservatory Drive, 117377 Singapore;*
[2]*Department of Ecology and Evolutionary Biology and Biodiversity Institute, University of Kansas, Dyche Hall, 1345 Jayhawk Blvd, Lawrence, Kansas 66045-7561, USA;*
[3]*Museum of Natural Sciences and Department of Biological Sciences, 119, Foster Hall. Louisiana State University, Baton Rouge, LA 70803, USA;*
[4]*Department of Biological Sciences & Museum of Natural History, Auburn University, 101 Rouse Life Sciences Building, Auburn, Alabama 36849;*
[5]*Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, 100, Shih-Chuan 1st Rd, Kaohsiung City, 80708, Taiwan*
**Correspondence to be sent to: Lee Kong Chian National History Museum, Faculty of Science, National University of Singapore, 2 Conservatory Drive, 117377 Singapore;*
E-mail: cko@nus.edu.sg.

November 12 2020; reviews returned 29 April 2021; accepted 6 May 2021
Associate Editor: Frank Burbrink

Abstract.—In cryptic amphibian complexes, there is a growing trend to equate high levels of genetic structure with hidden cryptic species diversity. Typically, phylogenetic structure and distance-based approaches are used to demonstrate the distinctness of clades and justify the recognition of new cryptic species. However, this approach does not account for gene flow, spatial, and environmental processes that can obfuscate phylogenetic inference and bias species delimitation. As a case study, we sequenced genome-wide exons and introns to evince the processes that underlie the diversification of Philippine Puddle Frogs—a group that is widespread, phenotypically conserved, and exhibits high levels of geographically based genetic structure. We showed that widely adopted tree- and distance-based approaches inferred up to 20 species, compared to genomic analyses that inferred an optimal number of five distinct genetic groups. Using a suite of clustering, admixture, and phylogenetic network analyses, we demonstrate extensive admixture among the five groups and elucidate two specific ways in which gene flow can cause overestimations of species diversity: 1) admixed populations can be inferred as distinct lineages characterized by long branches in phylograms; and 2) admixed lineages can appear to be genetically divergent, even from their parental populations when simple measures of genetic distance are used. We demonstrate that the relationship between mitochondrial and genome-wide nuclear *p*-distances is decoupled in admixed clades, leading to erroneous estimates of genetic distances and, consequently, species diversity. Additionally, genetic distance was also biased by spatial and environmental processes. Overall, we showed that high levels of genetic diversity in Philippine Puddle Frogs predominantly comprise metapopulation lineages that arose through complex patterns of admixture, isolation-by-distance, and isolation-by-environment as opposed to species divergence. Our findings suggest that speciation may not be the major process underlying the high levels of hidden diversity observed in many taxonomic groups and that widely adopted tree- and distance-based methods overestimate species diversity in the presence of gene flow. [Cryptic species; gene flow; introgression; isolation-by-distance; isolation-by-environment; phylogenetic network; species delimitation.]

As taxonomic knowledge increases over time, new species discoveries have disproportionately and nonrandomly shifted from deep lineages to shallower nodes towards the tips of evolutionary trees (Blackburn et al. 2019). New species discoveries at shallow nodes are further facilitated by advances in genomic sequencing that have enabled the characterization of genetic structure at much finer scales, leading to the discovery of purportedly high levels of cryptic, or hidden species diversity throughout the tree of life (Bickford et al. 2007; Pfenninger and Schwenk 2007; Trontelj and Fišer 2009; Struck et al. 2018). The widespread use of genetic data has also accelerated the process of species discovery by providing a framework for phylogeny- and sequence-based species delimitation approaches (Vences et al. 2005a; Fouquet et al. 2007; Vieites et al. 2009; Yang and Rannala 2010; Puillandre et al. 2012; Fujisawa and Barraclough 2013; Jones et al. 2015; Kapli et al. 2017). These approaches are effective at delimiting allopatric and deeply divergent species, but, as attention shifts towards shallower lineages involving continuously distributed and/or recently diverged populations, species boundaries can be increasingly obfuscated by spatial (Barley et al. 2015; Bradburd and Ralph 2019) and microevolutionary processes that act at the population level (Chan et al. 2017, 2020b; Jackson et al. 2017; Leaché et al. 2019). Unfortunately, these potentially confounding processes are rarely taken into consideration during cryptic species delimitation; hence, the extent of their impact on species estimation remains poorly understood.

The increased use of molecular phylogenies to delineate morphologically cryptic groups has led to a growing trend that equates high levels of genetic structure with unrecognized or hidden species diversity. Typically, large-scale phylogenies are estimated using mitochondrial and/or a handful of nuclear genes, and cryptic species diversity is inferred based on phylogenetic structure, arbitrary genetic distance thresholds, or divergence-based species delimitation analyses. This practice is prevalent across various taxonomic groups including fishes (Kon et al. 2007; Thomas et al. 2014; Divya et al. 2017; Shelley et al. 2018), mammals (Manthey et al. 2011; Demos et al. 2018; Chen et al. 2020; Pozzi et al. 2020), arthropods (Crivellaro et al. 2018; Schäffer et al. 2019; Zhang et al. 2019;

MANUSCRIPT CATEGORY: Systematic Biology

Jusoh et al. 2020; Mignotte et al. 2020; Sánchez-Restrepo et al. 2020; Rubinoff et al. 2021), annelids (Cerca et al. 2020; Grosse et al. 2020; Martinsson and Erséus 2021), cnidarians (Schuchert 2014; Postaire et al. 2016), algae/plants (de Jesus et al. 2019; Díaz-Tapia et al. 2020), amphibians (Vieites et al. 2009; McLeod 2010; Nishikawa et al. 2012; Rowley et al. 2015; Matsui et al. 2016; Chen et al. 2017; Amador et al. 2018; Chan et al. 2018; Trevisan et al. 2020), and reptiles (Grismer et al. 2013; Siler et al. 2014; Blair and Bryson 2017; Mendes et al. 2018; Davis et al. 2020). Implicit within these findings is the assumption that speciation is the dominant process of diversification; yet, making this assumption ignores other factors that can also increase genetic structure. For example, extrinsic factors such as improved analytical methods and the expanded breadth of geographic and genetic sampling can also lead to increased genetic structure. Analyzing more genes, partitions, and geographic populations with more sophisticated methods can lead to the splitting of genetic divergences at finer scales (Huang 2020).

More robust examinations involving genome-scale data are beginning to reveal that species delimitation criteria predicated on phylogenetic structure, distance measures, or even multispecies coalescent models can inflate species diversity by delimiting metapopulation lineages as species (Chan et al. 2017, 2020b; Sukumaran and Knowles 2017; Leaché et al. 2019; Huang 2020). This is largely due to the violation of model assumptions that assumes no gene flow between diverging species, instantaneous speciation, and that the reference phylogeny accurately represents the true species tree (Kapli et al. 2017; Liu et al. 2009). Furthermore, the likelihood of violating these assumptions is expected to be higher among continuously occurring or recently diverged taxa, or in cases where gene flow is plausible (Carstens et al. 2017), and to complicate matters, it has recently been shown that a small amount of gene flow is enough to cause substantial changes in gene histories and mislead species tree estimations (Jiao et al. 2020). To overcome these issues, approaches that explicitly test for population-level processes, particularly gene flow, have been shown to produce more accurate estimates of species boundaries (Chan et al. 2017, 2020b; Jackson et al. 2017; Morales and Carstens 2018; Smith and Carstens 2019). These findings indicate that gene flow could be prevalent in many cryptic taxa and that the purportedly high levels of hidden diversity within these groups could be artificially inflated.

The systematics of Southeast Asian Puddle Frogs (family Dicroglossidae; genus *Occidozyga*) have not been thoroughly reviewed, in part due to the ubiquitous and overlapping distribution ranges of many named species that have been recognized in the absence of discrete diagnostic morphological characters. Ambiguous species diagnoses have resulted in widespread misidentifications and taxonomic uncertainty that has persisted for decades (Inger 1954, 1966; Iskandar 1998; Ohler 2003). More recently, the application of genetic data has alluded to the possibility of numerous undescribed cryptic species in Southeast Asian Puddle Frogs. Preliminary studies, predominantly using mitochondrial data, identified three putative new species and high levels of hidden cryptic diversity in the *O. lima-martensii* complex (Chan 2013; Bogisich 2019). Similarly, the *Occidozyga laevis* complex from the Philippines has been suggested to comprise up to eight undescribed cryptic species (Chan et al. 2021). Within the *O. laevis* complex, notable intra- and interisland variations have been documented (Inger 1954) but few follow-up studies have been performed to determine the underlying source(s) of variation (but see Chan et al. 2021). In this study, we conducted dense sampling of *O. laevis* throughout all major islands in the Philippine archipelago (Fig. 1) and generated novel genomic sequence capture data consisting of exons, introns, and exonic single nucleotide polymorphisms (SNPs) to 1) explicate the causal processes that underlie high levels of genetic diversity in *O. laevis* and to 2) test the accuracy of phylogeny- and distance-based methods for estimating cryptic species diversity. We found that most genetic diversity in *O. laevis* can be attributed to genetic admixture (as opposed to speciation) and that phylogenetic structure resulting from gene flow can easily and erroneously be misconstrued as species divergence. The empirical evidence presented in this suggests that a large portion of genetic diversity in many cryptic species complexes comprises admixed metapopulation lineages as opposed to distinct cryptic species, and calls into question the prevalence and degree of hidden cryptic species diversity throughout the tree of life.

Materials and Methods

Sampling and Sanger Sequencing

We compiled a comprehensive genetic panel based on the 16S rRNA mitochondrial gene to provide a preliminary estimate of phylogenetic relationships to guide our selection of samples for subsequent genomic sequencing. This panel consists of 403 sequences (322 newly sequenced for this study and 81 publicly available sequences from GenBank to increase geographic coverage; Supplementary Table S1 available on Dryad at https://doi.org/10.5061/dryad.34tmpg4j1) from numerous populations throughout the entire distribution of *Occidozyga*, with particularly dense sampling across all major islands in the Philippines (Fig. 1). We used the primers 16Sc-L (5′-GTRGGCCTAAAAGCAGCCAC-3′), and 16Sd-H (5′-CTCCGGTCTGAACTCAGATGACGTAG-3′) to amplify and sequence the 16S gene (Evans et al. 2003). Amplification was done using the following PCR thermal profile: 95°C for 4 min, followed by 35 cycles of 95°C for 30 s, 52°C for 30 s, 72°C for 70 s, and a final extension phase at 72°C for 7 min (Chan and Grismer 2019). Amplified DNA products were subsequently

MANUSCRIPT CATEGORY: Systematic Biology

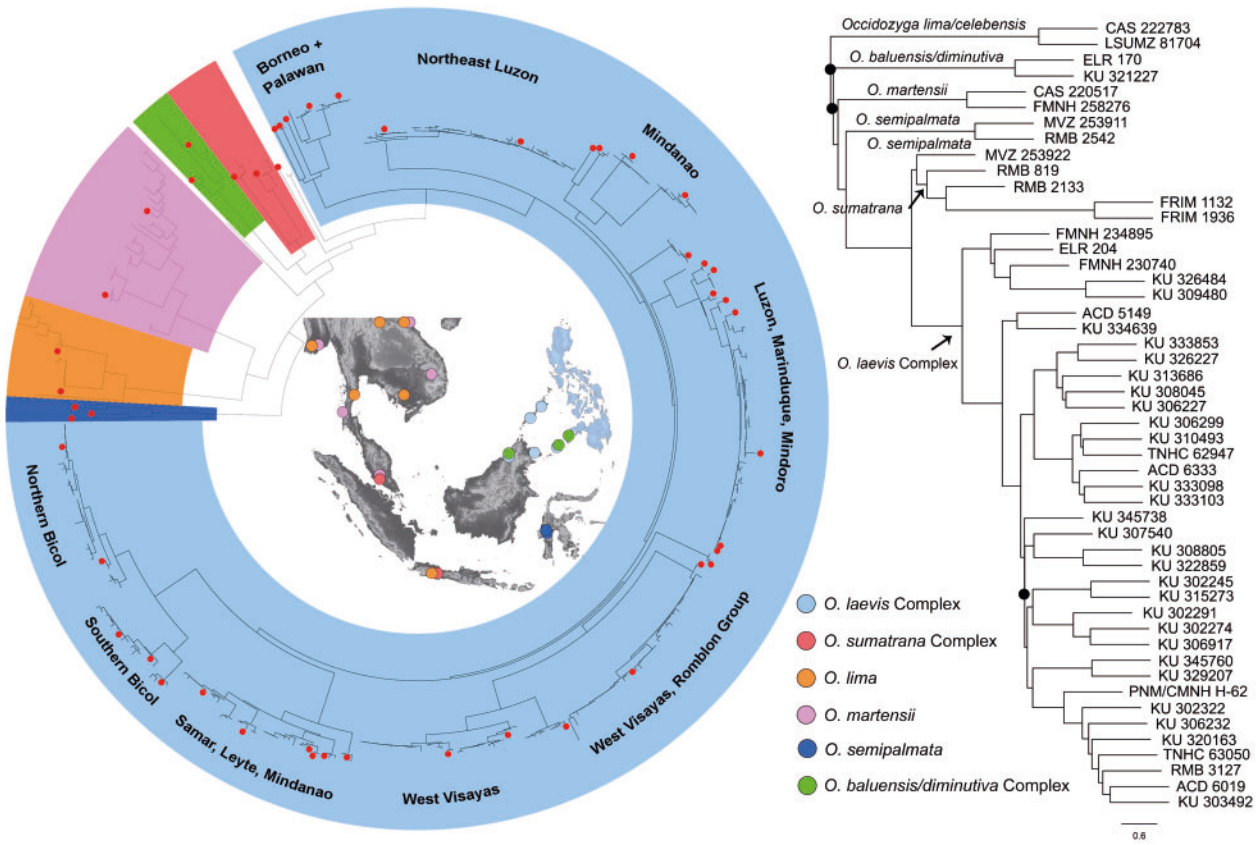

FIGURE 1. Left: Maximum likelihood phylogeny based on 1003 base pairs of the 16S rRNA mitochondrial gene and distribution map of samples used in this study. Red dots at tips represent samples that were selected for genomic sequencing using the FrogCap protocol. Right: ASTRAL-III phylogeny from the combined (intron+exon) data set (6274 gene trees). Black circles at nodes represent poorly supported branches (posterior probability < 0.95). All other branches were strongly supported (pp $= 1.0$).

visualized on 1.0% agarose gels and sequenced at GENEWIZ, South Plainfield, NJ. Sequences were assembled and aligned (MUSCLE algorithm, default settings) in Geneious Pro 5.3 (Kearse et al. 2012) prior to phylogenetic estimation.

For genomic sequencing, 50 samples were selected representing each major mitochondrial clade to maximize geographic and genetic coverage across all divergent clades. We used the modular FrogCap sequence capture marker set, creating a novel subset of markers shown to be successful in *Occidozyga* to increase capture success and reduce missing data (Hutter et al. 2019). A brief overview of the FrogCap protocol is provided below and additional details are provided in the Supplementary Material available on Dryad.

FrogCap Sequence Capture and Genomic Data Sets

We designed a reduced set of markers for population genetics (Reduced-Ranoidea probe set), where we selected exons from the main FrogCap marker set described in Hutter et al. (2019). Our goal was to select markers that were not part of the same gene and were separated by at least 100,000 base pairs to reduce the chances of linkage due to genomic proximity. We selected markers that were found in both *Occidozyga* and *Kaloula* (family Microhylidae), as our goal was to have this reduced set of markers work for both groups of taxa. We also included 47 ultralong exons (>5000 bp) previously not included in the Ranoidea V1 set. Furthermore, we included 86 "Legacy" markers from a recent study across the radiation of frogs (Feng et al. 2017). Final marker sequences were designed from the consensus sequences across the multiple sequence alignments from Feng et al. (2017) and were then used for probe design.

The Reduced-Ranoidea set was designed after the 40K Ranoidea V1 set had been tested, which enabled the selection of markers that were captured successfully across the entire superfamily Ranoidea (Hutter et al. 2019). To select markers that had already been tested, we began with alignments from the 24 sample Ranoidea set evaluated below and reduced the probe set to markers successfully captured across 75% or greater of the samples, which was 10,274 markers. Next, we filtered the markers as follows: 1) UCEs were excluded; 2) the largest exon within 100,000 bp of another exon on the *Nanorana* genome was retained reducing potential genetic linkage due to recombination; and 3) all exons

MANUSCRIPT CATEGORY: Systematic Biology

greater than 500 bp were included. This final candidate set included ~4312 exons. To accommodate the probe limit, exons were randomly deleted until the 20,020 bait limit was reached, resulting in 3161 exons retained. After combining the 86 Legacy markers, our final Reduced-Ranoidea marker set included 3247 markers targeting 1,519,233 bp of data.

The bioinformatics pipeline for filtering adapter contamination, assembling contigs, trimming, and alignment is scripted in R (R Core Team 2014), available at (https://github.com/chutter/FrogCap-Sequence-Capture; bioinformatics-pipeline_stable-v1), and described in detail in the Supplementary material available on Dryad. Our final set of matching markers was aligned on a marker-by-marker basis using MAFFT (Katoh and Standley 2013) local pair alignment. We screened each alignment for samples ≥40% divergent from consensus sequences, which were almost always incorrectly assigned contigs. Alignments were retained if they included four or more taxa, had ≥100 bp length, and mean sample specificities (i.e., the "breadth of coverage" of the sample; see below) ≥50% across the alignment (to prevent nonoverlapping segments of the alignment). We then separated alignments into two initial data sets: 1) "Exons-Only," which included only exon contigs with intronic regions trimmed from each alignment using the *Nanorana* genome sequence reference exon as a guide; and 2) "All-Markers," which included the entire matching contig to the reference marker. The Exons-Only and All-Markers alignment sets were further trimmed into usable phylogenetic analyses data sets, and data type comparisons, resulting in 1) "Exons," each exons-only alignment was adjusted to be in an open-reading frame in multiples of three bases and trimmed to the largest reading frame that accommodated >90% of the sequences; 2) "Introns," a consensus sequence of the exon previously delimited was aligned to the All-Markers data set and the aligning region was removed leaving only the two intron ends, which were concatenated; 3) "Combined," the exons and introns were not separated from each other; 4) "Legacy," markers from Feng et al. (2017) were saved separately for ease of access and comparison; Finally, the Introns data sets were internally trimmed using trimAl (Capella-Gutiérrez et al. 2009).

Variant Discovery.—variant calling for SNPs (single nucleotide polymorphisms) was conducted through a custom pipeline in R and is available at (https://github.com/chutter/FrogCap-Sequence-Capture; variant-pipeline_stable-v1). We used GATK v4.1 (McKenna et al. 2010), following developer best practices recommendations for discovering and calling variants (Van der Auwera et al. 2013). To discover potential variant data (e.g., SNPs, indels), we used a consensus sequence from each alignment from the target group as a reference and mapped cleaned reads back to reference markers from each sample. We used BWA (Li 2013) to map cleaned reads (cleaned-reads data set explained in Supplementary material available on Dryad) to our reference markers, adding the read group information (e.g., Flowcell, Lane, Library) obtained from the fastq header files. Next, we used SAMTOOLS (Liu et al. 2009) to convert the mapped reads SAM file to a cleaned BAM file, and merged BAM files with our unmapped reads, as required for downstream analyses. We used the program PICARD to mark exact duplicate reads that may have resulted from optical and PCR artifacts and reformatted each data set for variant calling. To locate variant and invariant sites, we used GATK to generate a preliminary variant data set using the GATK program *HaplotypeCaller*, to call haplotypes, in GVCF format, for each sample individually.

After processing each sample, we used the GATK *GenomicsDBImport* program to aggregate samples from separate data sets into their own combined database. Using these databases, we used the *GenotypeGVCF* function to genotype the combined sample data sets and output separate ".vcf" files for each marker, containing variant data, from all samples, for final filtration. The preliminary variant set was filtered into a final data set refining as follows: 1) All variants were kept after moderate filtering to remove probable errors filtered at a quality score > 5; 2) High-quality variants were kept including SNPs, MNPs (multinucleotide polymorphisms), and indels filtered at a quality > 20; 3) SNPs specifically were chosen after high-quality filtering (quality > 20); and 4) our final data set consisted of one high-quality SNP from each exon that was most variable across samples.

Phylogenetic Inference

A preliminary mitochondrial phylogeny (16S rRNA marker) was estimated using the maximum likelihood (ML) program IQ-TREE v1.6 (Nguyen et al. 2015). The best-fit substitution model was determined using ModelFinder (Kalyaanamoorthy et al. 2017) and branch support was assessed using 1000 ultrafast bootstrap replicates (Hoang et al. 2017). For genomic data, legacy, intron, exon, and combined (introns+exons) data sets were analyzed separately to assess variability in phylogenetic signal that could potentially stem from different classes of genomic data. Maximum likelihood and summary coalescent methods were performed separately on these data sets. Sequences were concatenated and IQ-TREE was used to estimate an ML phylogeny using a partitioned analysis (partitioned by marker) with the GTR+I+GAMMA substitution model applied to each partition (Chan et al. 2020a). Branch support was assessed using 1000 ultrafast bootstrap replicates (Hoang et al. 2017). IQ-TREE was also used to estimate individual gene trees within each data set using the GTR+I+GAMMA substitution model. These gene trees were used to calculate the gene concordance factor (gCF) (Minh et al. 2020) and also as input for the summary coalescent species tree analysis using ASTRAL-III (Zhang et al. 2017). The ASTRAL analysis was not performed on the legacy data set due to the low number of markers.

MANUSCRIPT CATEGORY: Systematic Biology

Species Network Analysis

Species networks were estimated under the multispecies coalescent model using NANUQ (Allman et al. 2019) and PhyloNet v3.8 (Wen et al. 2018). The complete set of gene trees from the intron data set (3142 gene trees; outgroups removed) was used in both analyses. NANUQ was implemented through the R package MSCQuartets v1.1.0 using small alpha (0.0001) and large beta (0.95) values for hypothesis testing as recommended by the developers. Test results on quartet counts were used to calculate a network quartet distance matrix between taxa. A splits graph under the Neighbor-Net algorithm was inferred from the quartet distance matrix using the software SplitsTree4 (Huson and Bryant 2006). The PhyloNet species network was inferred using maximum pseudolikelihood inference. We performed eight separate analyses (10 runs per analysis) with the number of maximum reticulations ranging from 3 to 10. Likelihood scores for the best run in each analysis were compared to select the network with the optimal number of reticulations.

Species Delimitation

We compared genetic distances and performed single-locus species delimitation analysis based on the mitochondrial 16S ribosomal RNA to generate a preliminary hypothesis of species boundaries (Vences et al. 2005a, 2005b; Hillis 2019). To characterize genetic divergences among clades, pairwise genetic distances (16S rRNA gene) were calculated using uncorrected *p*-distances and the Kimura two-parameter model in MEGA-X (Kumar et al. 2018). Putative species were then identified using the single-locus species delimitation program mPTP (Kapli et al. 2017). The mitochondrial phylogeny was used as the input tree and the minimum branch length was calculated using the *–minbr_auto* function in the mPTP program. The confidence of delimitation schemes was assessed using two independent MCMC chains at 10,000,000 generations each. To test whether mitochondrial divergence predicts nuclear genomic divergence, we performed linear regression and a Pearson's correlation test. Pairwise *p*-distances for the 37 samples from the *O. laevis* clade were calculated based on the 16S (1003 bp) and combined intron + exon alignments (2,709,020 bp). *P*-distances were scaled (subtracting the mean and dividing by the standard deviation) prior to analysis.

We then used genomic SNP data to validate these boundaries by implementing a variety of clustering, population ancestry, and hybrid analyses. The program GENODIVE v3.0 (Meirmans 2020) was used to perform a K-means analysis to divide samples into groups that maximize the among-group Sum of Squares calculated from an Analysis of Molecular Variance. Convergence was assessed by simulated annealing using 50,000 MCMC steps and the optimal number of clusters was determined using a pseudo-F index and Bayesian information criterion (BIC). The dimensionality of the SNP data set was then reduced using a principal component analysis (PCA) to uncover the underlying group structure. The PCA was implemented through the R package *adegenet* (Jombart and Ahmed 2011). Because applying PCA on high-dimensional data can be difficult to interpret (due to possible high numbers of retained PCs), we implemented the t-distributed stochastic neighbor embedding algorithm (t-SNE) that is better able to capture the local structure of high-dimensional data within 2–3 dimensions (van der Maaten and Hinton 2008; Li et al. 2017; Derkarabetian et al. 2019). This analysis was performed with the R package *Rtsne* (Krijthe 2015) using the following settings: perplexity = 11, max iterations = 1,000,000, theta = 0.0.

Population ancestry was estimated using a program based on sparse non-negative matrix factorization, sNMF (Frichot et al. 2014). Ancestry coefficients were estimated for 1–10 ancestral populations (K) using 100 replicates for each K. The crossentropy criterion was then used to determine the best K based on the prediction of masked genotypes. The sNMF analysis was implemented through the R package *LEA* (Frichot and François 2015).

Gene Flow

The program HyDe (Blischak et al. 2018) was used to estimate the amount of gene flow among individuals. Similar to the NANUQ and PhyloNet analyses, the intron alignment (1,423,892 sites) was used as input and the *individual_hyde* script was used to examine gene flow at the individual level with bootstrapping (100 replicates). The hypothesis-testing framework of HyDe differentiates between hybrid speciation and gene flow based on the underlying assumption that all individuals in a hybrid population are admixed. Although this distinction is not of primary interest, estimations of gene flow (γ) resulting from this test are useful to assess the integrity of putative species boundaries.

Isolation-by-distance and Isolation-by-environment

We tested for isolation-by-distance (IBD) using distance-based redundancy analysis (dbRDA) as opposed to the Mantel test that has been shown to be inappropriate when geographic distances are derived from spatial coordinates (Legendre et al. 2015). Pairwise Euclidean genetic distances among individuals were calculated in GENODIVE and used as a response variable. Geographic distances were derived from GPS coordinates and transformed into distance-based Moran's eigenvector maps (dbMEM) using the R package *adespatial*. First, a Euclidean distance matrix between all individuals was calculated with the threshold value (*thresh*) for the truncation of the distance matrix set as the length of the longest edge of the minimum spanning tree. Distances that were longer than the truncation threshold were modified to 4**thresh*. A principal coordinates analysis

TABLE 1. Summaries for the data sets analyzed in this study

Data set	Total markers	Basepairs	Variable sites	Total PIS	Prop. PIS
16S	1	1003	698	494	0.4930
Legacy	85	86,925	17,093	9374	0.1028
Exon	3143	1,285,128	347,696	178,613	0.1260
Intron	3131	1,423,892	880,658	496,324	0.3479
Combined	6274	2,709,020	1,228,354	674,937	0.2367

PIS = parsimony informative sites.

was then performed on the modified matrix and eigenfunctions that model positive spatial correlation of the dbMEMs were retained as spatial variables. The first Moran eigenvector (MEM) was used as the independent variable. Environmental data consist of WorldClim bioclimatic variables at a resolution of 30 arc-seconds (Fick and Hijmans 2017). Correlated variables were identified using Pearson's correlation coefficient and removed via the *removeCollinearity* function implemented in the *virtualspecies* R package (Leroy et al. 2016). A redundancy analysis (RDA) was then performed using the R function *capscale* from the *vegan* package (Oksanen et al. 2017). Statistical significance was assessed using 999 permutations. Both R^2 and adjusted $R^2(R^2_{adj})$ were calculated to assess the explanatory power of the model. To test for IBD, the RDA was performed with genetic distances as the response variable and the first MEM as the independent variable. For isolation-by-environment (IBE), uncorrelated and scaled bioclimatic variables were used as the independent variable, while partialing out geographic distance.

RESULTS

Genomic Data and Phylogenetic Relationships

Summary statistics of mitochondrial and genomic data sets are presented in Table 1. For variant discovery, a mean of 546.7 variants and 361.8 SNPs were recovered per exon. Results for variant discovery at every locus (Supplementary Table S2 available on Dryad) and all phylogenetic trees can be obtained from the Supplementary material available on Dryad. Phylogenetic relationships inferred from mitochondrial data differed significantly from those inferred with genomic data, both at the species level (Fig. 1) as well as within the *O. laevis* complex (Fig. 2). Branch support for the mitochondrial phylogeny was largely weak along the backbone of the tree (Fig. 2), but relatively strong close to the tips (Supplementary material available on Dryad). For the genomic data sets, IQ-TREE and ASTRAL-III analyses inferred different topological arrangements among the *O. diminutiva/baluensis, O. lima,* and *O. martensii* clades, but relationships among *O. semipalmata*, *O. sumatrana*, and *O. laevis* were congruent. All genomic IQ-TREE analyses (legacy, exon, intron, and combined) inferred *O. diminutiva/baluensis* as the first branching clade with weak support, followed by *O. lima* and *O. martensii*, respectively (UFB = 87 for exons; UFB = 100 for legacy and combined data sets). The intron data set inferred *O. lima* and *O. martensii* as sister lineages with moderate support (UFB = 91). All ASTRAL summary coalescent trees had identical quartet scores (0.67) and inferred *O. lima* as the first branching lineage with weak support, followed by *O. diminutiva/baluensis* and *O. martensii* respectively with high support. *Occidozyga semipalmata* was not monophyletic, which requires further investigation. Within the *O. laevis* clade, IQ-TREE and ASTRAL analyses inferred congruent topologies across all data sets (Figs. 1–3) and all nodes were highly supported (UFB = 100; ASTRAL local posterior probability, pp = 1.0) except for Clade F in ASTRAL analyses (pp = 0.48). In terms of gCF, branches leading to clades E–J were substantially lower (1.56–7.4) compared to clades A–B (>40) and C+D (20.5; Fig. 2).

Species Delimitation

The mPTP species delimitation analysis on the *O. laevis* complex inferred 20 species that were between 3 and 10% divergent based on uncorrected *p*-distances from the 16S rRNA marker (Fig. 2). Genetic distances calculated based on the Kimura 2-parameter model were congruent but slightly higher overall (Supplementary Fig. S1 available on Dryad). In contrast, the K-means analysis using genomic SNP data only inferred six and eight clusters as determined by the pseudo-F and BIC criteria respectively (Fig. 2). The smaller number of clusters was supported by the sNMF analysis, which inferred an optimal number of 4–7 ancestral populations (Fig. 3). Admixture was detected in Clades B, E, H, and I, whereas Clades A, C, G, and J were largely nonadmixed. In agreement with the K-means and sNMF analyses, the PCA and t-SNE analyses showed that the nonadmixed clades formed four well-separated clusters: A, C, D, I+J (Fig. 4). Both analyses also detected a slight separation within clade G which corresponds to fine-scale phylogenetic and population substructuring that was also inferred from the sNMF analysis at K = 7 (Fig. 3). Clades E, F, and H were clustered together and showed an affinity with either Clade G or Clade I+J. Compared to PCA, the t-SNE analysis was able to reveal more subtle data structures that were consistent with quartet distances and splits inferred from NANUQ (see below).

MANUSCRIPT CATEGORY: Systematic Biology

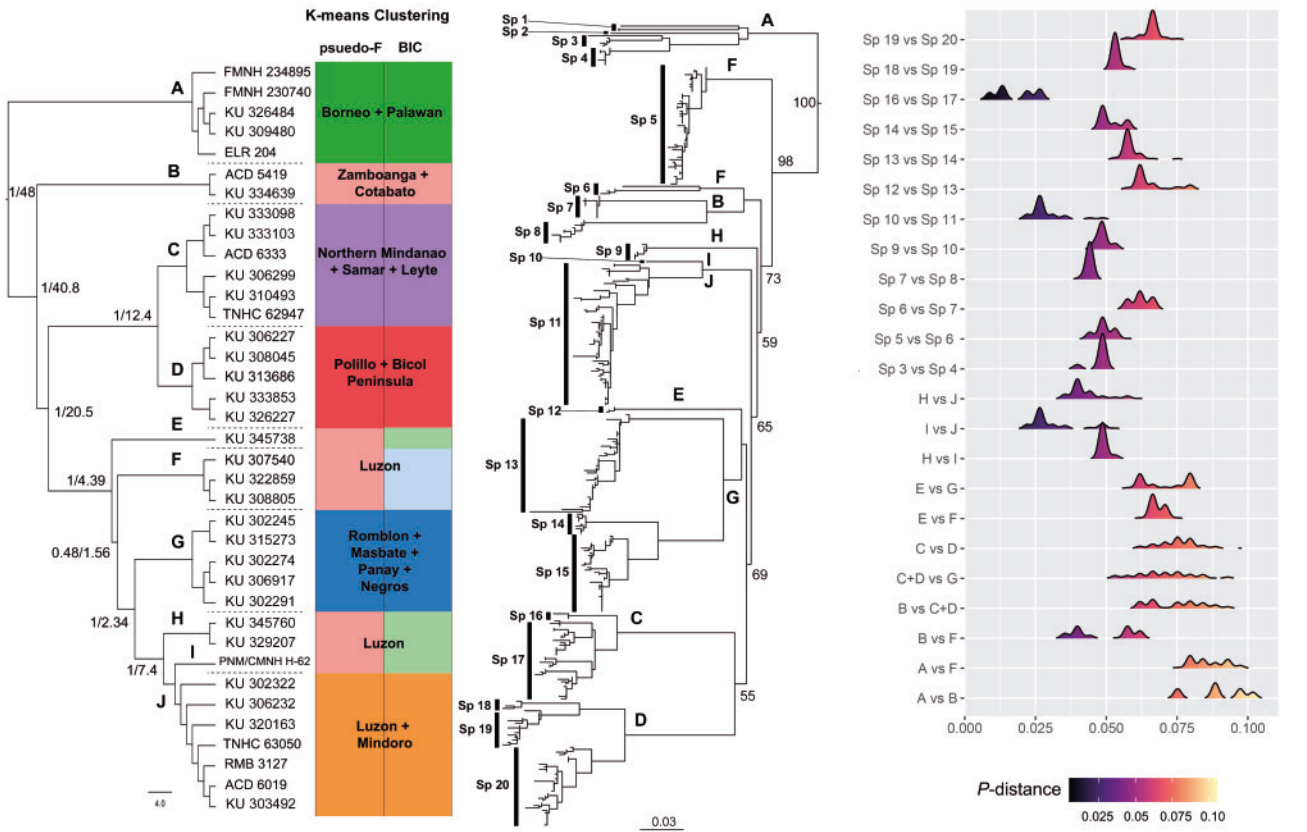

FIGURE 2. Left: ASTRAL phylogeny (combined data set) of the *Occidozyga laevis* complex juxtaposed with results of the K-means clustering analysis using pseudo-F (six clusters: A, B+E+F+I, C, D, G, and J) and Bayesian Information Criterion (eight clusters: A, B, C, D, E+H+I, F, G, J). Major clades of interest are labeled A–J. Numbers along major nodes are posterior probabilities followed by gCF (separated by "/"). Center: corresponding ML phylogeny (16S rRNA) of the *O. laevis* complex showing strongly supported putative species (Spp. 1–20) as inferred from the mPTP species delimitation analysis. Numbers along major nodes are Ultrafast bootstrap support values. Right: Pairwise comparisons of 16S uncorrected p-distances among pairs of closely related clades.

Species Network and Gene Flow

At $\alpha = 0.0001$ and $\beta = 0.95$, the NANUQ quartet hypothesis test showed that most quartet counts concordance factors (qcCFs) were clustered close to the tree model (3 line segments) suggesting that quartets were tree-like with little error. A relatively large portion of qcCFs rejected the tree and star model in support of a 4-cycle network (red triangles; Fig. 5a). This is illustrated by the neighbor-net splits graph that showed high levels of discordance, especially among Clades E, F, G, and H (Fig. 5b–f). The most highly weighted splits grouped Clade A+B+C+D and Clade H+I+J (Fig. 5b). Clade E was grouped with Clade A+B+C+D with high weight (Fig. 5c) but was also grouped with Clade F with lower weight (Fig. 5f). This is consistent with the sNMF analysis that showed Clade E containing alleles from all those clades (Fig. 3). Similarly, Clade G was grouped with Clade H+I+J with high weight (Fig. 5c) but was also inferred to group with Clade A+B with lower weight (Fig. 5f). The t-SNE analysis also showed similar results where Clade G was clustered with Clade A+B along Axis 2 and with Clade H+I+J along Axis 3 (Fig. 4). Clade F was also grouped with either Clade H+I+J and G (Fig. 5d,e) or Clade E (Fig. 5f). The splits graph also showed slight separation within Clades F and G, which was also evinced in the t-SNE analysis.

Based on likelihood scores, a species network comprising seven reticulation events was found to be optimal for the PhyloNet analysis (Fig. 6; Supplementary Fig. S2 available on Dryad). Similar to results from NANUQ, Clade F was shown to be connected to Clade E as well as Clade H+I+J, while Clade E was also connected to Clade C+D (Fig. 6). Two separate introgression events were inferred between Clade G and H+I+J: one was more recent, while an older event involved the ancestor of Clade G. Similarly, Clade B was connected to Clades C+D and A through multiple introgressive events. Introgression between C and D was shown to be recent.

The HyDe analysis at the individual level was able to provide a more nuanced characterization of gene flow. Within Clade D, two individuals from Camarines Sur (KU 306227) and Camarines Norte (KU 313686) were highly admixed ($0.41 < \gamma < 0.77$) with Clade C and Clades E, H, and F (Table 2). These two individuals and another sample from Quezon (KU 333853) showed moderate levels of admixture with Clade C and Clade J

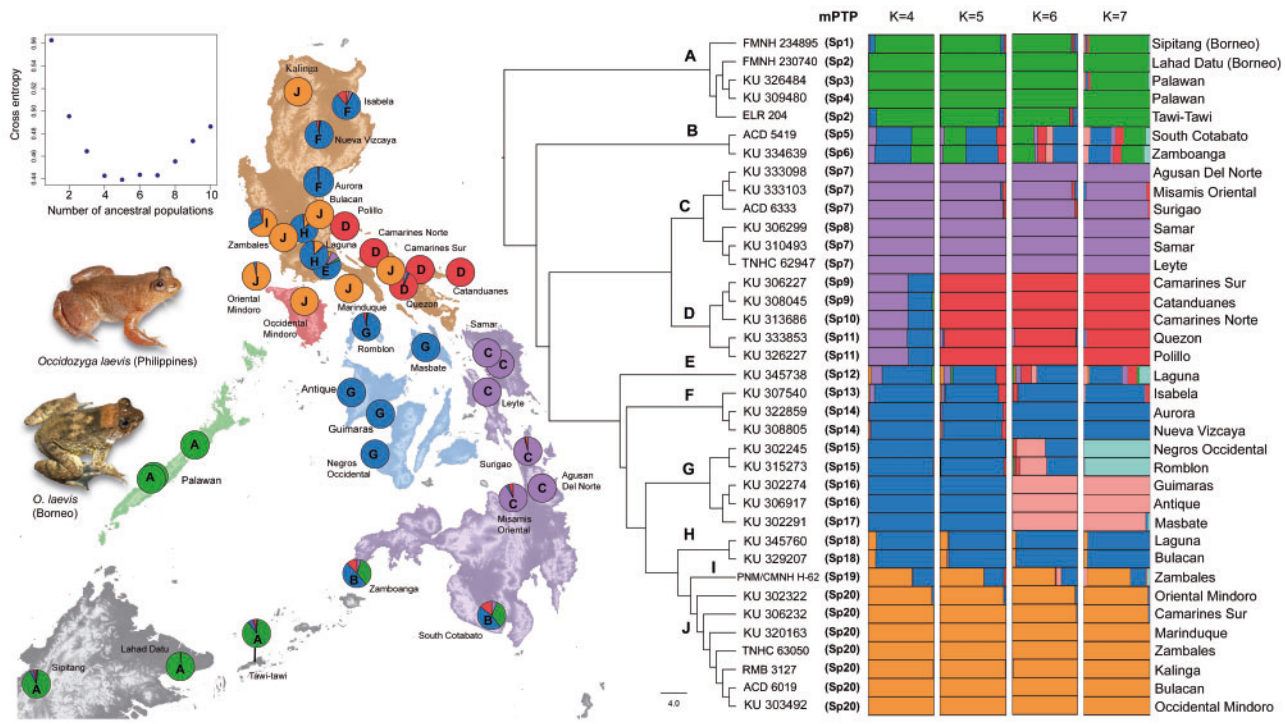

FIGURE 3. Ancestry coefficient bar plots from the sNMF analysis for the *Occidozyga laevis* complex (K = 4–7) juxtaposed with the ASTRAL phylogeny (combined data set) and putative species identified from the mPTP species delimitation analysis. Piecharts of ancestry coefficient proportions of the most optimal K = 5 are spatially visualized on the distribution map and labeled according to the major clades (A–J). The distribution map is shaded according to major Pleistocene Aggregate Island Complexes: orange = Luzon, red = Mindoro, green = Palawan, blue = West Visayan, purple = Mindanao. Crossentropy results for the sNMF analysis are shown in the top left corner and inset images represent topotypic *O. laevis* from population C and a population from Borneo (population A).

($0.13 < \gamma < 0.28$). Interestingly, samples from the islands of Polillo (KU 326227) and Catanduanes (KU 308045) were not admixed. Clade E was moderately admixed between Clade H and J ($\gamma = 0.8$). All individuals in Clade F had high levels of admixture between Clade J and Clade H ($0.42 < \gamma < 0.63$) and moderate to low admixture between Clade D and Clades H and E ($0.73 < \gamma < 0.92$). Clade H had very low levels of admixture between Clades I, E, and F ($0.94 < \gamma < 0.95$). In Clade B, one individual from Zamboanga (KU 334639) was moderately admixed between Clades C and G ($\gamma = 0.24$), whereas the individual from South Cotabato (ACD 5419) showed moderate admixture between Clades C and A ($\gamma = 0.73$). The bootstrap analysis showed that levels of admixture are heterogeneously distributed across individuals and clades (Fig. 6).

IBD and IBE

The dbRDA analysis to test for IBD was significant ($p < 0.01$) indicating a correlation between genetic and geographic distance (Table 3). For the IBE analysis, four of eight uncorrelated bioclimatic variables were significant and interestingly, these four variables are associated with precipitation. However, R^2 and R^2_{adj} were very low for all tested variables indicating high variability (noise) in the data and low explanatory power of the IBD and IBE model.

Relationship between Mitochondrial and Nuclear p-distances

For Clade D, a nonadmixed individual (KU 308045) was selected for *p*-distance calculations. The relationship between mitochondrial and nuclear *p*-distances was strong and significantly correlated (Pearson's correlation coefficient, $R = 0.55$, $P < 0.001$; P value for the linear regression was $P = 0.000$) in clades that had little to no admixture (Clades A+C+D+G+J). In contrast, correlation was weak ($R = 0.17$, $P = 0.025$) in highly admixed clades (Clades B+E+F+H+I; Fig. 4). At $\alpha = 0.01$, separate comparisons of individual clades revealed that all nonadmixed clades showed strong and significant correlations, whereas admixed clades were not correlated (Fig. 7).

DISCUSSION

Impacts of Gene Flow on Phylogenetic Inference and Species Delimitation

In contrast to the substantially higher number of species inferred from a conventional tree- and

MANUSCRIPT CATEGORY: Systematic Biology

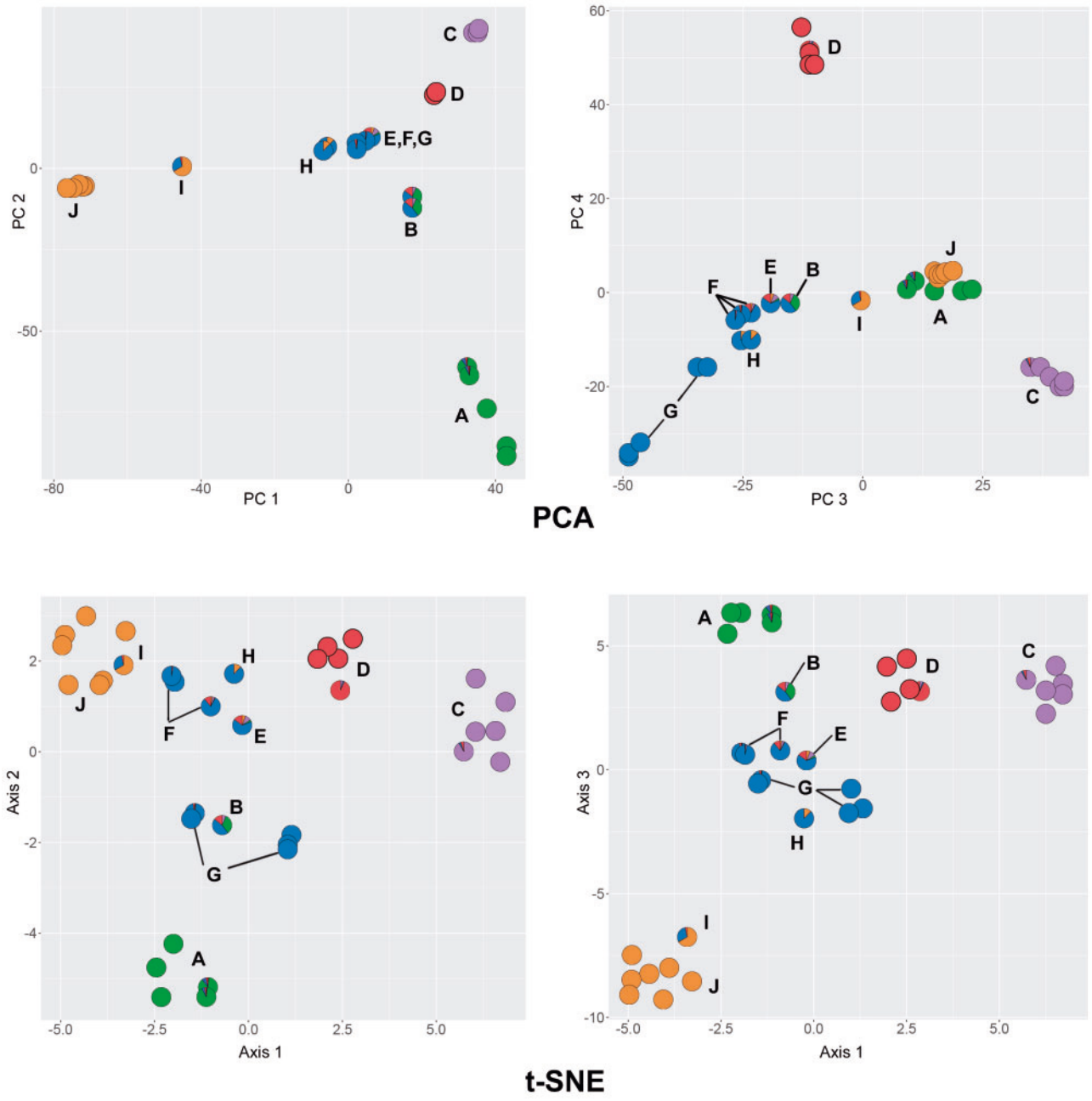

FIGURE 4. Results of the PCA and t-SNE analyses on SNP data with datapoints substituted with piecharts of ancestry coefficients from the sNMF analysis at K = 5.

divergence-based species delimitation approach (up to 20 species), our genomic validation analyses consistently inferred a much lower number of distinct genetic clusters. Our results demonstrate two specific ways in which gene flow can confound phylogenetic inference and species delimitation. First, admixed populations can be erroneously inferred as distinct clades using mitochondrial as well as genomic data. However, by implementing a suite of clustering, population structure, gene flow, and network analyses, we were able to demonstrate that admixed clades were genetically similar and connected to one or multiple parental populations through gene flow. Despite high bootstrap and local posterior probability support, gCF values were markedly lower for the admixed Clades E–I (Fig. 2), indicating high levels of phylogenetic discordance. These results recapitulate and contribute to a growing corpus of literature demonstrating the misleading and inadequate properties of traditional measures of branch support such as bootstrapping when applied to genome-scale data (Kumar et al. 2012; Smith et al. 2015; Roycroft et al. 2019; Chan et al. 2020a). Taken together, we

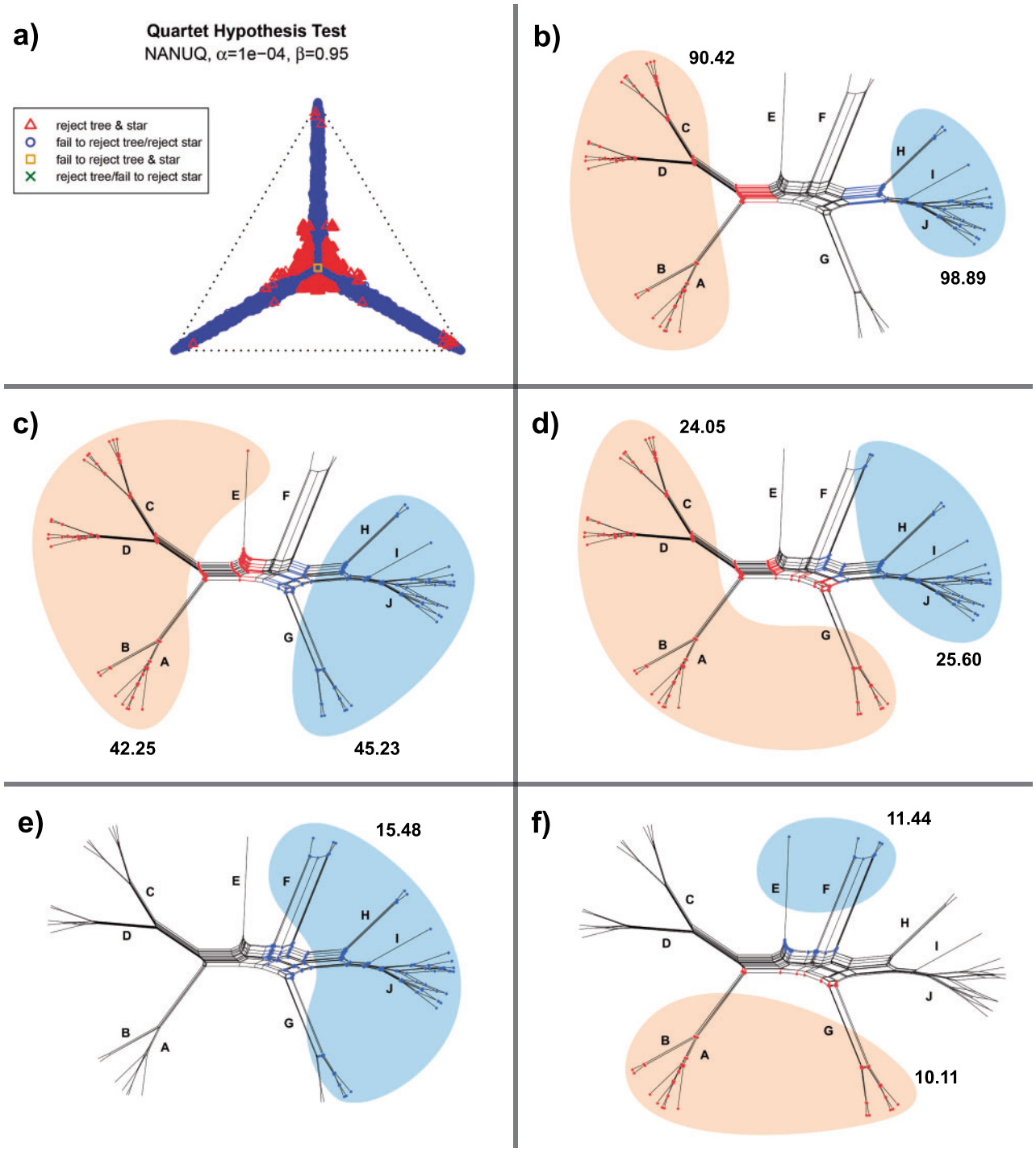

FIGURE 5. Simplex plot of NANUQ's quartet hypothesis test at $\alpha = 0.0001$ and $\gamma = 0.95$ (a); Neighbor-Net splits graphs based on quartet distances (b–f). All graphs are identical, and different colors represent alternative groupings. Numbers represent split weights, which can be interpreted as relative support for the group. Clade labels A–J correspond to major clades of interest (see Figs. 2 and 3). The inclusion of a clade in more than one group is interpreted as admixture.

showed that admixture caused by gene flow can increase phylogenetic structure (and by implication, genetic divergence) that is not entirely consistent with species divergence, and that a single consensus phylogram is unable to accurately depict the genetic affinity of admixed clades to multiple parental populations.

Secondly, admixed lineages can appear to be genetically divergent, even from their parental

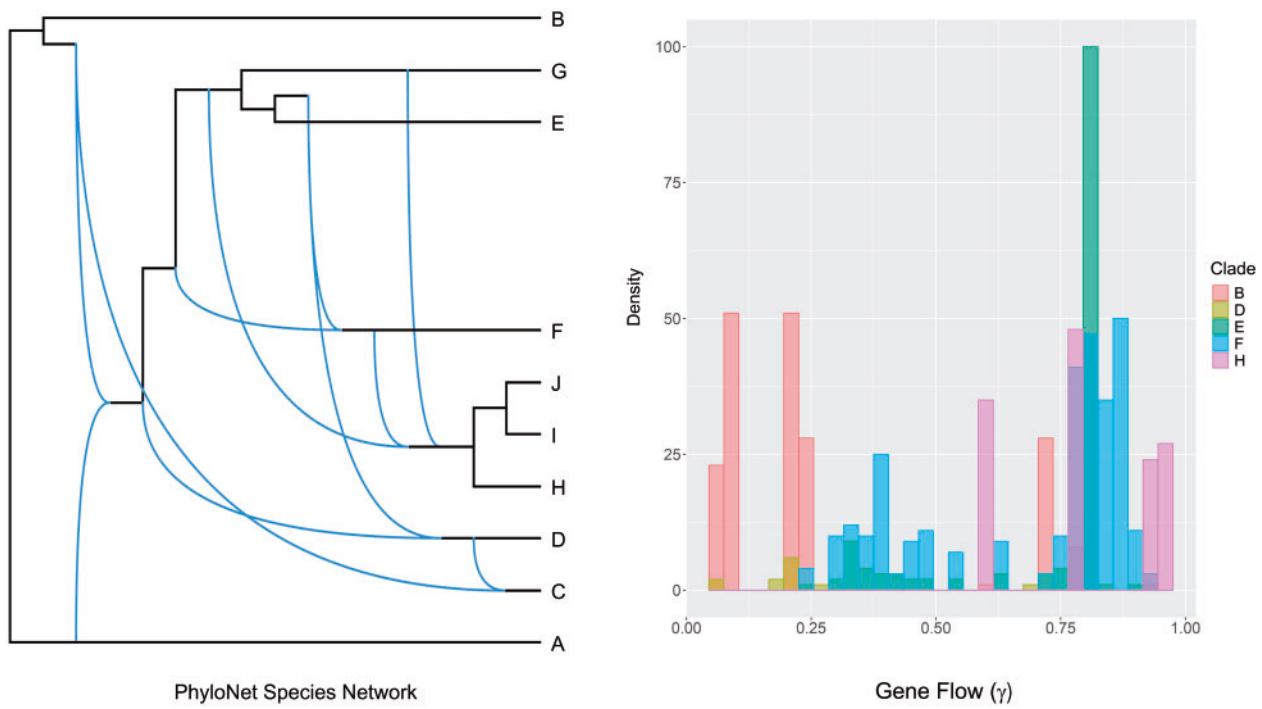

FIGURE 6. Left: Most optimal PhyloNet species network inferred using maximum pseudolikelihood Inference from 3015 intron-derived gene trees. Right: Gene flow estimations (derived from HyDe analysis at the individual level, obtained using 100 bootstrap replicates.

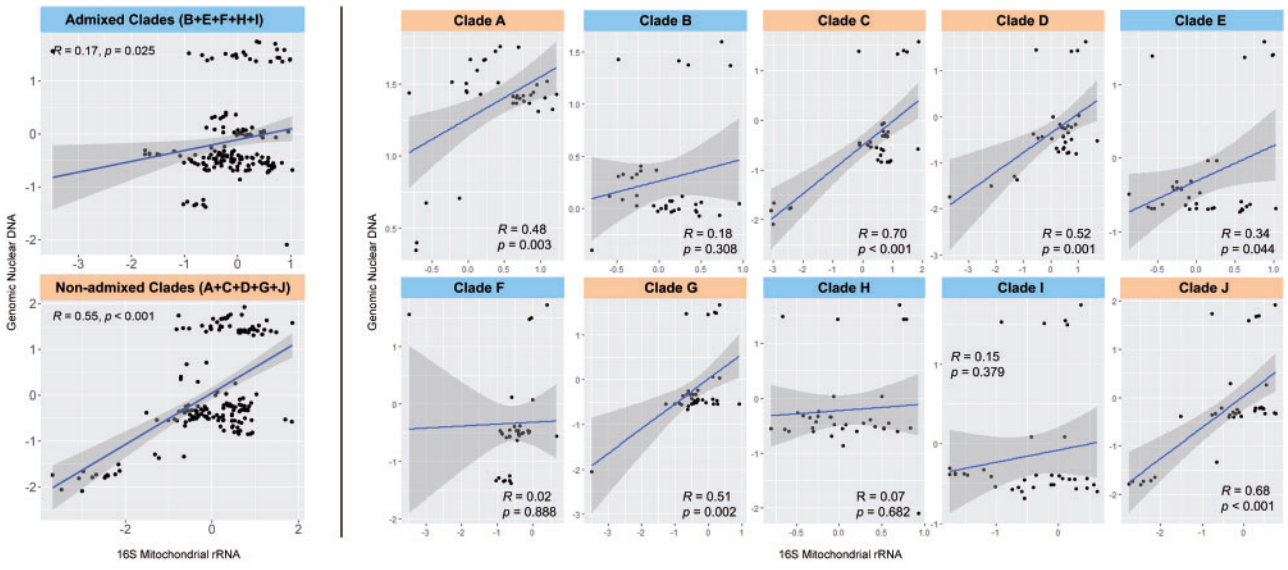

FIGURE 7. Scatter plots of mitochondrial and nuclear p-distances for admixed and nonadmixed clades combined (left) and separated (right). Blueline and gray shading represent the regression line and 95% confidence intervals of a linear model. R and p are the regression coefficient and significance values, respectively. Clade labels A–J correspond to major clades of interest (see Figs. 2 and 3). Clade D is represented by nonadmixed samples.

populations when simple measures of divergence are calculated based on a single mitochondrial gene (Fig. 2). We contend that the issue is not with the use of divergence measures *per se*, but rather, the limited amount of data that can be obtained from a single mitochondrial gene (or a handful of Sanger-derived markers), and the shallow amount of information contained in simple measures such as uncorrected *p*-distances or even model-based distances such as the Kimura 2-parameter model. More importantly,

TABLE 2. Results of the HyDe analysis at the individual level

P1	Hybrid	P2	Z-score	*P* value	Gamma (γ)
C	(D) KU326227	E	11.24662813	0	0.677903622
C	(D) KU313686	E	17.38488215	0	0.584934731
C	(D) KU326227	H	6.002662092	9.74e-10	0.769793995
C	(D) KU313686	H	10.79508126	0	0.675015729
C	(D) KU326227	F	12.63074896	0	0.410136043
C	(D) KU313686	F	14.34479627	0	0.437389205
C	(D) KU333853	J	6.517210175	3.60e-11	0.275268398
C	(D) KU326227	J	4.422054227	4.89e-06	0.130321284
C	(D) KU313686	J	6.67133536	1.27e-11	0.195259231
H	(E) KU345738	J	8.248272379	0	0.814049556
H	(F) RMB6117	D	8.925312189	0	0.90221198
H	(F) KU308805	D	23.01866112	0	0.731866338
H	(F) KU322859	D	15.30642043	0	0.835472497
E	(F) RMB6117	D	13.76687875	0	0.924135234
E	(F) KU308805	D	27.33226215	0	0.740295546
E	(F) KU322859	D	18.68411272	0	0.80612333
J	(F) RMB6117	H	8.732226424	0	0.227277673
J	(F) KU308805	H	8.146439771	2.22e-16	0.631819219
J	(F) KU322859	H	16.90365691	0	0.422869504
I	(H) KU329207	E	6.286835818	1.63e-10	0.955818546
I	(H) KU329207	F	7.302644735	1.42e-13	0.942091794
G	(B) KU334639	C	18.02390945	0	0.237620065
A	(B) ACD5419	C	18.25021737	0	0.729032803

Only significant results are shown ($P < 0.01$). Admixed populations are designated as hybrids (clade name in parenthesis followed by sample ID) and plausible parental populations were selected based on geographic proximity and results from the sNMF analysis. Gamma values represent the relative amount of admixture in the hybrid individual.

TABLE 3. Results of the redundancy analysis to test for isolation-by-distance (IBD) and isolation-by-environment (IBE)

	P value	R^2	R^2_{adj}
IBD			
Geographic distance	0.001	0.01	0.04
IBE			
Annual mean temperature	0.379	0.02695	0.00042
Mean diurnal range	0.23	0.02973	0.00336
Temperature annual range	0.029	0.03935	0.01355
Mean temperature of driest quarter	0.473	0.02582	−0.0008
Annual precipitation	0.001	0.06574	0.04149
Precipitation of wettest month	0.011	0.04364	0.0181
Precipitation of driest quarter	0.001	0.05644	0.03165
Precipitation of coldest quarter	0.001	0.05646	0.03166

Bioclimatic variables were obtained from WorldClim.

for the first time, we demonstrated the decoupling of the relationship between mitochondrial and genome-wide nuclear *p*-distances when admixture is present. This result is significant because genetic distance thresholds derived from short mitochondrial barcodes are routinely used to justify and support the recognition of new species across many taxonomic groups (Vences et al. 2005a; Kvist 2016; Brennan et al. 2017; Lukhtanov 2019; Zhou et al. 2019; Jusoh et al. 2020; Martinsson and Erséus 2021). We showed that while the use of distance-based approaches may be viable when gene flow is absent, it does not accurately reflect divergence patterns in the nuclear genome when gene flow occurs, leading to erroneous estimates of genetic distances and species boundaries. As such, divergence thresholds for species delimitation should be used with caution when the absence of gene flow has not been ruled out. Our results also highlight the potential pitfalls of confidently relying on resolved species trees when gene flow is present (Hahn and Nakhleh 2016). For such cases, we demonstrate that phylogenetic networks can provide a more accurate and holistic evolutionary framework for downstream inferences (Cao et al. 2019). We anticipate that the application of genomic data and more robust analyses will eventually reveal gene flow to be prevalent in many co-occurring or recently diverged cryptic groups, thereby leading to a paradigm shift in our understanding of cryptic species diversity and the use of conventional phylogeny- and distance-based approaches for cryptic species delimitation.

Interestingly, there were some discrepancies in group clustering between nonparametric (K-means, PCA, t-SNE) and model-based methods (sNMF, NANUQ, PhyloNet). Nonparametric methods directly analyze SNP data, and it is conceivable that gene flow affected allele frequencies in a way that can bias clustering

(Francois et al. 2010). This is corroborated by the fact that nonadmixed clades were consistently and accurately clustered, whereas discrepancies only occurred in admixed clades. The impact of gene flow on the efficacy of nonparametric versus model-based methods has not been thoroughly explored; this represents a novel avenue for future research.

Spatial Patterns of Genetic Structure across an Island Archipelago

Despite the prevalence of gene flow across multiple populations, several distinct populations were detected (Clades A, C/D, G/E/F, and J). The distribution of these distinct populations corresponds well with one of the primary predictions of the Pleistocene Aggregate Island Complex model (PAIC; Fig. 3), namely the geographical pattern of lineage distribution and endemism (Inger 1954; Heaney 1985; Brown and Diesmos 2009), coincident with the archipelago's five major landmass amalgamations (geological platforms) and *in situ* diversification on larger islands (Brown et al. 2013).

Within the Luzon PAIC, for example, admixture patterns elucidated here are congruent with previous studies depicting fine-scale geographical patterns of endemism, situated on the various geological components of Luzon (Sanguila et al. 2011; Barley et al. 2013; Brown and Siler 2014; Siler et al. 2014; Brown et al. 2016), which have accreted together in recent geological time scales, and now about one another, in close proximity, on southern Luzon where the most complex patterns of admixture were observed in this study. For example, Lineage F, to date has only been recorded in the Sierra Madre mountain range, a semi-isolated, elongated mountain range of eastern Luzon (Fig. 3), which abuts the Caraballo Mountains of Nueva Viscaya Province, where our only other F individual was collected. Clade D is a Bicol faunal region lineage found at high elevation mountains of the Bicol Peninsula or on the small islands of Polillo and Catanduanes. Although the sNMF analysis at the most optimal K $=5$ did not detect signatures of admixture, the HyDe analysis revealed that samples from Camarines Sur and Camarines Norte in the northern region of the Bicol Peninsula were highly admixed with samples from Clades E, H, and F on Luzon, which is congruent with results from the PhyloNet analysis (Fig. 6). Clade D is derived from a split with Lineage C, which is from the south to northern Samar island, and separated from the southern tip of Luzon and the Bicol faunal region by a single, narrow channel (Fig. 3). Divergence across this southern Luzon–northern Samar interface is not surprising and was presumably facilitated by sea-level vicariance or colonization of Luzon from the south, followed by isolation. Lineage J is found unadmixed at higher elevations in isolated mountains of western Luzon (the Cordillera and Zambales mountains) and on small islands to the south and southwest of Luzon (Marinduque, Mindoro, and Lubang). One plausible interpretation is that this genotype initially evolved in isolation—either in one or more of Western Luzon's isolated mountain ranges or on the deepwater island of Mindoro or Lubang—but has subsequently expanded its range to move into the highly disturbed agricultural matrix of southern Luzon, where it now comes into close contact and found to be admixed with E, H, I, F, and to a lesser degree, D (Table 2) in intermediate localities of Laguna, Bulacan, and Quezon provinces, and nearby southern Zambales. The occurrence and admixture between a sample from Clade J (KU 306232) at moderate elevations on the same mountains (foothills of Mt. Isarog, Camarines Sur) as a sample from Clade D (nearby Municipality of Pili, vicinity of Naga City) (KU 333853; Table 2) exemplifies the ability of these lineages to exchange genes upon contact.

In contrast, patterns of demonstrated south-to-north admixture between the Mindanao and Luzon PAICs necessitates interpretations of over-water dispersal and colonization, which have frequently been detected (Siler et al. 2011; Brown et al. 2013). The same appears likely for gene flow from Mindoro, West Visayan islands, and the Romblon island group in the central portions of the archipelago (Fig. 3). Interestingly, admixture patterns inferred from this study indicate that gene flow only occurred in a northward direction; Luzon-endemic genotypes were not admixed with West Visayan, Mindanao, Palawan, or Sulu/Borneo genomes. Because PAICs are separated by deep marine barriers and have never been connected by land, the most plausible explanation for gene flow among PAICs is overwater dispersal (Brown et al. 2013). We find this explanation to be particularly cogent for the Philippines, which receives an average of 20 typhoons every year, with southern Luzon being the most general area of storm landfall (Cinco et al. 2016). This is further corroborated by multiple studies that have inferred overwater dispersal across Philippine PAICs, many of which include amphibians (Evans et al. 2003; Brown and Diesmos 2009; Siler et al. 2011; Blackburn et al. 2013; Brown et al. 2013; Brown 2016).

Admixture is most evident in southern Luzon (Laguna Province, where mainland Luzon joins the Bicol Peninsula) and southwestern Mindanao (Zamboanga and the Cotobato Coast), alluding to the possible existence of hybrid zones in those areas (Fig. 3). Despite gene flow occurring among multiple populations, our results revealed the existence of several distinct lineages whose occurrence correspond with well-defined biogeographic zones, sutures, and faults (Brown et al. 2013), suggesting that the highly fragmented archipelagic landscape may counteract the homogenizing effects of gene flow, thereby allowing distinct populations to persist on isolated landmasses and even PAICs separated by narrow, but deep, marine channels of the central regions of the archipelago (Welton et al. 2013; Brown et al. 2016). However, the explicit delimitation of these lineages will require denser

sampling and more robust analyses to examine temporal patterns of diversification. Quantifying gene flow and obtaining realistic estimates of beta diversity across these barriers in the remaining, but imperiled, natural habitats of the central Philippines, provides compelling opportunities for future conceptual studies of genomic data and the applied conservation actions that may be derived from them.

Reconceptualizing Cryptic Diversity with Implications on Taxonomy and Conservation

The rise in new cryptic species discoveries across a diverse range of taxon complexes has led to the notion that cryptic species may be prevalent across the tree of life and is responsible for a significant portion of the earth's yet undescribed biodiversity (Beheregaray and Caccone 2007; Bickford et al. 2007; Pfenninger and Schwenk 2007; Adams et al. 2014). Although true cryptic species do undoubtedly occur (Colborn et al. 2001; Brown et al. 2007; Elmer et al. 2007; Qin et al. 2016; Perry et al. 2018; Rosser et al. 2019), our results indicate that its prevalence in nature could be overestimated. The majority of studies that reported high levels of cryptic diversity based their supposition on phylogenetic structure and distance-based analysis. This study provides compelling empirical evidence showing that gene flow can produce diversification patterns that mimic species divergence. In particular, admixture can produce highly pectinate and asymmetrical topologies (e.g., Clades E–H), which is apparent in most cryptic groups (McLeod 2010; Adams et al. 2014; Matsui et al. 2016; Chen et al. 2017). Despite very high levels of discordance (as evinced by concordance factors), such lineages can be strongly supported by conventional measures (e.g., bootstrap values). These nonmonophyletic but genetically cohesive metapopulation lineages can be challenging to resolve taxonomically due to the dissonance between the criteria of monophyly and genetic distinctness. This is exemplified by Clades E and F that are not reciprocally monophyletic with strong support. Robust analyses based on genomic data indicate that they are conspecific, yet lumping them would invoke an unnatural phylogenetic grouping. This taxonomic conundrum highlights the limitation of delineating species based on a single consensus phylogram, even when clades are highly supported by genomic data. Our results add to the rapidly growing number of studies showing that conflicting evolutionary histories can occur within and between loci (Mendes et al. 2019; Smith et al. 2020) and that a small number of genes can heavily influence the species tree topology (Walker et al. 2018). Therefore, a phylogeny-based species delimitation approach can yield erroneous results with spuriously high support when gene flow is present. Moving forward, delineating cryptic species boundaries should necessarily consider gene flow and gene tree variation within the species delimitation framework to avoid taxonomic inflation.

Our results also provide an alternative explanation for the formation of cryptic species—genetically divergent lineages that are not accompanied by morphological differentiation. Gene flow can act as a homogenizing force that reduces phenotypic variation, and simultaneously, as shown by our data, admixed populations can appear as distinct and divergent lineages on a phylogenetic tree. This can result in anomalous patterns such as highly divergent sympatric lineages and genetically similar but parapatric populations (Chan et al. 2020b). These patterns are not uncommon across the tree of life and have been touted as clear evidence of cryptic speciation (Chan et al. 2018; McLeod 2010; Landaverde-González et al. 2017; Dobson et al. 2018; Filippi-Codaccioni et al. 2018). Our study challenges this overly simplistic assumption and empirically demonstrates that gene flow can produce similar patterns.

Despite inferring a substantially lower number of species, our study nevertheless detected several genetically distinct cryptic lineages, thereby demonstrating that cryptic species do indeed occur. The importance of recognizing cryptic species, particularly for conservation is widely acknowledged (Trontelj and Fišer 2009; Delić et al. 2017). Although we do not question its importance, we argue that over-recognizing species diversity can be equally as detrimental to conservation (Folt et al. 2019). Oversplitting of existing species can lead to an increase in the number of endangered species due to a reduction in the distribution range of the associated species. The abundance and area occupied by each newly delimited species will inherently be a subset of its progenitor species. As widely accepted classifications such as the IUCN Red List are in part predicated on both abundance and range, the number of threatened species can be expected to increase, leading to a concomitant increase in devoted resources and education (Agapow et al. 2004).

This study demonstrates the specific mechanisms in which widely adopted species discovery approaches can inflate diversity estimates. These methods can be applied to large data sets and are relatively cheap and easier to generate. Hence, they are well-suited for identifying candidate species for future validation testing. Nevertheless, moving forward, delineating cryptic species boundaries should necessarily consider gene flow and gene tree variation within the species delimitation framework to avoid taxonomic inflation. For the *O. laevis* complex, denser geographic sampling in the southern island of Mindanao and the northern island of Luzon would be needed to provide better estimates of gene flow. Characterizing gene flow within a temporal context would also aid in assessing the integrity and durability of species boundaries through evolutionary time (Singhal et al. 2018).

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.34tmpg4j1.

Acknowledgments

We thank Paul Hime, Rich Glor, Jim McGuire, Frank Burbrink, an anonymous reviewer, and the participants of KU Biodiversity Institute's Molecular Genomics Discussion Group, as well as UC Berkeley's MVZ Herpetology Group, for feedback on analyses and interpretation, and for constructive reviews of previous drafts of this manuscript. Awards from the University of Kansas Provost's Research Investment Council (RIC Level II Award No. 2300207 to R.M.B. and R. G. Moyle), and KU's Docking Scholar Fund (to R.M.B.) supported library preparation and sequence capture, and KU's Genome Sequencing Core (to C.R.H. and R.M.B.) provided funds for genomic sequencing. We thank U.S. National Science Foundation support to CRH (GRF 1540502) during the development of FrogCap, and to RMB during field work over the last two decades (DEB 1654388, 1557053, and 0743491). We thank J. McGuire (University of California, Berkeley), T. LaDuc, D. Cannatella (University of Texas, Austin), J. Vindum (California Academy of Sciences), A. Resetar, H. Voris, and R. Inger (Field Museum) for access to genetic resources and the Philippine Department of the Environment and Natural Resources, Biodiversity Management Bureau for research, collection, and export permits (to R.M.B. at KU; Animal handling protocols approved by KU IACUC 185-05) necessary for this and related studies. The contributions of P.L.W. to this article constitute contribution number 933 of the Auburn University Museum of Natural History. Thanks to C. Siler, K. Cobb, and A. Diesmos for assistance collecting specimens and preliminary data.

References

Adams M., Raadik T.A., Burridge C.P., Georges A. 2014. Global biodiversity assessment and hyper-cryptic species complexes: More than one species of elephant in the room? Syst. Biol. 63:518–533.

Agapow P.M., Bininda-Emonds O.R.P., Crandall K.A., Gittleman J.L., Mace G.M., Marshall J.C., Purvis A. 2004. The impact of species concept on biodiversity studies. Q. Rev. Biol. 79:161–179.

Allman E.S., Baños H., Rhodes J.A. 2019. NANUQ: A method for inferring species networks from gene trees under the coalescent model. Algorithms Mol. Biol. 14:1–25.

Amador L., Parada A., D'Elía G., Guayasamin J.M. 2018. Uncovering hidden specific diversity of Andean glassfrogs of the *Centrolene buckleyi* species complex (Anura: Centrolenidae). PeerJ. 6:e5856.

Van der Auwera G.A., Carneiro M.O., Hartl C., Poplin R., del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., Banks E., Garimella K. V., Altshuler D., Gabriel S., DePristo M.A. 2013. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43:11.10.1-11.10.33.

Barley A.J., Monnahan P.J., Thomson R.C., Grismer L.L., Brown R.M. 2015. Sun skink landscape genomics: assessing the roles of microevolutionary processes in shaping genetic and phenotypic diversity across a heterogeneous and fragmented landscape. Mol. Ecol. 24:1696–1712.

Barley A.J., White J., Diesmos A.C., Brown R.M. 2013. The challenge of species delimitation at the extremes: diversification without morphological change in Philippine Sun Skinks. Evolution (NY) 67:3556–3572.

Beheregaray L.B., Caccone A. 2007. Cryptic biodiversity in a changing world. J. Biol. 6:9.

Bickford D., Lohman D.J., Sodhi N.S., Ng P.K.L., Meier R., Winker K., Ingram K.K., Das I. 2007. Cryptic species as a window on diversity and conservation. Trends Ecol. Evol. 22:148–155.

Blackburn D.C., Giribet G., Soltis D.E., Stanley E.L. 2019. Predicting the impact of describing new species on phylogenetic patterns. Integr. Org. Biol. 1:1–12.

Blackburn D.C., Siler C.D., Diesmos A.C., Mcguire J.A., Cannatella D.C., Brown R.M. 2013. An adaptive radiation of frogs in a southeast asian island archipelago. Evolution (NY). 67:2631–2646.

Blair C., Bryson R.W. 2017. Cryptic diversity and discordance in single-locus species delimitation methods within horned lizards (Phrynosomatidae: *Phrynosoma*). Mol. Ecol. Resour. 17:1168–1182.

Blischak P.D., Chifman J., Wolfe A.D., Kubatko L.S. 2018. HyDe: a python package for genome-scale hybridization detection. Syst. Biol. 67:821–829.

Bogisich A. 2019. Frogs Hiding in plain sight: phylogenetic systematics of Myanmar's Occidozyga species complex, and the identification of a novel species [Masters Theses] 1208.

Bradburd G.S., Ralph P.L. 2019. Spatial population genetics: it's about time. Annu. Rev. Ecol. Evol. Syst. 50:427–449.

Brennan I.G., Bauer A.M., Van Tri N., Wang Y.Y., Wang W.Z., Zhang Y.P., Murphy R.W. 2017. Barcoding utility in a mega-diverse, cross-continental genus: keeping pace with *Cyrtodactylus* geckos. Sci. Rep. 7:1–11.

Brown D.M., Brenneman R.A., Koepfli K.P., Pollinger J.P., Milá B., Georgiadis N.J., Louis E.E., Grether G.F., Jacobs D.K., Wayne R.K. 2007. Extensive population genetic structure in the giraffe. BMC Biol. 5:1–13.

Brown R.M. 2016. Biogeography of land vertebrates. In: Kliman R.M., editor. Encyclopedia of evolutionary biology. Oxford: Academic Press/Elsevier Inc. p. 211–220.

Brown R.M., Diesmos A.C. 2009. Philippines, Biology. In: Gillespie R., Clague D., editors. Encyclopedia of islands. Berkeley: University of California Press. p. 723–732.

Brown R.M., Siler C.D. 2014. Spotted stream frog diversification at the Australasian faunal zone interface, mainland versus island comparisons, and a test of the Philippine "dual-umbilicus" hypothesis. J. Biogeogr. 41:182–195.

Brown R.M., Siler C.D., Oliveros C.H., Esselstyn J.A., Diesmos A.C., Hosner P.A., Linkem C.W., Barley A.J., Oaks J.R., Sanguila M.B., Welton L.J., Blackburn D.C., Moyle R.G., Townsend Peterson A., Alcala A.C. 2013. Evolutionary processes of diversification in a model island archipelago. Annu. Rev. Ecol. Evol. Syst. 44:411–435.

Brown R.M., Su Y.C., Barger B., Siler C.D., Sanguila M.B., Diesmos A.C., Blackburn D.C. 2016. Phylogeny of the island archipelago frog genus *Sanguirana*: another endemic Philippine radiation that diversified "Out-of-Palawan." Mol. Phylogenet. Evol. 94:531–536.

Cao Z., Liu X., Ogilvie H.A., Yan Z., Nakhleh L. 2019. Practical aspects of phylogenetic network analysis using PhyloNet. bioRxiv.:746362.

Capella-Gutiérrez S., Silla-martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Carstens B.C., Morales A.E., Jackson N.D., O'Meara B.C. 2017. Objective choice of phylogeographic models. Mol. Phylogenet. Evol. 116:136–140.

Cerca J., Meyer C., Purschke G., Struck T.H. 2020. Delimitation of cryptic species drastically reduces the geographical ranges of marine interstitial ghost-worms (Stygocapitella; Annelida, Sedentaria). Mol. Phylogenet. Evol. 143:106663.

Chan H.K. 2013. Phylogeography and cryptic diversity of Occidozyga lima (Gravenhorst 1829) [Thesis]. University of Hong Kong, Pokfulam, Hong Kong SAR. b5060583.

Chan K.O., Alexander A.M., Grismer L.L., Su Y.-C., Grismer J.L., Quah E.S.H., Brown R.M. 2017. Species delimitation with gene flow: a methodological comparison and population genomics approach to elucidate cryptic species boundaries in Malaysian Torrent Frogs. Mol. Ecol. 26:5435–5450.

Chan K.O., Grismer L.L. 2019. To split or not to split? Multilocus phylogeny and molecular species delimitation of southeast Asian toads (family: Bufonidae). BMC Evol. Biol. 19:95.

Chan K.O., Grismer L.L., Brown R.M. 2018. Comprehensive multilocus phylogeny of Old World tree frogs (Anura: Rhacophoridae) reveals taxonomic uncertainties and potential cases of over- and underestimation of species diversity. Mol. Phylogenet. Evol. 127:1010–1019.

Chan K.O., Hutter C.R., Wood P.L., Grismer L.L., Brown R.M. 2020a. Larger, unfiltered datasets are more effective at resolving phylogenetic conflict: introns, exons, and UCEs resolve ambiguities in Golden-backed frogs (Anura: Ranidae; genus *Hylarana*). Mol. Phylogenet. Evol. 151:106899.

Chan K.O., Hutter C.R., Wood P.L., Grismer L.L., Das I., Brown R.M. 2020b. Gene flow creates a mirage of cryptic species in a Southeast Asian spotted stream frog complex. Mol. Ecol. 29:3970–3987.

Chan K.O., Schoppe S., Rico E.L.B., Brown R.M. 2021. Molecular systematic investigation of Philippine Puddle Frogs (Anura: Dicroglossidae: *Occidozyga* Kuhl and Van Hasselt 1822) reveal new candidate species and a novel pattern of species dyads. Philipp. J. Syst. Biol. 14:1–14.

Chen J.M., Zhou W.W., Poyarkov N.A., Stuart B.L., Brown R.M., Lathrop A., Wang Y.Y., Yuan Z.Y., Jiang K., Hou M., Chen H.M., Suwannapoom C., Nguyen S.N., Duong T. Van, Papenfuss T.J., Murphy R.W., Zhang Y.P., Che J. 2017. A novel multilocus phylogenetic estimation reveals unrecognized diversity in Asian horned toads, genus *Megophrys sensu lato* (Anura: Megophryidae). Mol. Phylogenet. Evol. 106:28–43.

Chen S., Qing J., Liu Z., Liu Y., Tang M., Murphy R.W., Pu Y., Wang X., Tang K., Guo K., Jiang X., Liu S. 2020. Multilocus phylogeny and cryptic diversity of white-toothed shrews (Mammalia, Eulipotyphla, *Crocidura*) in China. BMC Evol. Biol. 20:29.

Cinco T.A., de Guzman R.G., Ortiz A.M.D., Delfino R.J.P., Lasco R.D., Hilario F.D., Juanillo E.L., Barba R., Ares E.D. 2016. Observed trends and impacts of tropical cyclones in the Philippines. Int. J. Climatol. 36:4638–4650.

Colborn J., Crabtree R.E., Shaklee J.B., Pfeiler E., Bowen B.W. 2001. The evolutionary enigma of bonefishes (*albula* spp.): cryptic species and ancient separations in a globally distributed shorefish. Evolution (NY). 55:807–820.

Crivellaro M.S., Zimmermann B.L., Bartholomei-Santos M.L., Crandall K.A., Pérez-Losada M., Bond-Buckup G., Santos S. 2018. Looks can be deceiving: species delimitation reveals hidden diversity in the freshwater crab *Aegla longirostri* (Decapoda: Anomura). Zool. J. Linn. Soc. 182:24–37.

Davis H.R., Chan K.O., Das I., Brennan I.G., Karin B.R., Jackman T.R., Brown R.M., Iskandar D.T., Nashriq I., Grismer L.L., Bauer A.M. 2020. Multilocus phylogeny of Bornean Bent-Toed Geckos (Gekkonidae: *Cyrtodactylus*) reveals hidden diversity, taxonomic disarray, and novel biogeographic patterns. Mol. Phylogenet. Evol. 147:106785.

Delić T., Trontelj P., Rendoš M., Fišer C. 2017. The importance of naming cryptic species and the conservation of endemic subterranean amphipods. Sci. Rep. 7:3391.

Demos T.C., Webala P.W., Bartonjo M., Patterson B.D. 2018. Hidden diversity of African yellow house bats (Vespertilionidae, *Scotophilus*): insights from multilocus phylogenetics and lineage delimitation. Front. Ecol. Evol. 6:86.

Derkarabetian S., Castillo S., Koo P.K., Ovchinnikov S., Hedin M. 2019. A demonstration of unsupervised machine learning in species delimitation. Mol. Phylogenet. Evol. 139:106562.

Díaz-Tapia P., Ly M., Verbruggen H. 2020. Extensive cryptic diversity in the widely distributed *Polysiphonia scopulorum* (Rhodomelaceae, Rhodophyta): molecular species delimitation and morphometric analyses. Mol. Phylogenet. Evol. 152:106909.

Divya P.R., Mohitha C., Rahul G.K., Rajool Shanis C.P., Basheer V.S., Gopalakrishnan A. 2017. Molecular based phylogenetic species recognition in the genus *Pampus* (Perciformes: Stromateidae) reveals hidden diversity in the Indian Ocean. Mol. Phylogenet. Evol. 109:240–245.

Dobson C.M., Howarth M.A., Redfield C. 2018. Molecular evidence for new sympatric cryptic species of *Aedes albopictus* (Diptera: Culicidae) in China: a new threat from *Aedes albopictus* subgroup? Parasit. Vectors 11:228.

Elmer K.R., Dávila J.A., Lougheed S.C. 2007. Cryptic diversity and deep divergence in an upper Amazonian leaflitter frog, *Eleutherodactylus ockendeni*. BMC Evol. Biol. 7:1–14.

Evans B.J., Brown R.M., McGuire J.A., Supriatna J., Andayani N., Diesmos A., Iskandar D., Melnick D.J., Cannatella D.C. 2003. Phylogenetics of fanged frogs: testing biogeographical hypotheses at the interface of the Asian and Australian faunal zones. Syst. Biol. 52:794–819.

Feng Y.-J., Blackburn D.C., Liang D., Hillis D.M., Wake D.B., Cannatella D.C., Zhang P. 2017. Phylogenomics reveals rapid, simultaneous diversification of three major clades of Gondwanan frogs at the Cretaceous–Paleogene boundary. Proc. Natl. Acad. Sci. 114:E5864–E5870.

Fick S.E., Hijmans R.J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 4315:4302–4315.

Filippi-Codaccioni O., Beugin M.P., De Vienne D.M., Portanier E., Fouchet D., Kaerle C., Muselet L., Queney G., Petit E.J., Regis C., Pons J.B., Pontier D. 2018. Coexistence of two sympatric cryptic bat species in French Guiana: insights from genetic, acoustic and ecological data. BMC Evol. Biol. 18:175.

Folt B., Bauder J., Spear S., Stevenson D., Hoffman M., Oaks J.R., Wood P.L., Jenkins C., Steen D.A., Guyer C. 2019. Taxonomic and conservation implications of population genetic admixture, mito-nuclear discordance, and male-biased dispersal of a large endangered snake, *Drymarchon couperi*. PLoS One 14:1–21.

Fouquet A., Gilles A., Vences M., Marty C., Blanc M., Gemmell N.J. 2007. Underestimation of species richness in neotropical frogs revealed by mtDNA analyses. PLoS One 2:e1109.

Francois O., Currat M., Ray N., Han E., Excoffier L., Novembre J. 2010. Principal component analysis under population genetic models of range expansion and admixture. Mol. Biol. Evol. 27:1257–1268.

Frichot E., François O. 2015. LEA: an R package for landscape and ecological association studies. Methods Ecol. Evol. 6: 925–929.

Frichot E., Mathieu F., Trouillon T., Bouchard G., François O. 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics 196:973–983.

Fujisawa T., Barraclough T.G. 2013. Delimiting species using single-locus data and the generalized mixed yule coalescent approach: a revised method and evaluation on simulated data sets. Syst. Biol. 62:707–724.

Grismer L.L., Wood P.L., Anuar S., Muin M.A., Quah E.S.H., McGuire J.A., Brown R.M., Ngo V.T., Pham H.T., Van Tri N., Hong Thai P., Ngo V.T., Hong Thai P., Pham H.T. 2013. Integrative taxonomy uncovers high levels of cryptic species diversity in *Hemiphyllodactylus* Bleeker, 1860 (Squamata: Gekkonidae) and the description of a new species from Peninsular Malaysia. Zool. J. Linn. Soc. 169:849–880.

Grosse M., Bakken T., Nygren A., Kongsrud J.A., Capa M. 2020. Species delimitation analyses of NE Atlantic *Chaetozone* (Annelida, Cirratulidae) reveals hidden diversity among a common and abundant marine annelid. Mol. Phylogenet. Evol. 149:106852.

Hahn M.W., Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution (NY). 70:7–17.

Heaney L.R. 1985. Zoogeographic evidence for middle and late Pleistocene land bridges to the Philippine Islands. Mod. Quat. Res. Southeast Asia. 9:127–143.

Hillis D.M. 2019. Species delimitation in herpetology. J. Herpetol. 53:3–12.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Le S.V. 2017. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518–522.

Huang J. 2020. Is population subdivision different from speciation? From phylogeography to species delimitation. Ecol. Evol. 1–7.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Hutter C.R., Cobb K.A., Portik D.M., Travers S.L., Wood P.L., Brown R.M. 2019. FrogCap: a modular sequence capture probe set for phylogenomics and population genetics for all frogs, assessed across multiple phylogenetic scales. bioRxiv. 825307.

Inger R.F. 1954. Systematics and zoogeography of Phillipine amphibia. Fieldiana (Zool.). 33:183–531.

Inger R.F. 1966. The systematics and zoogeography of the Amphibia of Borneo. Fieldiana (Zool.). 52:1–402.
Iskandar D.T. 1998. The Amphibians of Java and Bali. Bogor, Indonesia: Research and Development Centre for Biology—LIPI and GEF Biodiviersity Collections Project.
Jackson N.D., Carstens B.C., Morales A.E., O'Meara B.C. 2017. Species delimitation with gene flow. Syst. Biol. 66:799–812.
de Jesus P.B., Costa A.L., Nunes J.M. de C., Manghisi A., Genovese G., Morabito M., Schnadelbach A.S. 2019. Species delimitation methods reveal cryptic diversity in the *Hypnea cornuta* complex (Cystocloniaceae, Rhodophyta). Eur. J. Phycol. 54:135–153.
Jiao X., Flouri T., Rannala B., Yang Z. 2020. The impact of cross-species gene flow on species tree estimation. Syst. Biol. 69:830–847.
Jombart T., Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27:3070–3071.
Jones G., Aydin Z., Oxelman B. 2015. DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics 31:991–998.
Jusoh W.F.A., Ballantyne L., Chan K.O. 2020. DNA-based species delimitation reveals cryptic and incipient species in synchronous flashing fireflies (Coleoptera: Lampyridae) of Southeast Asia. Biol. J. Linn. Soc. 130:520–532.
Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 14:587–589.
Kapli P., Lutteropp S., Zhang J., Kobert K., Pavlidis P., Stamatakis A., Flouri T. 2017. Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. Bioinformatics 33:1630–1638.
Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–9.
Kon T., Yoshino T., Mukai T., Nishida M. 2007. DNA sequences identify numerous cryptic species of the vertebrate: a lesson from the gobioid fish *Schindleria*. Mol. Phylogenet. Evol. 44:53–62.
Krijthe J.H. 2015. Rtsne: T-distributed stochastic neighbor embedding using a Barnes-Hut implementation. Available from: https//github.com/jkrijthe/Rtsne.
Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.
Kumar S., Stecher G., Li M., Knyaz C., Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35:1547–1549.
Kvist S. 2016. Does a global DNA barcoding gap exist in Annelida? Mitochondrial DNA. Part A, DNA mapping, Seq. Anal. 27:2241–2252.
Landaverde-González P., Moo-Valle H., Murray T.E., Paxton R.J., Quezada-Euán J.J.G., Husemann M. 2017. Sympatric lineage divergence in cryptic Neotropical sweat bees (Hymenoptera: Halictidae: *Lasioglossum*). Org. Divers. Evol. 17:251–265.
Leaché A.D., Zhu T., Rannala B., Yang Z. 2019. The spectre of too many species. Syst. Biol. 68:168–181.
Legendre P., Fortin M.-J., Borcard D. 2015. Should the Mantel test be used in spatial analysis? Methods Ecol. Evol. 6:1239–1247.
Leroy B., Meynard C.N., Bellard C., Courchamp F. 2016. virtualspecies, an R package to generate virtual species distributions. Ecography (Cop.). 39:599–607.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 1303.3997v.
Liu H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.
Li W., Cerise J.E., Yang Y., Han H. 2017. Application of t-SNE to human genetic data. J. Bioinform. Comput. Biol. 15:1750017.
Liu L., Anderson C., Pearl D., Edwards S. V. 2019. Modern phylogenomics: building phylogenetic trees using the multispecies coalescent model. In: Anisimova M., editor. Evolutionary genomics. Methods in molecular biology. New York: Humana. p. 211–239.
Lukhtanov V.A. 2019. Species delimitation and analysis of cryptic species diversity in the XXI century. Entomol. Rev. 99:463–472.
van der Maaten L., Hinton G. 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9:2579–2605.
Manthey J.D., Klicka J., Spellman G.M. 2011. Cryptic diversity in a widespread North American songbird: phylogeography of the brown creeper (*Certhia americana*). Mol. Phylogenet. Evol. 58:502–512.
Martinsson S., Erséus C. 2021. Cryptic clitellata: molecular species delimitation of clitellate worms (annelida): an overview. Diversity 13:1–13.
Matsui M., Kuraishi N., Eto K., Hamidy A., Nishikawa K., Shimada T., Yambun P., Vairappan C.S., Hossman M.Y. Bin. 2016. Unusually high genetic diversity in the Bornean *Limnonectes kuhlii*-like fanged frogs (Anura: Dicroglossidae). Mol. Phylogenet. Evol. 102:305–319.
McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.
McLeod D.S. 2010. Of Least Concern? Systematics of a cryptic species complex: *Limnonectes kuhlii* (Amphibia: Anura: Dicroglossidae). Mol. Phylogenet. Evol. 56:991–1000.
Meirmans P.G. 2020. GENODIVE version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids. Mol. Ecol. Resour. 00:1–6.
Mendes F.K., Livera A., Hahn M.W. 2019. The perils of intralocus recombination for inferences of molecular convergence. Philos. Trans. R. Soc. B. 374:20180244.
Mendes J., Salvi D., Harris D.J., Els J., Carranza S. 2018. Hidden in the Arabian Mountains: Multilocus phylogeny reveals cryptic diversity in the endemic *Omanosaura* lizards. J. Zool. Syst. Evol. Res. 56:395–407.
Mignotte A., Garros C., Gardès L., Balenghien T., Duhayon M., Rakotoarivony I., Tabourin L., Poujol L., Mathieu B., Ibañez-Justicia A., Deniz A., Cvetkovikj A., Purse B. V., Ramilo D.W., Stougiou D., Werner D., Pudar D., Petrić D., Veronesi E., Jacobs F., Kampen H., Pereira Da Fonseca I., Lucientes J., Navarro J., De La Puente J.M., Stefanovska J., Searle K.R., Khallaayoune K., Culverwell C.L., Larska M., Bourquia M., Goffredo M., Bisia M., England M., Robin M., Quaglia M., Miranda-Chueca M.Á., Bødker R., Estrada-Penã R., Carpenter S., Tchakarova S., Boutsini S., Sviland S., Schäfer S.M., Ozolina Z., Seglina Z., Vatansever Z., Huber K. 2020. The tree that hides the forest: cryptic diversity and phylogenetic relationships in the Palaearctic vector *Obsoletus/Scoticus* complex (Diptera: Ceratopogonidae) at the European level. Parasit. Vectors 13:265.
Minh B.Q., Hahn M.W., Lanfear R. 2020. New Methods to calculate concordance factors for phylogenomic datasets. Mol. Biol. Evol. 37:2727–2733.
Morales A.E., Carstens B.C. 2018. Evidence that *Myotis lucifugus* "subspecies" are five nonsister species, despite gene flow. Syst. Biol. 67:756–769.
Nguyen L.T., Schmidt H.A., Von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.
Nishikawa K., Matsui M., Yong H. Sen, Ahmad N., Yambun P., Belabut D.M., Sudin A., Hamidy A., Orlov N.L., Ota H., Yoshikawa N., Tominaga A., Shimada T. 2012. Molecular phylogeny and biogeography of caecilians from Southeast Asia (Amphibia, Gymnophiona, Ichthyophiidae), with special reference to high cryptic species diversity in Sundaland. Mol. Phylogenet. Evol. 63:714–723.
Ohler A. 2003. Comments on a new book on the Amphibia of Thailand, with a tentiative allocation of the figured species. Alytes 21:100–102.
Oksanen J., Blanchet F.G., M. F., Kindt R., Legendre P., McGlinn D., Minchin P.R., O'Hara R.B., Simpson G.L., Solymos P., Henry M., Stevens H., Szoecs E., Wagner H. 2017. Vegan: community ecology package. Version 2.4-4. R package.

Perry K.D., Baker G.J., Powis K.J., Kent J.K., Ward C.M., Baxter S.W. 2018. Cryptic *Plutella* species show deep divergence despite the capacity to hybridize. BMC Evol. Biol. 18:1–17.

Pfenninger M., Schwenk K. 2007. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. BMC Evol. Biol. 7:1–6.

Postaire B., Magalon H., Bourmaud C.A.F., Bruggemann J.H. 2016. Molecular species delimitation methods and population genetics data reveal extensive lineage diversity and cryptic species in Aglaopheniidae (Hydrozoa). Mol. Phylogenet. Evol. 105:36–49.

Pozzi L., Penna A., Bearder S.K., Karlsson J., Perkin A., Disotell T.R. 2020. Cryptic diversity and species boundaries within the *Paragalago zanzibaricus* species complex. Mol. Phylogenet. Evol. 150:106887.

Puillandre N., Modica M. V., Zhang Y., Sirovich L., Boisselier M.C., Cruaud C., Holford M., Samadi S. 2012. Large-scale species delimitation method for hyperdiverse groups. Mol. Ecol. 21:2671–2691.

Qin L., Pan L.L., Liu S.S. 2016. Further insight into reproductive incompatibility between putative cryptic species of the *Bemisia tabaci* whitefly complex. Insect Sci. 23:215–224.

R Core Team. 2014. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rosser N., Freitas A.V.L., Huertas B., Joron M., Lamas G., Mérot C., Simpson F., Willmott K.R., Mallet J., Dasmahapatra K.K. 2019. Cryptic speciation associated with geographic and ecological divergence in two Amazonian *Heliconius* butterflies. Zool. J. Linn. Soc. 186:233–249.

Rowley J.J.L., Tran D.T.A., Frankham G.J., Dekker A.H., Le D.T.T., Nguyen T.Q., Dau V.Q., Hoang H.D. 2015. Undiagnosed cryptic diversity in small, microendemic frogs (*Leptolalax*) from the Central Highlands of Vietnam. PLoS One 10:1–21.

Roycroft E.J., Moussalli A., Rowe K.C. 2019. Phylogenomics uncovers confidence and conflict in the rapid radiation of Australo-Papuan rodents. Syst. Biol.

Rubinoff D., San Jose M., Hundsdoerfer A.K. 2021. Cryptic diversity in a vagile Hawaiian moth group suggests complex factors drive diversification. Mol. Phylogenet. Evol. 155:107002.

Sánchez-Restrepo A.F., Chifflet L., Confalonieri V.A., Tsutsui N.D., Pesquero M.A., Calcaterra L.A. 2020. A species delimitation approach to uncover cryptic species in the South American fire ant decapitating flies (Diptera: Phoridae: *Pseudacteon*). PLoS One 15:e0236086.

Sanguila M.B., Siler C.D., Diesmos A.C., Nuñeza O., Brown R.M. 2011. Phylogeography, geographic structure, genetic variation, and potential species boundaries in Philippine slender toads. Mol. Phylogenet. Evol. 61:333–350.

Schäffer S., Kerschbaumer M., Koblmüller S. 2019. Multiple new species: cryptic diversity in the widespread mite species *Cymbaeremaeus cymba* (Oribatida, Cymbaeremaeidae). Mol. Phylogenet. Evol. 135:185–192.

Schuchert P. 2014. High genetic diversity in the hydroid *Plumularia setacea*: a multitude of cryptic species or extensive population subdivision? Mol. Phylogenet. Evol. 76:1–9.

Shelley J.J., Swearer S.E., Adams M., Dempster T., Le Feuvre M.C., Hammer M.P., Unmack P.J. 2018. Cryptic biodiversity in the freshwater fishes of the Kimberley endemism hotspot, northwestern Australia. Mol. Phylogenet. Evol. 127:843–858.

Siler C.D., Alex Dececchi T., Merkord C.L., Davis D.R., Christiani T.J., Brown R.M. 2014. Cryptic diversity and population genetic structure in the rare, endemic, forest-obligate, slender geckos of the Philippines. Mol. Phylogenet. Evol. 70:204–209.

Siler C.D., Diesmos A.C., Alcala A.C., Brown R.M. 2011. Phylogeny of Philippine slender skinks (Scincidae: *Brachymeles*) reveals underestimated species diversity, complex biogeographical relationships, and cryptic patterns of lineage diversification. Mol. Phylogenet. Evol. 59:53–65.

Singhal S., Hoskin C.J., Couper P., Potter S., Moritz C. 2018. A framework for resolving cryptic species: a case study from the lizards of the Australian wet tropics. Syst. Biol. 67:1061–1075.

Smith M.L., Carstens B.C. 2019. Process-based species delimitation leads to identification of more biologically relevant species. Evolution (NY). 66:37–39.

Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15:1–15.

Smith S.A., Walker-Hale N., Walker J.F. 2020. Intragenic conflict in phylogenomic data sets. Mol. Biol. Evol. 37:3380–3388.

Struck T.H., Feder J.L., Bendiksby M., Birkeland S., Cerca J., Gusarov V.I., Kistenich S., Larsson K.H., Liow L.H., Nowak M.D., Stedje B., Bachmann L., Dimitrov D. 2018. Finding evolutionary processes hidden in cryptic species. Trends Ecol. Evol. 33:153–163.

Sukumaran J., Knowles L.L. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. USA 114: 1607–1612.

Thomas R.C., Willette D.A., Carpenter K.E., Santos M.D. 2014. Hidden diversity in sardines: Genetic and morphological evidence for cryptic species in the goldstripe sardinella, Sardinella gibbosa (Bleeker, 1849). PLoS One 9:e84719.

Trevisan C.C., Batalha-Filho H., Garda A.A., Menezes L., Dias I.R., Solé M., Canedo C., Juncá F.A., Napoli M.F. 2020. Cryptic diversity and ancient diversification in the northern Atlantic Forest *Pristimantis* (Amphibia, Anura, Craugastoridae). Mol. Phylogenet. Evol. 148:106811.

Trontelj P., Fišer C. 2009. Cryptic species diversity should not be trivialised. Syst. Biodivers. 7:1–3.

Vences M., Thomas M., Bonett R.M., Vieites D.R. 2005a. Deciphering amphibian diversity through DNA barcoding: chances and challenges. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360: 1859–1868.

Vences M., Thomas M., van der Meijden A., Chiari Y., Vieites D.R. 2005b. Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. Front. Zool. 2:5.

Vieites D.R., Wollenberg K.C., Andreone F., Kohler J., Glaw F., Vences M. 2009. Vast underestimation of Madagascar's biodiversity evidenced by an integrative amphibian inventory. Proc. Natl. Acad. Sci. USA 106:8267–8272.

Walker J.F., Brown J.W., Smith S.A. 2018. Analyzing contentious relationships and outlier genes in phylogenomics. Syst. Biol. 67:916–924.

Welton L.J., Siler C.D., Oaks J.R., Diesmos A.C., Brown R.M. 2013. Multilocus phylogeny and Bayesian estimates of species boundaries reveal hidden evolutionary relationships and cryptic diversity in Southeast Asian monitor lizards. Mol. Ecol. 22:3495–3510.

Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring phylogenetic networks using PhyloNet. Syst. Biol. 67:735–740.

Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. USA 107:9264–9269.

Zhang B., Chen T.W., Mateos E., Scheu S., Schaefer I. 2019. DNA-based approaches uncover cryptic diversity in the European *Lepidocyrtus lanuginosus* species group (Collembola: Entomobryidae). Invertebr. Syst. 33:661–670.

Zhang C., Sayyari E., Mirarab S. 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. Comp. Genomics. 53–75.

Zhou Z., Guo H., Han L., Chai J., Che X., Shi F. 2019. Singleton molecular species delimitation based on COI-5P barcode sequences revealed high cryptic/undescribed diversity for Chinese katydids (Orthoptera: Tettigoniidae). BMC Evol. Biol. 19:79.