The Challenges of Resolving a Rapid, Recent Radiation: Empirical and Simulated Phylogenomics of Philippine Shrews

THOMAS C. GIARLA1,2,* AND JACOB A. ESSELYN1,2

1Museum of Natural Science and 2Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA

Correspondence to be sent to: 119 Foster Hall, Louisiana State University, Baton Rouge, LA 70803, USA;
E-mail: giarla@lsu.edu

Received 24 February 2015; reviews returned 1 May 2015; accepted 7 May 2015

Associate Editor: John McCormack

Abstract—Phylogenetic relationships in recent, rapid radiations can be difficult to resolve due to incomplete lineage sorting and reliance on genetic markers that evolve slowly relative to the rate of speciation. By incorporating hundreds to thousands of unlinked loci, phylogenomic analyses have the potential to mitigate these challenges. Here, we attempt to resolve phylogenetic relationships among eight shrew species (genus Crocidura) from the Philippines, a phylogenetic problem that has proven intractable with small (<10 loci) data sets. We sequenced hundreds of ultraconserved elements and whole mitochondrial genomes in these species and estimated phylogenies using concatenation, summary coalescent, and hierarchical coalescent methods. The concatenated approach recovered a maximally supported and fully resolved tree. In contrast, the coalescent-based approaches produced similar topologies, but each had several poorly supported nodes. Using simulations, we demonstrate that the concatenated tree could be positively misleading. Our simulations also show that the tree shape we tend to infer, which involves a series of short internal branches, is difficult to resolve, even if substitution models are known and multiple individuals per species are sampled. As such, the low support we obtained for backbone relationships in our coalescent-based inferences reflects a real and appropriate lack of certainty. Our results illuminate the challenges of estimating a bifurcating tree in a rapid and recent radiation, providing a rare empirical example of a nearly simultaneous series of speciation events in a terrestrial animal lineage as it spreads across an oceanic archipelago. [Coalescence; concatenation; Crocidura; Philippines; SNPs; Soricidae; species tree; ultraconserved elements.]

Understanding the underlying processes that drive rapid radiations is an important objective in evolutionary biology (Schluter 1996; Baldwin and Sanderson 1998; Kozak et al. 2006; Moyles et al. 2009; Rundell and Price 2009), but it cannot be achieved without first characterizing the evolutionary history of such radiations. Unfortunately, disentangling phylogenetic relationships among lineages that diversified quickly and recently remains a significant challenge (Braian and Kimball 2001; Slosninski 2001; Maddison and Knowles 2006; Townsend 2007; Knowles and Chan 2008). If a set of species has diversified recently, many molecular markers will not have evolved quickly enough to provide signal for a fully bifurcating phylogeny. That problem is further compounded if diversification occurred rapidly. When the intervals between speciation events are short and effective population sizes are large, the most likely gene trees can conflict with the underlying species branching history (Degnan and Rosenberg 2006), hampering phylogenetic inference. In phylogenetic analyses of suspected recent and rapid radiations, it is often unclear whether poor support for a series of splits in a tree should be attributed to its explosive evolutionary history or to methodological artifacts (e.g., inappropriate selection of markers, models, or inference methods).

In theory, combining many independently segregating loci sampled from multiple loci, individuals per species and applying a coalescent-based approach should improve phylogenetic inference by reducing sampling error and allowing one to distinguish evolutionary signal from methodological artifact (McCormack et al. 2009; Kumar et al. 2012).

Recent advances in laboratory techniques (e.g., sequence capture; Gnrke et al. 2009) and falling prices for high-throughput sequencing have made the use of phylogenomic data sets more common (Lemmon et al. 2012; McCormack et al. 2013; Leaché et al. 2014; Pyron et al. 2014; Smith et al. 2014). Researchers now routinely sequence hundreds to thousands of loci in non-model organisms using target capture or reduced representation approaches. However, the sizes of these data sets sometimes force researchers to apply suboptimal inference methods like locus concatenation (Kubatko and Degnan 2007). Consequently, phylogeneticists are debating which methods are most appropriate for inferring phylogenies from phylogenomic data sets (Huang and Knowles 2009; Huang et al. 2010; Lanier et al. 2014; Gatesy and Springer 2014; Mirarab et al. forthcoming). At its core, this struggle revolves around the trade-offs among computational efficiency, data set size, and simplifying assumptions.

A typical study in which phylogenies are inferred from large, multilocus data sets include at least one of the four methodologies: (i) concatenation, in which loci are combined into one analyzable element under the assumption that individual loci are the product of the same underlying topology (de Queiroz and Gatesy 2007); (ii) summary coalescent methods, in which gene trees are estimated for each locus before a species tree is summarized from those gene trees (Maddison 1997; Kubatko et al. 2009; Liu et al. 2010; Chaudhary et al. 2013; Mirarab et al. 2014); (iii) hierarchical coalescent methods, in which the estimation of gene trees and species trees are integrated under the multispecies
coalescent model (Edwards et al. 2007; Heled and Drummond 2010; Bryant et al. 2012); and (iv) statistical binning, a hybrid approach in which subsets of loci are concatenated, and these “super loci” are then analyzed with summary or hierarchical coalescent methods (Song et al. 2012; Bayzid and Warnow 2013). Concatenation of loci has been criticized for failing to accommodate stochastic variation in gene histories and for producing inflated support metrics (Edwards et al. 2007; Kubatko and Degnan 2007; Knowles 2009). Summary coalescent approaches have also been critiqued—primarily by proponents of concatenation—for sometimes relying on poorly resolved gene trees (Gatesy and Springer 2014).

Summary methods are designed to accommodate gene tree variation, but they assume that input gene trees are estimated accurately. They do not explicitly model mutational variance, the stochasticity inherent to the estimation of gene trees themselves (Huang and Knowles 2009; Huang et al. 2010; Knowles et al. 2012). Hierarchical coalescent methods take mutational variance into account but can be computationally difficult when many loci are available, even for small radiations (Bayzid and Warnow 2013). These difficulties can be eased by only analyzing biallelic markers (e.g., single nucleotide polymorphisms (SNPs)), as implemented, for example, in the program SNAPP (Bryant et al. 2012). Finally, statistical binning offers a compromise between concatenation and summary approaches and potentially circumvents the computational burden of hierarchical coalescent approaches by reducing the number of analyzable “genes.” However, this technique has not been subjected to as much testing as the other approaches and has been labeled by critics (e.g., Springer and Gatesy 2014; Gatesy and Springer 2014) as combining the most problematic assumptions of concatenation (ignoring gene-tree discordance) and summary coalescent methods (unresolved gene trees).

Here, we sequence hundreds of ultraconserved elements (UCEs; Bejerano et al. 2004) and apply concatenated, summary coalescent, and hierarchical coalescent phylogenetic methods to resolve relationships among a small, recent radiation of shrews (genus Crocidura) from the Philippines. The Philippine archipelago is a biodiversity hotspot (Esselstyn et al. 2009). Mindoro Island is the sole known exception, where both C. grayi and C. mindorus occur (Fig. 1).

Previous studies of Philippine Crocidura have not yielded a consistent, well-supported phylogenetic hypothesis. Esselstyn et al. (2011) analyzed two mitochondrial genes across the Philippine endemic radiation and recovered a poorly supported tree with a shape that suggested a rapid series of speciation events. Concatenated analyses combining the same mitochondrial genes with three nuclear loci provided greater support (Esselstyn et al. 2009), but the resulting topology conflicted with the mitochondrial gene tree. Similarly, a species tree analysis of eight unlinked nuclear genes (Esselstyn et al. 2013) failed to confidently resolve any nodes within the Philippine endemic radiation (reproduced in Fig. 1). In this study, we again attempt to resolve relationships among these species by sequencing whole mitochondrial genomes (WMGs) and hundreds of nuclear loci in 18 individuals representing all recognized species from the Philippine endemic radiation. Our results indicate that adding DNA sequence data from hundreds of loci provides some clues regarding potential topologies but does not provide strong, consistent signal for a complete set of bifurcating relationships. Moreover, our analyses of simulated data show that some of our strongly supported empirical inferences should be regarded with skepticism.

**Materials and Methods**

**Ultraconserved Element Sequencing and Data Quality Control**

We sequenced UCEs in 19 individuals representing 7 species of Philippine shrews and an Indonesian out-group species, Crocidura orientalis (Table 1). One individual represents the first record of any shrew species from Lubang Island, Philippines. Based on the genetic evidence detailed below, we identify this specimen as C. grayi. Specimens included in this study are held in the collections of the Field Museum of Natural History (FMNH) or the University of Kansas Biodiversity Institute (KU). Most DNA samples were extracted as part of previous studies (Esselstyn et al. 2009, 2013); DNA extractions from additional individuals followed the same protocols. Aliquots of DNA extracts containing between 0.5 μg and 3.0 μg of DNA were purified using homemade solid-phase reversible immobilization beads (Rohland and Reich 2012), and each sample was rehydrated in 50 μL of TE buffer. Samples were sonicated on an EpiSonic Bioprocessor 1100 (Epigenetik) to an average fragment size of 600 bp. We used a Kapa “on bead” low throughput Library Preparation Kit (Kapa Biosystems) to prepare 19 libraries for sequencing on Illumina platforms.
FIGURE 1. a) A map of the Philippines and surrounding islands. Dark lines outlining light gray encompass composite islands that existed when sea levels were lower (~120 m) during Pleistocene glacial maxima. b) Partial species tree from Esselstyn et al. (2013) based on eight nuclear loci showing the Philippine endemic radiation of Crocidura and five Indonesian outgroup species. Numbers at nodes denote Bayesian posterior probabilities. “+” signs after island names indicate that the species is also found on nearby smaller islands. c) A specimen of Crocidura negrina captured and photographed on the island of Negros in 2006. Photograph by Jacob Esselstyn.

TABLE 1. Voucher numbers, collection localities, and results from a single 150 bp paired-end sequencing run on an Illumina MiSeq machine.

<table>
<thead>
<tr>
<th>Species</th>
<th>Catalog no.</th>
<th>Country</th>
<th>Island</th>
<th>Locality</th>
<th>Total reads sequenced</th>
<th>Reads in WMGs(^a) (% of total reads)</th>
<th>No. of UCE loci(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crocidura orientalis</td>
<td>FMNH212778</td>
<td>Indonesia</td>
<td>Java</td>
<td>Mt. Gede</td>
<td>733,728</td>
<td>17,237 (2.3)</td>
<td>1305</td>
</tr>
<tr>
<td>Crocidura beatus</td>
<td>FMNH166439</td>
<td>Philippines</td>
<td>Mindanao</td>
<td>Bukidnon</td>
<td>824,228</td>
<td>16,293 (2.2)</td>
<td>1156</td>
</tr>
<tr>
<td>Crocidura beatus</td>
<td>KU167037</td>
<td>Philippines</td>
<td>Mindanao</td>
<td>Mt. Balatukan</td>
<td>1,353,498</td>
<td>19,576 (1.4)</td>
<td>1462</td>
</tr>
<tr>
<td>Crocidura beatus</td>
<td>KU167039</td>
<td>Philippines</td>
<td>Mindanao</td>
<td>Mt. Balatukan</td>
<td>1,355,190</td>
<td>3545 (0.4)</td>
<td>1438</td>
</tr>
<tr>
<td>Crocidura beatus</td>
<td>KU169569</td>
<td>Philippines</td>
<td>Mindanao</td>
<td>Zamboanga</td>
<td>1,007,854</td>
<td>21,652 (2.1)</td>
<td>1321</td>
</tr>
<tr>
<td>Crocidura greyi</td>
<td>FMNH218425</td>
<td>Philippines</td>
<td>Mindanao</td>
<td>Mt. Ambulon</td>
<td>1,374,355</td>
<td>129,445 (8.2)</td>
<td>1461</td>
</tr>
<tr>
<td>Crocidura greyi</td>
<td>KU165176</td>
<td>Philippines</td>
<td>Mindoro</td>
<td>Mt. Calavite</td>
<td>683,790</td>
<td>18,567 (2.7)</td>
<td>1366</td>
</tr>
<tr>
<td>Crocidura greyi</td>
<td>KU165178</td>
<td>Philippines</td>
<td>Mindoro</td>
<td>Mt. Calavite</td>
<td>840,345</td>
<td>5432 (0.8)</td>
<td>1371</td>
</tr>
<tr>
<td>Crocidura greyi</td>
<td>KU165912</td>
<td>Philippines</td>
<td>Luzon</td>
<td>Mt. Labo</td>
<td>1,220,691</td>
<td>24,883 (2.0)</td>
<td>1414</td>
</tr>
<tr>
<td>Crocidura ninoyi</td>
<td>FMNH145685</td>
<td>Philippines</td>
<td>Sibuyan</td>
<td>Mt. Guitinguitin</td>
<td>597,678</td>
<td>8300 (1.4)</td>
<td>1246</td>
</tr>
<tr>
<td>Crocidura ninoyi</td>
<td>KU164877</td>
<td>Philippines</td>
<td>Sibuyan</td>
<td>Mt. Guitinguitin</td>
<td>697,678</td>
<td>8300 (1.4)</td>
<td>1246</td>
</tr>
<tr>
<td>Crocidura palawanensis</td>
<td>FMNH195992</td>
<td>Philippines</td>
<td>Palawan</td>
<td>Mt. Mantalingahan</td>
<td>1,661,403</td>
<td>5992 (0.9)</td>
<td>1310</td>
</tr>
<tr>
<td>Crocidura panayensis</td>
<td>KU164878</td>
<td>Philippines</td>
<td>Panay</td>
<td>Antique</td>
<td>1,360,066</td>
<td>12,514 (0.9)</td>
<td>1279</td>
</tr>
<tr>
<td>Crocidura panayensis</td>
<td>KU164877</td>
<td>Philippines</td>
<td>Panay</td>
<td>Antique</td>
<td>1,519,080</td>
<td>4099 (0.3)</td>
<td>1331</td>
</tr>
</tbody>
</table>

Note: Catalog numbers reference specimens held at the FMNH and KU.
\(^a\) Whole mitochondrial genomes.
\(^b\) Total number of UCE loci that were included in the final data set after passing the minimum coverage threshold.

Individual samples were barcoded using 19 TruSeq-style adapters with 10 bp indices (from Faircloth and Glenn 2012). We followed the Kapa protocol, but during the end repair, A-tailing, and adapter-ligation steps we used one-fourth of Kapa’s recommended reagent volumes. We otherwise adhered to their protocol. We used a MYbaits kit (MYcroarray) containing 2560 probes to enrich our samples for over 2000 UCEs (Faircloth
et al. 2012). Libraries were combined in equimolar proportions into three pools prior to UCE enrichment, following the MYaBaits protocol. After enrichment, we amplified the products using 15 cycles of PCR and then combined the 3 enrichment pools in equimolar ratios. We sequenced the enrichment products at the Georgia Genomics Facility (Athens, GA) using one lane of Illumina MiSeq 150 bp paired-end sequencing. We associated particular sequences with individual specimens using strict matching of their unique bar codes in CASAVA (Illumina, Inc.). Demultiplexed sequence reads were subjected to quality control in Trimmomatic (Bolger et al. 2014), which removes contaminating adapter sequences and low quality bases, using the parallel wrapper script Illumiprocessor (Faircloth 2013). De novo read assembly was performed in ABySS v1.3.7 (Simpson et al. 2009) with a k-mer value of 35. This step and all subsequent UCE processing steps were completed using Phyvuce v1.4 (Faircloth 2014).

Whole Mitochondrial Genome Assembly and Mitochondrial Gene Tree Estimation

Residual mitochondrial DNA persisted after enrichment and was sequenced along with the UCE loci. We assembled WMGs for each sample using MiToBIM v1.6 (Hahn et al. 2013), with the Crocidura shantungensis WMG sequence (Genbank Accession No. NC021398) used as a reference. The D-loop region of the mitochondrial genome was not included in the final assembly due to complications associated with assembling this highly variable region. We aligned WMGs in Geneious v7.1.7 (Biomatters) using the MUSCLE algorithm (Edgar 2004). We assessed alternative WMG partitioning schemes using PartitionFinder v1.11 (Lanfear et al. 2012), considering combinations of 64 potential data subsets: 22 subsets representing each tRNA, 39 representing each codon position in each of the 13 protein-coding genes, 2 representing each tRNA gene, and 1 subset that grouped the short spacer sequences between functional elements. Using the optimal partitioning strategy from PartitionFinder, we inferred a mitochondrial genealogy using WMG sequences in MrBayes v3.2.2 (Ronquist et al. 2012). We unlinked substitution models and rate parameters across subsets and initialized two MrBayes runs with seven incrementally heated chains and one cold chain. Each analysis ran for 5 million generations, with samples drawn every 1000 generations. We checked effective sample sizes (ESSs) for all parameters in Tracer v1.6 (Rambaut et al. 2013) and ensured that topological convergence was achieved by checking that the standard deviation of split frequencies across runs was less than 0.01. We discarded the first 25% of trees as burn-in and summarized the remaining posterior distribution of trees as a majority-rule consensus tree.

Inferring Species Trees from UCE Gene Trees Using Summary Coalescent Methods

We obtained a data set of 1112 UCEs in which all Crocidura species were represented by at least one individual per locus. One hundred and ninety-three of these loci contained no phylogenetically informative variation, and we excluded them from species tree analyses, leaving 919 loci. We then used CloudForest beta v0.1 (Crawford and Faircloth 2014) to fit nucleotide substitution models, infer genealogies in PhyML v3 (Guindon and Gascuel 2003), and generate 100 bootstrap pseudoreplicates for each phylogeny. Genealogies were used to infer species trees using the summary coalescent methods implemented in MulRF v1.2 (Chaudhary et al. 2013) and ASTRAL v4.7.6 (Miraab et al. 2014). Bootstrap pseudoreplicates were also analyzed in MulRF and ASTRAL. Unlike most other summary coalescent methods, both programs allow unrooted, multifurcating gene trees as input, which, for our low-variation loci, is a more accurate reflection of topological uncertainty in the underlying genealogy than the fully bifurcating trees produced by PhyML (where polytomies are randomly resolved with very short branches). We therefore applied the `di2multi` function in APE (Paradis et al. 2004) to collapse these arbitrarily resolved relationships (branch lengths < 10^{-6}) into polytomies in all gene trees and bootstrap pseudoreplicates. To apply nodal support metrics to the species trees, we summarized bootstrap support on the estimated topologies from MulRF and ASTRAL in DendroPy v3.12 (Sukumaran and Holder 2010).

Inferring Species Trees under the Multilocus Coalescent Model

We used *BEAST v2.1.3 (Bouckaert et al. 2014) to jointly estimate gene trees and species trees under the multispecies coalescent model (Heled and Drummond 2010). This approach is computationally intensive, and analyses using all 919 phylogenetically informative alignments were intractable. To ease the computational burden, and after testing runs analyzing data sets containing 25–100 loci, we divided the data set into 19 random subsets (eighteen 50-locus subsets and one 19-locus subset). Individuals were assigned to one of nine possible species based on morphological diagnoses, collection localities, and the results of the WMG analysis, with one exception: _C. ninoci_ was split into two putative species (see section “Results”). All loci were assigned the best-fitting nucleotide substitution model identified by CloudForest, except those that were assigned a GTR or SYM model. GTR/SYM rate parameter estimates failed to converge in initial runs, so we replaced GTR models with HKY models and SYM models with K2P models. For models that allow unequal base frequencies, empirical values were applied to reduce the number of estimated parameters. We assigned each locus a strict clock model with an exponential prior distribution
(mean = 1.0) and applied a Yule species tree prior with a piecewise-linear and constant-root population-size model. For each of the subsets, we conducted a single BEAST run until all parameters exhibited adequate mixing, up to $2 \times 10^7$ generations, and sampled the chain every $2 \times 10^6$ generations. We used Tracer to assess ESSs for all parameters. Species trees from all runs were individually summarized in TreeAnnotator v2.1.3 (Bouckaert et al. 2014) after discarding 10% of the trees as burn-in. We computed a 50% majority-rule consensus tree from the 19 species trees using the sumtrees tool in DendroPy.

**Analysis of Concatenated UCE Loci**

We used Phyluce to extract the best-fitting nucleotide substitution model from the CloudForest output for each of the 1112 UCE alignments. We then concatenated all of the UCE loci, grouping all loci sharing the same best-fitting model into model-specific, partitioned subsets. We inferred a phylogeny from the concatenated data set in MrBayes using the same run settings as for the WMG analysis described above.

**Analysis of SNPs within UCE Loci**

We compiled a data set of biallelic, unlinked SNPs by mapping Illumina reads to UCE reference sequences and extracting a single SNP from each locus. A complete list of commands and settings used in our analysis pipeline is provided in Online Appendix 1 (available on Dryad at http://dx.doi.org/10.5061/dryad.b756), and a summary is provided here. First, we selected as our reference the sample for which the largest number of UCE loci was successfully sequenced as our reference the sample for which the largest number of UCE loci was successfully sequenced (C. beatus, KU167037; Table 1). We used BWA v0.7.7 (Li and Durbin 2009), Picard v1.106 (available at http://broadinstitute.github.io/picard/; last accessed May 26, 2015), and SAMtools v0.119 (Li et al. 2009) to create reference indices and sequence dictionaries from the UCE fasta files generated by Phyluce for KU167037. Then, we used snps.py from the Phyluce package to align reads from each sample to the reference, resulting in BAM files for each of the 19 individuals. We indexed the individual BAM files using SAMtools, and then, for each of those BAM files, we used the Genome Analysis Tool Kit v3.3 (McKenna et al. 2010) to locate regions suspected of containing indels, realign these regions, call SNPs and indels (including heterozygotes), and merge calls across samples. We extracted the SNPs from the resulting VCF file and annotated those that were potential false positives or of low quality. Finally, we used VCFTools v0.1.12 ( Danecek et al. 2011) to remove all of the SNPs that did not pass our quality thresholds (Online Appendix 1 available on Dryad at http://dx.doi.org/10.5061/dryad.b756) and to select up to one SNP per UCE locus. Only SNPs genotyped for all 19 individuals were considered for inclusion in the final data set.

We analyzed the SNP data set using SNAPP v1.1.6 (Bryant et al. 2012), a software package available in BEAST. Using a Phyluce script (convert_vcf_to_snapp.py), we converted our filtered VCF files into a format that is interpretable by SNAPP. We conducted two analyses in SNAPP: one in which we assigned individuals to species just as we did for the BEAST analyses, and another in which each of the 19 individuals was treated as a distinct species. To estimate the backward mutation rate $u$, we sampled the MCMC chain for that parameter instead of relying on the default value. We did not alter the other default parameters and ran the MCMC chain for 5 million generations, sampling every 1000 generations. We assessed ESS values in Tracer when the run was complete. After removing 10% of the sample as burn-in, we constructed a species tree and cloudogram using TreeAnnotator and DensiTree (Bouckaert 2010), respectively.

**Evaluating Empirical Results with Simulations**

Several of our initial phylogenetic results conflicted with one another, and some relationships consistently received poor support (see section “Results”). To assess the extent to which inconsistent and poor resolution was due to the Philippine endemic radiation of Crocidura being a truly rapid, difficult-to-resolve problem, we generated a series of simulated data sets matching and modifying the characteristics of our empirical data. We then inferred phylogenies from these sequences and compared how well different approaches could identify the known branching history: ASTRAL, MulRF, and the concatenated MrBayes analysis all produced the same empirical topology (see section “Results”), which we used as the basis for the “true” species tree from which we simulated character data. We used BP&P v3 (Yang and Rannala 2010, 2014) to estimate node depths ($r$) and mutation-rate-scaled population sizes ($\theta$) on our chosen topology. A single BP&P run with 919 loci failed to converge, so we divided the data set into eight random subsets (seven subsets of 115 loci and one subset of 114 loci) and analyzed each subset independently with a single run. For all runs, we applied diffuse prior distributions [$\Gamma(2, 0.001)$; mean = 0.001] for both the $\theta$ and $r$ parameters. We drew 150,000 samples from the MCMC chain of each analysis, sampling every 10 generations and discarding the first 50,000 samples as burn-in. We used the mean values of $\theta$ and $r$ taken across all eight runs of BP&P in the “true” tree that served as the basis for our simulated data. Because we cannot obtain empirical $\theta$ estimates for taxa with sequences from a single individual, we applied the mean $\theta$ across all extant and ancestral branches to the four branches represented by single individuals.

We constructed three different data sets using BP&P’s MCCoal sequence simulator. The first, Sim-Matching, matches the characteristics of our empirical results. We simulated data under the same topology as the BP&P guide tree, with empirical estimates for $\theta$ and $r$.
defining the species tree’s population sizes and branch lengths, respectively. We generated data with the same pattern of individuals sampled per species as in our empirical data set (19 individuals total, with between 1 and 5 individuals sampled per species). Our goal here was simply to see if we could accurately estimate the tree that generated the data, thereby suggesting how confident we should be in our empirical estimates. The second data set, Sim-Multi-Individual, increased the number of simulated sequences to five individuals per species. In the other simulations, four of the species (Crocidura sp., C. mindorus, C. ninogi, and the outgroup C. orientalis) included only one individual each, whereas the remaining species had two (C. nigrina, C. palawanensis, and C. panayensis), four (C. boasii), or five (C. grayi) individuals each. Sim-Multi-Individual increases the total number of simulated individuals to 45, testing the extent to which increased sampling within species improves our ability to infer an accurate species tree. Our third simulated data set, Sim-3x-Rate, is identical to Sim-Matching, except branch lengths in the underlying MCcoal guide tree were tripled (population sizes stayed constant) to test the extent to which using more variable loci would improve our ability to infer the tree.

For each simulation, we generated 500 alignments under the Jukes—Cantor nucleotide substitution model (JC69), each 700 bp long. We conducted a separate concatenated analysis for each of the three simulated data sets in MrBayes, using the same settings as with the concatenated empirical analyses, but assigning the JC69 model to the entire concatenated alignment. We then estimated species trees for the three simulated data sets in MulRF and ASTRAL, following the same procedure as for our empirical data set. Finally, to use *BEAST, we divided the 500 loci in each data set into ten 50-locus subsets, analyzed each in *BEAST, and summarized the ten resulting species trees in a 50% majority-rule consensus tree. We applied the JC69 substitution model to every locus, but otherwise used the same settings as in our empirical analyses. The Sim-Matching and Sim-3x-Rate MCMC chains exhibited successful mixing, but none of the Sim-Multi-Individual data sets converged in *BEAST. To ease the computational burden but still maintain more than 1 sequence per species, we pruned Sim-Multi-Individual for the *BEAST input files to include 23 sequences per locus; 2 sequences for each species that in the empirical data had only one individual, and the empirical number of samples for the remaining taxa.

### Results

#### Sequence Characteristics

The MiSeq run produced 19.2 million reads, with 0.6–1.7 million reads per individual (Table 1). A small proportion of the reads from each individual (2.5% on average; Table 1) is similar to the reference.

#### Table 2. PartitionFinder results for WMG

<table>
<thead>
<tr>
<th>PartitionFinder subset</th>
<th>Best-fitting model</th>
<th>WMG regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset 1</td>
<td>GTR+I+G</td>
<td>ATP6 (3); ATP8 (1,3); COX1 (1); COX2 (1); COX3 (3); CYTB (1); ND1 (1); ND2 (1); ND3 (1); ND4 (3); ND5 (1); ND6 (3,1); all tRNAs except tRNA-Tyr; 12S-rRNA, 16S-rRNA, Non-coding spacer regions</td>
</tr>
<tr>
<td>Subset 2</td>
<td>HKY+I</td>
<td>ATP6 (1); ATP8 (2); COX1 (2); COX2 (2); COX3 (1); CYTB (2); ND1 (2); ND2 (2); ND3 (1); ND4 (1); ND5 (2); ND6 (2); RNAs-Tyr; 12S-rRNA, 16S-rRNA, Non-coding spacer regions</td>
</tr>
<tr>
<td>Subset 3</td>
<td>GTR+G</td>
<td>ATP6 (2); COX1 (3); COX2 (2); COX3 (2); CYTB (3); ND1 (3); ND2 (3); ND3 (3); ND4 (1); ND5 (2); ND6 (2)</td>
</tr>
</tbody>
</table>

Note: Numbers in parentheses after gene names denote the codon position included in the subset.

C. shantungensis mitochondrial genome. These reads were assembled into a nearly complete WMG for each specimen (the D-loop region was excluded due to poor coverage in assemblies). The WMG alignment contains 19 individuals and 1730 phylogenetically informative sites (GenBank Accession Nos KR537873–KR537891). PartitionFinder identified an optimal partitioning strategy that includes three data subsets, each with a different nucleotide substitution model (Table 2). We assembled the remaining reads into contigs and extracted those contigs that matched our UCE probe set. We compiled a UCE data set that allowed for missing data as long as at least one individual from each species was sequenced per locus. It comprises 1112 loci with an average locus length of 663 bp (Online Appendix 2 available on Dryad at http://dx.doi.org/10.5061/dryad.b7156; GenBank Accession Nos KR537892–KR558635). Five alignments contain no variation, and 193 contain no phylogenetically informative characters. For all but the concatenated analyses, only the 919 phylogenetically informative alignments were included. Phylogenetically informative loci include, on average, 13.4 variable characters, 3.1 phylogenetically informative characters, and 1.5 sites coded as indels. In total, 9.5% of our UCE matrices comprise missing data. For most loci, the best-fitting nucleotide substitution model is HKY; the second most common model is F81 (Online Appendix 2 available on Dryad at http://dx.doi.org/10.5061/dryad.b7156).

As in other studies of UCEs (e.g., Smith et al., 2014), variation is distributed randomly along the length of individual alignments, with most sequence differences occurring in the distal portions of aligned loci (Supplementary Fig. S1 available on Dryad at http://dx.doi.org/10.5061/dryad.b7156). We also compiled a data set of 1170 biallelic SNP loci (up to one from each UCE) for all 19 individuals. Among all the SNPs, 3.6% of the calls are heterozygous.
Phylogenetic Results

The mitochondrial phylogeny (Fig. 2), based on WMGs for 19 individuals, contains weak conflicts with the two-gene mitochondrial tree from Esselstyn et al. (2011). All individuals identified morphologically as the same species form monophyletic groups with high support. However, the only interspecific relationship that receives high support is the clade uniting *C. panayensis*, *C. ninoyi*, and *C. nigrina*. In contrast, the phylogeny based on concatenated UCE sequences (Fig. 3, Supplementary Fig. S2 available on Dryad at http://dx.doi.org/10.5061/dryad.b7156) is strongly supported at all nodes. Although the same sister relationship between *C. nigrina* and *C. panayensis* is recovered, it differs from the WMG gene tree at every other node. Individuals from each species form monophyletic groups, except for *C. ninoyi*, a species known only from the small island of Sibuyan (Esselstyn and Goodman 2010). One *C. ninoyi* individual (FMNH145685, a topotype) is part of a clade with *C. nigrina* and *C. panayensis* and the sister relationship to *C. nigrina* and *C. panayensis* and the sister relationship of *C. palawanensis*, which was not strongly supported in the WMG gene tree (Fig. 2), receives 100% bootstrap support as sister to all other members of the Philippine endemic radiation. Among the three nodes that do not receive 100% support, bootstrap values average 29% higher in the ASTRAL tree (Fig. 5) than in the MulRF tree (Fig. 4).

We conducted 19 *BEAST* runs, each incorporating a different random subset of loci from the pool of 919 variable loci. Individual species trees based on the 19 runs comprise 16 different topologies (Supplementary Fig. S3 available on Dryad at http://dx.doi.org/10.5061/dryad.b7156), each failing to support relationships along the backbone of the phylogeny. Strong conflict is rare, with only one instance of conflicting relationships receiving a posterior probability (PP) $\geq 0.95$ in different species trees (the placement of *Crocidura* sp. relative to *C. grayi* and *C. mindorus* in subset 2 vs. subsets 11 and 18; Supplementary Fig. 3 available on Dryad at http://dx.doi.org/10.5061/dryad.b7156). A majority-rule consensus tree summarizing the 19 individual
species trees (Fig. 6) does not strongly support the position of Crocidura sp. or C. beatus, nor does it strongly conflict with the topology recovered by the concatenated, MulRF, or ASTRAL analyses. The sister relationship between C. minorus and C. grayi is recovered in 13 of the 19 *BEAST trees, but support for this varies with PPs between 0.37 and 0.99. The core clade that contains all of the Philippine species except for C. palawanensis is more strongly supported (appearing in 18 of the 19 *BEAST trees with an average PP of 0.94), as is the clade that unites C. panayensis and C. negrina (appearing in all 19 *BEAST trees with an average PP of 0.94). The position of C. ninoyi is somewhat less certain, but its position as sister to C. panayensis + C. negrina appears in 16 of the 19 *BEAST trees, with an average PP of 0.89.

SNAPP analyses of 1170 biallelic SNPs (Fig. 7) did not produce better-supported species trees than other approaches. When each individual was allowed to represent its own putative species (Fig. 7a), all of the species (as defined above) with multiple individuals sampled are monophyletic. However, the clouddogram illustrates substantial conflict among species trees in the posterior distribution, and most relationships in the maximum clade credibility tree have low PPs (Fig. 7a). The MCMC chain exhibited successful mixing for all but three of the 37 $\theta$ parameters (results not shown), and the low ESS values (<100) for those three parameters may have suppressed nodal support. SNAPP estimates a different topology when individuals are assigned to species a priori (Fig. 7b; all ESS values >700), but strongly supported nodes do not conflict with corresponding interspecific nodes in Figure 7a. The only interspecific relationships to receive strong support are: (i) the sister relationship between C. negrina and C. panayensis and (ii) the node uniting all Philippine species except C. palawanensis. Unlike in other analyses, C. palawanensis is inferred as sister to the Indonesian out-group, C. orientalis, but this relationship is poorly supported.
In BP&P, we estimated branch lengths and population sizes across a fixed guide tree (Online Appendix 3 available on Dryad at http://dx.doi.org/10.5061/dryad.b7156), and we generated three simulated data sets based on the resultant species tree. We used those simulated data sets to test: (i) the extent to which different methods can successfully infer the true species tree, given our empirical circumstances and (ii) how we might have modified data collection to improve prospects for inferring the correct topology. Across all four analytical approaches (concatenated MrBayes, MulRF, ASTRAL, and *BEAST), the Sim-Matching and Sim-Multi-Individual data sets fail to recover the true topology (Figs. 3–6). However, only the concatenated analyses of these data sets include a node that is both incorrect and strongly supported (the sister relationship between Crocidura sp. and C. beatus; PP = 1.0). In contrast, all analyses involving the Sim-3x-Rate data set recover the correct topology (Figs. 3–6). Individual phylogenies from simulated data subsets used in *BEAST are shown in Supplementary Figs. S4–S6 (available on Dryad at http://dx.doi.org/10.5061/dryad.b7156).

**DISCUSSION**
Phylogenetic relationships within the Philippine endemic radiation of *Crocidura* have been difficult to resolve. Previous efforts using mitochondrial gene trees, mito-nuclear concatenation, and nuclear species-tree approaches with fewer than 10 loci have consistently
yielded poorly supported phylogenies (Esselstyn and Brown 2009; Esselstyn et al. 2013). Here, a phylogeny based on WMGs does little to improve resolution (Fig. 2). Adding substantially more nuclear sequence data, however, modestly improves recovery of interspecific relationships. Our empirical UCE trees, although often not well supported, are consistent across three of the four analytical methods we used (concatenation and the two summary coalescent approaches; Figs. 3–5). However, our analyses of simulated data sets, which mimic our empirical circumstances, led to the repeated inference of false topologies, casting doubt on the accuracy of all four inference methods (Sim-Matching and Sim-Multi-Individual; Figs. 3–6). Despite these issues, strongly supported conflict is rare—in our analyses of simulated data, only the concatenated results contain an incorrect and highly supported topology (Fig. 3). In all other cases, incorrect relationships received appropriately low support (Figs. 4–6). Recently, Gatesy and Springer (2014) criticized the use of summary coalescent methods, suggesting instead that an approach in which all sequences are concatenated is superior. For Philippine Crocidura, this is clearly not the case. Summary species tree approaches like MulRF (Fig. 4) and ASTRAL (Fig. 5) produced incorrect topologies when we analyzed simulated data, but the erroneous nodes are poorly supported. Our concatenated empirical tree may be fully resolved and well supported at every node, but the posterior probabilities are likely inflated and the topology may be unreliable due to the well-documented statistical inconsistency associated with concatenation (Kubatko and Degnan 2007; Cranston et al. 2009). Although we did not evaluate the accuracy of SNAPP using simulations, the low support at nearly all nodes in the SNAPP species trees suggests that we are not being positively misled (Fig. 7).

None of our coalescent-based analyses produces a phylogeny that is well supported at all nodes, but some interspecific relationships do emerge fairly consistently with moderate to strong support: (1) C. palawanensis as sister to all other Philippine species, (2) C. negrina as sister to C. panayensis, and (3) C. ninoyi as sister to C. negrina + C. panayensis. Slightly less consistently supported are the sister relationship between C. mindorus and C. grayi and the position of that clade relative to others. The relative positions of Crocidura sp. and C. beatus are almost never well supported, except for the species tree derived from ASTRAL, which strongly supports their inclusion in a clade with C. negrina, C. panayensis, and C. ninoyi (bootstrap support of 98%; Fig. 5). ASTRAL may be the most effective species tree program we used, having the fewest incorrect nodes and the highest support values for the correct nodes when the Sim-Matching and Sim-Multi-Individual data sets are analyzed.

The low support we recovered at certain nodes could be an artifact of methodology. First, model
mis-specification might diminish support values (Buckley 2002). If that were the case, we would expect to see consistently higher support values on trees derived from the Sim-Matching data set (where models were specified exactly as they were simulated) than those derived from the empirical data. Slightly higher support values are in fact observed in the Sim-Matching *BEAST results (Fig. 6), but the opposite is true for our results from MuRF (Fig. 4) and ASTRAL (Fig. 5). Second, low support may have been caused by sampling too few individuals per species (Maddison and Knowles 2006; McCormack et al. 2009). Our empirical data set included only one individual per species for Crocidura sp., C. mindorae, C. ninogi, and C. orientalis, so we conducted a simulation in which multiple individuals were sampled within every species (Sim-Multi-Individual). However, in none of those analyses was the true topology identified (Figs. 3–6), suggesting that at this scale, individual sampling was not the primary factor leading to low support or inaccurate reconstruction of the topology. Finally, low support could be driven by the low variation present in individual alignments. Results from the Sim-3x-Rate data set support this hypothesis, with the recovered species trees always matching the true topology and having higher support values than seen in analyses of the other two simulated data sets (Figs. 4–6). Nonetheless, increasing the mutation rate 3-fold still results in poor support for placements of Crocidura sp. and C. panayensis in *BEAST (Fig. 6) and Crocidura sp. by ASTRAL (Fig. 5).

While hierarchical coalescent models like *BEAST are best able to account for the various kinds of error associated with estimating species trees, they tend to be computationally intractable for large data sets, and we were unable to analyze all UCE loci (or even a substantial portion of them) in a single *BEAST run. Improvements in computational capacity may eventually solve this problem, but with current technology, we observed little evidence of convergence in analyses of more than 50 loci in *BEAST, even after billions of MCMC generations. The piecemal approach we introduced represents a compromise between maximizing data usage and ensuring computational tractability but fails to yield a consistent phylogenetic hypothesis. The 19 species trees we recovered from analysis of subsets of our empirical data vary considerably; 16 different topologies are recovered, all with relatively long terminal branches and short internodes (Supplementary Fig. S3 available on Dryad at http://dx.doi.org/10.5061/dryad.b7156). Trees with this shape are particularly difficult to resolve and indicate a rapid burst of speciation (Degnan and Rosenberg 2006; Kubatko and Degnan 2007). In situations like this, ancestral polymorphisms can be carried through a series of cladogenesis events in short succession, leading to rampant incomplete lineage sorting. The mutation rate in UCE loci may simply be too slow to record the near-simultaneous branching history in parts of this clade. Other DNA sequence data types, particularly those with faster mutation rates, may prove useful, but obtaining homologous sequences of neutrally evolving loci in multiple species of non-model organisms remains challenging. Restriction site-associated DNA sequencing (RADseq; Baird et al. 2008) is now a commonly used tool for making population genetic inferences, and several studies have explored the utility of RADseq for phylogeny estimation (Rubin et al. 2012; Cariou et al. 2013; Eaton and Ree 2013; Cruaud et al. 2014). Accurately identifying orthologous loci is not trivial, however, and commonly applied locus filtering techniques can bias results (Huang and Knowles forthcoming). Perhaps the largest benefit of using sequence capture-based markers like UCEs over RADseq is the higher confidence in orthology and the ease of incorporating those loci into other studies (Harvey et al. 2013). Nonetheless, RADseq loci could help more fully resolve this radiation if those loci are more variable than the UCEs included here.

Crocidura ninogi is the only putative species defined at the outset of this project whose constituent individuals did not form a monophyletic group in the concatenated tree (Supplementary Fig. S2 available on Dryad at http://dx.doi.org/10.5061/dryad.b7156) or the fully expanded SNAPP tree (Fig. 7a). The phylogenetic position of Crocidura sp. was not consistent across our analyses, but in no case was it inferred as a sister to C. ninogi. Given these results, the nuclear data strongly indicate that these two individuals represent different species. In contrast, the mitochondrial data point to a close relationship with strong support (Fig. 2). Mitochondrial discordance can be caused by mitochondrial introgression, whereby sympatric populations exchange mitochondrial DNA (Ballard and Whitlock 2004; Toews and BreIsford 2012). This is possible when one mitochondrial haplotype confers a specific advantage over another (adaptive introgression). The observation of mito-nuclear discordance among Sibuyan shrews has been obscured in past analyses (e.g., Esselstyn et al. 2009) by the concatenation of mitochondrial and nuclear genes, where the more variable mitochondrial data overwhelmed signal in the small number (three) of nuclear genes. We therefore posit that Crocidura sp. from Sibuyan is an undescribed species (possibly occurring on other nearby small islands), and the individuals we sampled from Sibuyan experienced mitochondrial, but not nuclear, introgression. However, this finding is based on only two individuals, and our conclusions are therefore tentative. Additional fieldwork is needed to acquire new specimens and allow a comprehensive taxonomic and population genetic analysis.

Although we were unable to recover a consistent, fully bifurcating topology, our results support a broad biogeographic pattern, with the oldest cladogenesis event occurring between the Palawan endemic C. palawanensis and the species inhabiting the rest of the islands, and the second split occurring between the northern species (C. grayi and C. mindorae) on Luzon and Mindoro and the southern species (C. beatus, C. negrina, C. ninogi, C. panayensis, and C. sp.) on greater Mindanao.
Negros, Panay, and Sibuyan. This pattern is consistent with the idea that Palawan served as a biogeographic bridge for terrestrial animals from the Sunda Shelf islands to the oceanic Philippines (Diamond and Gilpin 1983; Esseletyn et al. 2010). Once this lineage arrived on Palawan, interisland dispersal and allopatric divergence appear to have occurred rapidly. Rapid radiation of shrews across an oceanic archipelago is remarkable, because it requires multiple, nearly simultaneous overwater colonization events. Philippine shrews are small (ca. 5–15 g) and may have fast metabolisms like other shrews (Churchfield 1990), which would seemingly make it difficult for them to disperse over water. Nevertheless, the presence of Crocidura on numerous oceanic islands in the Philippines and Indonesia (Kitchener 1994; Ruedi et al. 1996; Esselstyn and Olivares 2010; Esselstyn et al. 2013) implies that they are in fact capable colonists. If colonization did occur very rapidly, it raises the possibility of occasional gene flow between these mostly allopatric species, such as what we suspect happened between the two putative species on Sibuyan. Interspecific gene flow is incompatible with bifurcating phylogenetic relationships and would certainly obfuscate any attempt to estimate the phylogeny. Although we doubt this is a pervasive factor due to the mostly allopatric distributions of Philippine shrews, exploring the possibility among partially co-occurring species may prove fruitful.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.b7156.

FUNDING

This work was supported by a grant from the National Science Foundation [grant number DEB-1343517].

ACKNOWLEDGMENTS

We thank Jeremy Brown, Brant Faircloth, Chris Austin, Ryan Eldridge, Frank Anderson, John McCormack, and two anonymous reviewers for helpful comments on earlier drafts of the manuscript. Mike Harvey provided guidance as we conducted laboratory procedures. Jeremy Ash, Jeremy Brown, and Xin Chen provided us with tools and assistance as we explored the results of our summary species tree analyses. Brant Faircloth assisted in the development of the SNP processing pipeline. Portions of this research were conducted with high performance computing resources provided by Louisiana State University (http://www.hpc.lsu.edu). We thank the Field Museum of Natural History (Lawrence Heaney, John Phelps, and William Stanley) and the University of Kansas Biodiversity Institute (Robert Timm) for access to specimens. For their contributions to fieldwork, we also thank Rafe Brown, Danilo Balete, Steve Goodman, and Carl Oliveros.

REFERENCES


