Frontiers in Zoology

RESEARCH

Open Access

Using natural history to guide supervised machine learning for cryptic species delimitation with genetic data

Shahan Derkarabetian[1*], James Starrett[2] and Marshal Hedin[3]

Abstract

The diversity of biological and ecological characteristics of organisms, and the underlying genetic patterns and processes of speciation, makes the development of universally applicable genetic species delimitation methods challenging. Many approaches, like those incorporating the multispecies coalescent, sometimes delimit populations and overestimate species numbers. This issue is exacerbated in taxa with inherently high population structure due to low dispersal ability, and in cryptic species resulting from nonecological speciation. These taxa present a conundrum when delimiting species: analyses rely heavily, if not entirely, on genetic data which over split species, while other lines of evidence lump. We showcase this conundrum in the harvester *Theromaster brunneus*, a low dispersal taxon with a wide geographic distribution and high potential for cryptic species. Integrating morphology, mitochondrial, and sub-genomic (double-digest RADSeq and ultraconserved elements) data, we find high discordance across analyses and data types in the number of inferred species, with further evidence that multispecies coalescent approaches over split. We demonstrate the power of a supervised machine learning approach in effectively delimiting cryptic species by creating a "custom" training data set derived from a well-studied lineage with similar biological characteristics as *Theromaster*. This novel approach uses known taxa with particular biological characteristics to inform unknown taxa with similar characteristics, using modern computational tools ideally suited for species delimitation. The approach also considers the natural history of organisms to make more biologically informed species delimitation decisions, and in principle is broadly applicable for taxa across the tree of life.

Keywords: Integrative taxonomy, Multispecies coalescent, RADSeq, Short-range endemism, Southern Appalachians, Supervised machine learning, Ultraconserved elements

Background

Organismal diversity is underpinned by diversity in life history and ecological characteristics among taxa, which in turn produce different underlying genetic patterns at the population and species levels [1–5]. Biological characteristics can determine the process and type of speciation. For example, nonecological speciation (speciation without divergent natural selection) produces ecologically similar/identical species that are allo- or parapatric replacements of each other [6, 7] and is more likely in low dispersal taxa that also show niche conservatism [8]. These biological and ecological characteristics often lead to cryptic speciation across many plant and animal taxa, where they complicate the application of many commonly used species criteria, and species delimitation relies largely, if not entirely, on genetic data.

The underlying diversity in speciation processes challenges the idea that any single genetic species delimitation model can be universally applicable. For example,

*Correspondence: sderkarabetian@gmail.com
[1] Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, 26 Oxford St., Cambridge, MA 02138, USA
Full list of author information is available at the end of the article

multiple empirical studies have shown that commonly used multispecies coalescent (MSC) models can oversplit species level diversity in low dispersal taxa because such systems violate the underlying assumption of panmixia [9–14], a sentiment echoed in theoretical literature [15]. Regardless of ongoing debates (e.g., [16]), we argue simply that MSC implementations taken at face value, and with well-resolved genomic or sub-genomic data, have strong potential to inflate species numbers in low dispersal systems (e.g., [17]). A solution to overreliance on genetics is integrative species delimitation [18, 19]. However, in many poorly known or "cryptic biology" taxa (e.g., [20]), like minute animals that live under rocks and logs, integrating behavioral, ecological, and/or phenotypic data is challenging to impossible. Morphological conservatism resulting from niche conservatism [21] means that distinct species are often not morphologically diagnosable. In an integrative framework, some lines of evidence cannot be feasibly studied, while others are clearly conservative. This leads to a fundamental conundrum—how can we rigorously delimit species when genetic analyses are biased to inflate, and other evidence is inaccessible or lumps evolutionarily significant diversity?

Many taxa in the arachnid order Opiliones present challenges for species delimitation. Their microhabitat specificity and low dispersal ability leads to nonecological speciation and high population genetic structure, where related congeners rarely co-occur in sympatry ruling out direct tests for reproductive isolation (e.g., [10, 22–29]). The biological characteristics associated with nonecological speciation in low dispersal Opiliones make several commonly used species criteria inapplicable or inappropriate in these systems. For example, morphological conservatism diminishes the utility of the morphological species criteria, and niche conservatism precludes ecological species criteria. In these cases, genetic data become the primary data type for species delimitation. However, these complexes represent classic cases of "too little gene flow", where distinguishing population genetic structure from species level divergence is not easy, and genetic species delimitation analyses overestimate species diversity (e.g., [10, 30, 31]).

Dispersal-limited microhabitat specialists are found in a diverse array of other taxa, including vertebrates and plants, with equal difficulty in resolving species-population boundaries (e.g., [12, 14]). This under-appreciated issue remains one of the most difficult challenges for empirical species delimitation and its implications extend to a diverse array of taxa regardless of biological characteristics (e.g., [32]). A possible solution to this issue is to use information from known systems to infer the unknown. In practice, this means using information derived from taxa with robust well-established species limits to infer species limits in a difficult cryptic species complex that shares similar biological and ecological characteristics and mode of speciation. Supervised machine learning is ideally suited for this approach, as known labeled data sets can be used to train a model that is then applied to an unknown unlabeled data set.

Here we use a combination of somatic and reproductive morphology, mitochondrial DNA, double-digest RAD-Seq (ddRAD; [33]), and ultraconserved elements (UCEs; [34, 35]) to illustrate the species conundrum in *Theromaster brunneus* [36], a widely distributed species from the southern Appalachian Mountains with high potential for cryptic speciation (Fig. 1). Our goal is not to exhaustively test species limits using every data and analysis type, but instead to demonstrate the difficulty of delimiting species in such taxa using common genetic species delimitation approaches. We highlight and emphasize a novel supervised machine learning approach, using training data from known taxa with similar biological characteristics as *T. brunneus*, to effectively (and conservatively) delimit cryptic species using phylogenomic data.

Results

Taxon sample

Our taxon sample included specimens from 76 different localities (Additional file 1: Table S1), 18 of which were only available for morphological study (i.e. preserved in 70–80% ethanol). All specimens are deposited in the San Diego State University Terrestrial Arthropod Collection. Locality data for all specimens housed in this collection (with SDSU_TAC or SDSU_OP catalog numbers) have been deposited at the Symbiota Collections of Arthropods Network (https://scan-bugs.org/portal).

Historically, the type specimens of *Theromaster brunneus* and *T. archeri* have never been directly compared by those who described or made any taxonomic acts affecting these species [37–39]. Further complicating the issue, these two species are reported from the same cave (McFarland Cave, AL) [37, 38]. Sympatry among congeneric Laniatores is extremely rare in northern temperate taxa. Several observations of *Theromaster* exist on iNaturalist from caves in northeast Alabama, however all of the individuals photographed are juveniles, which do not possess diagnostic species-specific characters.

Our examination of the male holotype of *T. archeri* did not alleviate any uncertainty. The genitalia and both pedipalps were previously dissected from the specimen and are no longer associated with the vial. As such, we cannot conduct comparative genitalic analyses or even confirm the sex of the specimen; all female *Theromaster* lack the cheliceral projections that are diagnostic for *T. brunneus*. It is possible that *T. archeri* is a local cave-adapted form

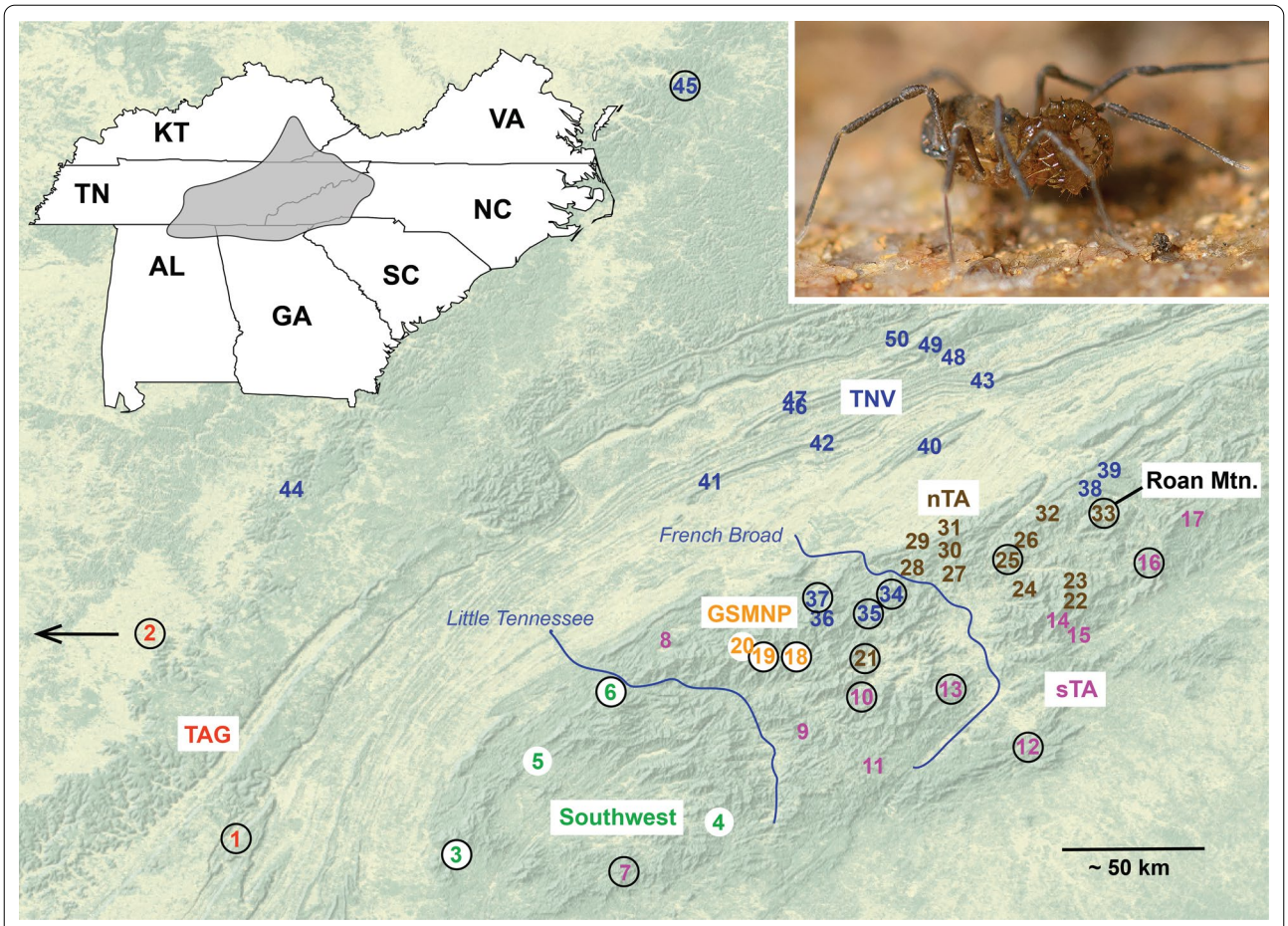

Fig. 1 Distribution of *Theromaster brunneus* sampled in this study. Colors correspond to lineages identified in phylogenomic analyses (see Fig. 3). Numbers correspond to locations included in ddRAD analyses (Additional file 1: Table S1); sites with an enclosed circle were also sequenced for UCEs. Left inset: general distribution of *Theromaster*. Right inset: live photo of *T. brunneus*. Clade names: Great Smoky Mountain National Park (GSMNP), Tennessee Valley (TNV), northern Trans Asheville (nTA), southern Trans Asheville (sTA)

of *T. brunneus* at the southern limits of the distribution, however neither the description nor our examinations suggest any cave adaptation. We unfortunately cannot confirm or deny species limits and therefore must retain *T. archeri* as a distinct species based on historical work. The final determination will likely only be possible with DNA sequencing of the holotype specimen (i.e., [40]), although we do not expect either outcome (a valid species or local population of *T. brunneus*) to affect species delimitation results in this study. It could be possible that our TAG lineage (see below) corresponds to *T. archeri.*

Morphological analyses

Voucher specimen data and accession numbers are provided in Additional file 1: Table S1. Representative morphological images (Figs. 2, 3, Additional file 2: Figs. S1–S5) confirm the highly conservative morphology across *Theromaster*, including male genitalia and male cheliceral modifications, a sexually dimorphic feature. However, there is clear differentiation in the habitus morphology of Roan Mountain (OP322, location #33), which has a more pointed eye mound, dorsal tergites clearly separated by grooves, more obvious dorsal spines, and more defined pigmentation patterns (Fig. 3). This specimen is clearly distinguished based on the morphological species criterion, suggesting at least two putative species.

Phylogenetic and clustering analyses

Accession numbers for all samples sequenced in this study are provided in Additional file 1: Table S1. RAxML analysis of cytochrome c oxidase subunit I (COI) was mostly poorly supported but revealed highly divergent Roan Mountain and "Southwest" lineages. RAxML analyses of ddRAD data (~75% taxon occupancy matrix with 1001 loci and 89,663 nucleotides) show at least seven main lineages (Fig. 3, Additional file 2: Fig. S7, Additional

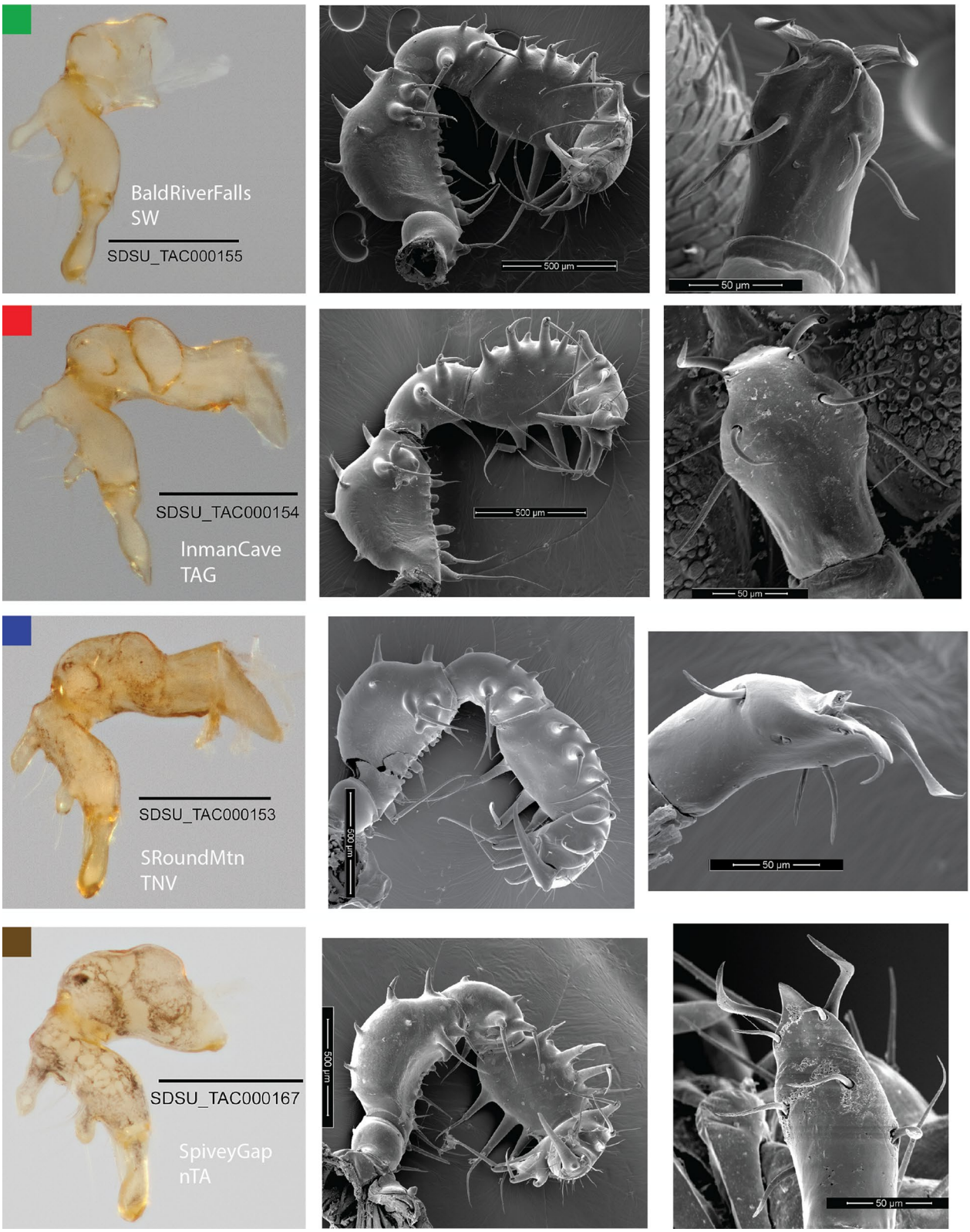

Fig. 2 Representative morphological images for select lineages of *Theromaster brunneus*. Each row of images corresponds to a single male individual, with the voucher information indicated in the left-most image. Columns of morphological characters correspond to chelicerae, pedipalp, and penis (L–R). Colors correspond to lineages identified in phylogenomic analyses (see Fig. 3)

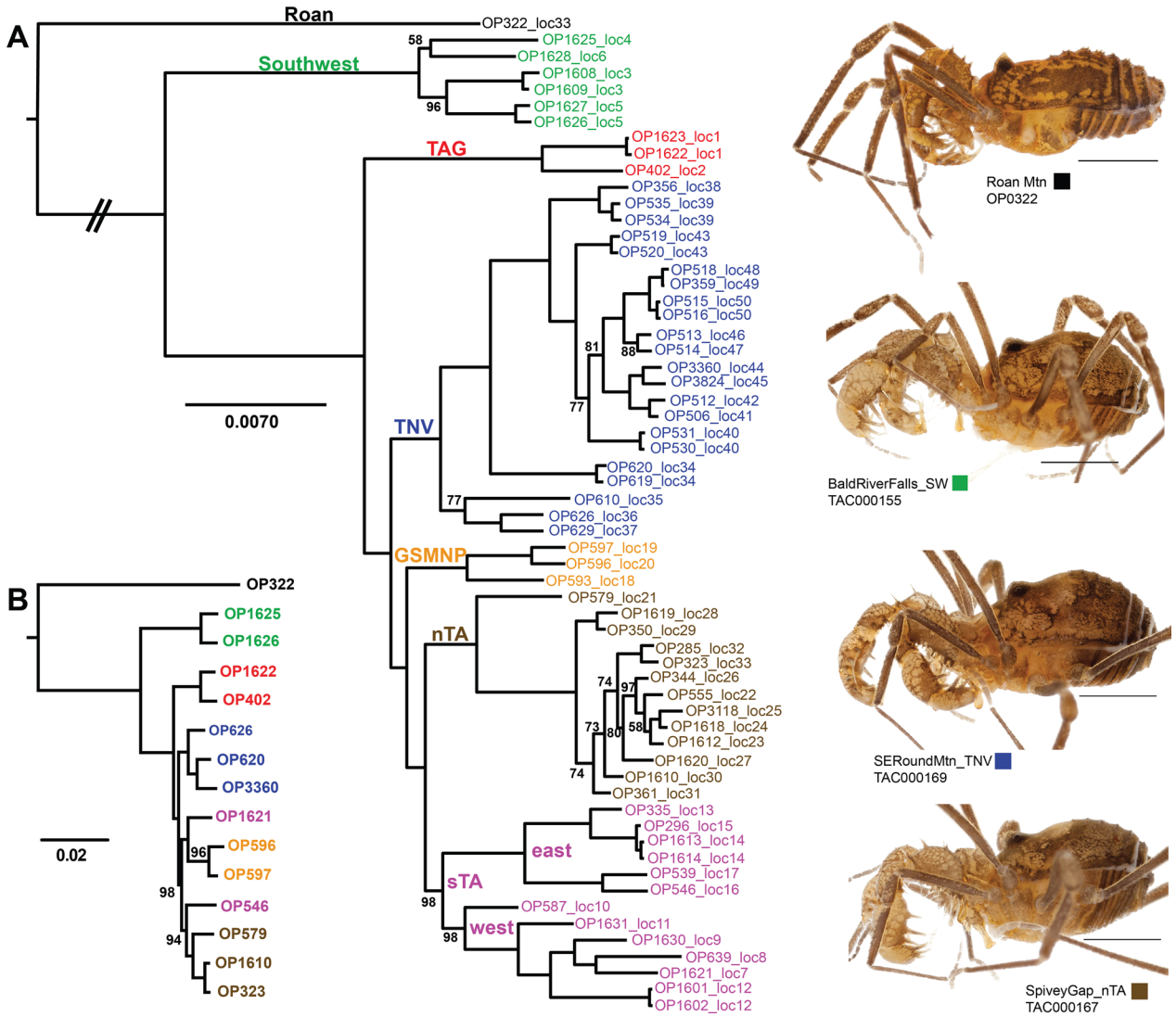

Fig. 3 Phylogenomic results based on RAxML analysis of ddRAD and UCE data, with representative habitus images of select lineages. All nodes have 100% bootstrap support unless indicated. **A** ddRAD phylogeny of the 61_48 dataset (see Additional file 1 for matrix details). TNV, GSMNP, nTA, and sTA lineage samples were included in DAPC and STRUCTURE analyses. **B** UCE RAxML phylogeny based on the 70% taxon occupancy matrix

file 1: Tables S2–3). Relationships among these lineages are all highly supported (bootstrap of 100), but a handful of nodes within some clades show lower support, although all but one of these nodes are still above 70. SVDQuartets analyses result in similar lineage composition and interrelationships, but with generally weaker support (Additional file 2: Fig. S8). Excluding a single instance of sympatry at Roan Mountain, all main lineages are allopatric (Fig. 1).

Both partitioned RAxML and SVDQuartets analyses of the concatenated UCE matrix (70% taxon occupancy with 324 loci and 108,875 nucleotides) generally supported the ddRAD topology, with Roan Mountain, Southwest, and TAG as divergent and early-diverging lineages (Fig. 3, Additional file 2: Fig. S9). However, in both the RAxML and SVDQuartets trees, the "southern Trans Asheville" (sTA) clade was not recovered as a monophyletic group. All nodes in the UCE RAxML tree have a posterior probability greater than 0.95, with the exception of a single node with 0.94. DAPC clustering of the ddRAD SNPs reveal optimal K values (Additional file 2: Fig. S10) that correspond to the main lineages, with further subdivisions for the sTA and "Tennessee Valley" (TNV) groups recovered in phylogenetic analyses. STRUCTURE analyses of ddRAD SNPs likewise recover the main lineages, but as K values increase, further subclusters are

found that correspond to geographic and/or phylogenetic groupings (Additional file 2: Fig. S11). The best-fit K value is K = 3 using the ΔK method [41], but K = 10 using the Pritchard et al. [42] method. VAE analyses cluster samples as in other analyses, and UCEs show more overlap among groups (Additional file 2: Fig. S12).

Phylogenomic species delimitation

The COI bPTP analyses supported 11 species (Additional file 2: Fig. S6), splitting most main genetic lineages into at least two species. BFD* hypothesis testing results are presented in Fig. 4 and Additional file 1: Tables S4–S5. Matrices consisted of 1828 and 415 SNPs for ddRAD and UCE datasets, respectively. The ddRAD dataset favored K = 3 corresponding to Roan Mountain, Southwest, and all other samples, however, the likelihood continuously increased with increasing number of species (Fig. 4A, B). BFD* analyses for the UCE dataset favored all samples as species (K = 15), with a second less likely peak at K = 3 (Fig. 4C, D).

CLADES requires that every predefined population, defined here as the main ddRAD genetic lineages, have at least one representative sequence present at every given locus in the input file. As such, the final *T. brunneus* dataset included 52 UCE loci. CLADES analyses of the UCE SNP dataset based on the "general" training dataset favored all populations as species, while analyses based on the "custom" training dataset favored two cryptic species (Fig. 5), corresponding to Roan Mountain and Southwest clade. While a "ground truth" for what the

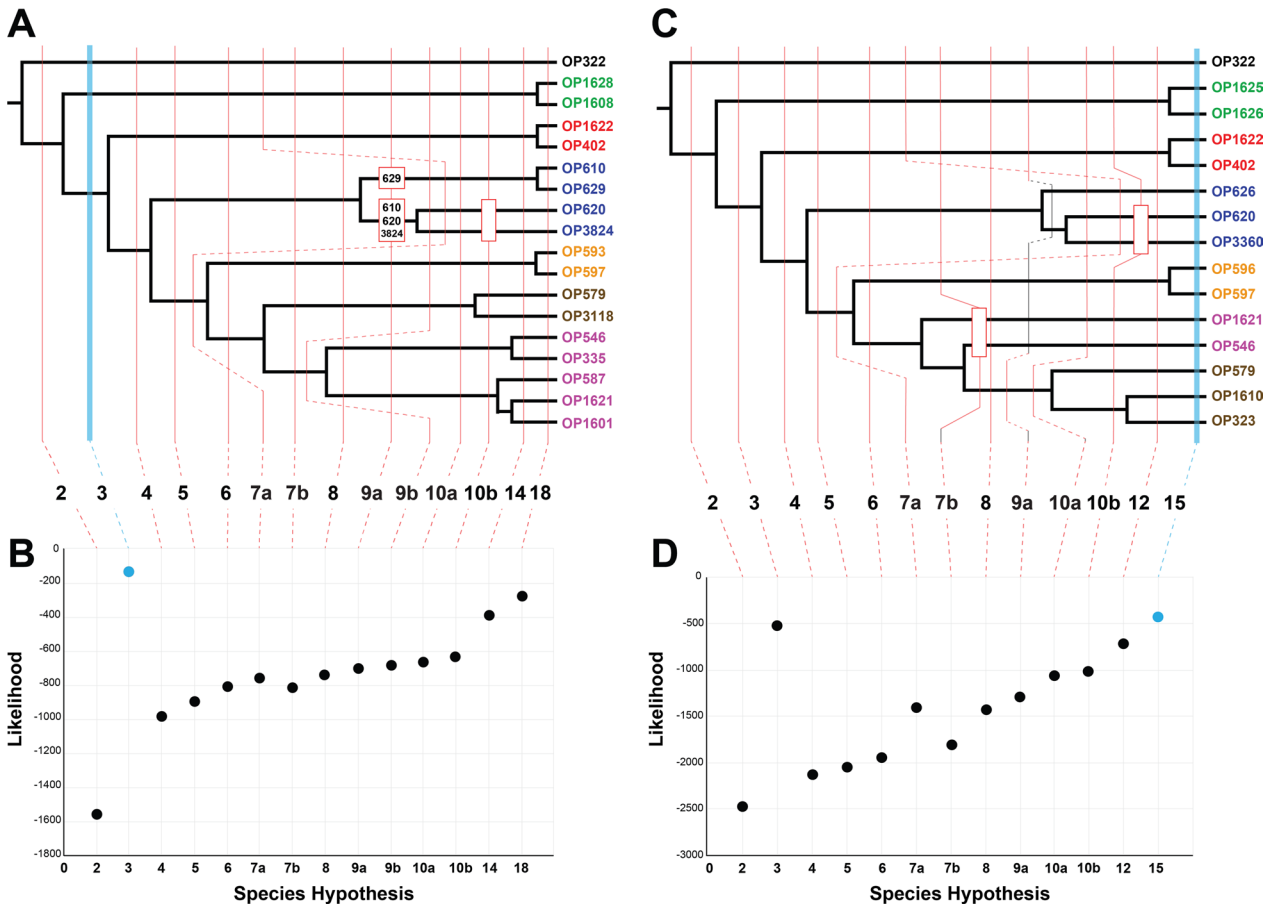

Fig. 4 Results of SNAPP/BFD* species delimitation analyses based on SNP datasets. **A**, **B** ddRAD phylogeny based on RAxML analysis of the 18_12 taxon occupancy matrix, with vertical lines indicating species hypotheses tested, and likelihood estimates for each hypothesis (averaged from two runs). **C**, **D** UCE phylogeny based on RAxML analysis of the 70% taxon occupancy matrix, with vertical lines indicating species hypotheses tested, and likelihood estimates for each hypothesis (averaged from two runs). Note: branch lengths for both cladograms were adjusted to fit vertical lines for easier interpretability and have no relative meaning; the phylogeny is purely to depict branching pattern and hypotheses tested. Blue vertical lines in **A** and **C** and blue data points in **B** and **D** indicate species hypothesis decisively favored by Bayes Factor analyses (see Additional file 1: Table S5 for values). Colors correspond to lineages identified in phylogenomic analyses (see Fig. 3)

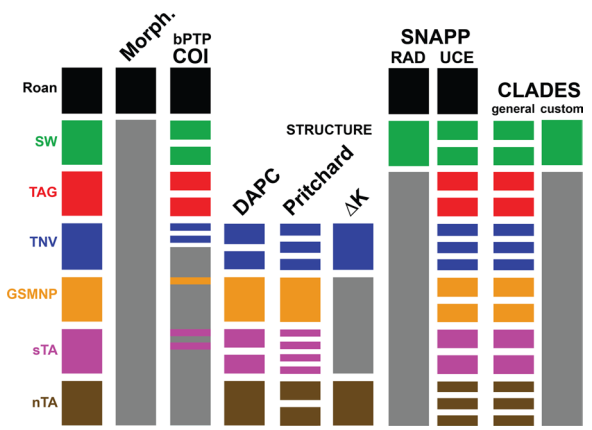

Fig. 5 Species delimitation results supported across data and analysis types. For any given analysis, multiple bars of the same color indicate lineages that were split in that analysis, and any lineages that were lumped in any given analyses are colored gray. Colors correspond to lineages identified in phylogenomic analyses (see Fig. 3)

actual cryptic species are is lacking, the effectiveness of a "custom" dataset relative to a "general" dataset can be assessed when comparing assignment probabilities of specimens that are from the same geographic location. In these cases, the probability that specimens from the same location belong to the same species should be relatively high due to the allopatric nature of species in low dispersal taxa undergoing nonecological speciation. When using the "general" training dataset, specimens from the same geographic location were classified on average as the same species with 0.116 probability, while with the "custom" dataset, specimens from the same population were classified on average as the same species with 0.753 probability.

Although we have high confidence in our species delimitations, we refrained from formally describing the two new species until more specimens of the undescribed species from Roan Mountain can be collected.

Discussion

Major portions of the tree of life include branches (species) that are unknown or poorly known, and integrative studies in such taxa are challenging for many reasons. At the same time, collecting sub-genomic data for these taxa has become increasingly easy, leading to species delimitations founded largely, if not entirely, upon genetic data. Many studies have documented overestimation of species numbers by commonly used genetic delimitation methods both in empirical (e.g., [9–11, 13, 14, 31, 32, 43, 44]), and theoretical/simulation studies [12, 15, 17, 45, 46]). Increasing the number of loci may also increase support for an incorrect hypothesis in Bayesian analyses [47], and most relevant here, supporting species level divergences for intraspecific populations [44]. Researchers studying cryptic species with inherently high population structure and allo- or parapatric distributions face a dilemma when delimiting species. Current phylogenomic species delimitation analyses overestimate species numbers, and without morphology, behavior, or distribution to assist, the degree of over-splitting is unknown. Systematists researching these taxa must be conservative in final species hypotheses (e.g., [17]), while simultaneously acknowledging that the actual number of species may still be underestimated.

In our study, most genetic analyses recover at least six lineages as species (Fig. 4), but the consensus favors three species (Roan, SW, all others), the latter two of which are morphologically cryptic. The supervised machine learning approach we employed with a "custom" training data set provided a reasonable and biologically informed hypothesis of three total species. Phylogenies derived from ddRAD and UCEs were essentially congruent, but the inferred number of species in the BFD* analyses differed. Both had a peak at K = 3, but UCE data had a higher likelihood at K = 15, and we argue that leading to this peak at K = 15 in the UCE data, the MSC oversplits because these taxa clearly violate the assumption of panmixia within species. BFD* analyses using ddRAD data did not overestimate species. However, like UCE analyses, there is an obvious increasing trend in likelihood with increasing species numbers leading us to ask if K = 3 would still be the most likely hypothesis if more sequenced samples from different collecting localities (i.e., populations) had been included. As a validation approach, only a limited set of species-level hypotheses are tested in empirical studies using BFD*. It might be beneficial for researchers using this approach, especially for low dispersal taxa, to fully explore hypotheses to see if "false positives" are more prevalent.

Training data justification and considerations

In this study we present a first attempt at demonstrating the potential of a supervised machine learning approach to genetic species delimitation using biologically relevant customized training data sets. Here we provide some justification for our training data set choice and considerations for future work. In the case of Opiliones, which are largely understudied from a modern genetic perspective, there is scant genome-scale species level data available to serve as training data. Many Opiliones studies focusing on species delimitation using genetic data to identify or conclude the presence of cryptic species, resulting in uncertainty across species boundaries (e.g., [30, 31, 43]). The level of congruence and support for species delimitations seen in *Metanonychus* is exceptional [10], making

this the most suitable training dataset for low-dispersal Opiliones to date, with potential application to many other unresolved putative cryptic species complexes.

Genetic statistics can be useful in identifying the overlap of actual and potential species boundaries between data sets. For UCE loci, the mean K2P-corrected genetic distance across all *T. brunneus* samples is 2.98% (range 0.37–8.92) and falls within the range of genetic distances seen across species of *Metanonychus* (mean = 14.746, range = 1.59–26.91). Speciation events within *Metanonychus* span recent and older divergences, where the mean K2P-corrected divergence of COI across the shallowest species-level split is 13.15% and the deepest split has a mean divergence of 27.15%. Mean COI divergence across *Theromaster* samples is 6.51%, with a maximum of 20.2%. Taken together, these genetic measures indicate that any potential species-level divergences in *T. brunneus* fall within the range of actual species divergences seen in *Metanonychus*. Sensitivity to the underlying model is an obvious and related consideration. In this regard, in taxa with better representation of genome-scale species level data, it would be worthwhile to explore varying combinations of suitable taxa in the training data set. For example, including shallower or deeper species divergences, or data from a phylogenetically more diverse range of taxa (i.e., from other genera or families with similar biological characteristics).

One concern relates to the differences in dispersal dynamics between the regions each taxon is found in (Pacific Northwest for *Metanonychus* and southern Appalachians for *Theromaster*). Dispersal dynamics and bio-/phylogeographic histories are different across these regions, where geologic and other abiotic factors (e.g., river formation) can drive the speciation process, especially for the ancient and more topographically complex southern Appalachian Mountains. Despite these differences, given the biological and ecological similarity of these taxa, we hypothesize that while the dynamics differ across regions, their *responses* and associated genetic signatures to any abiotic factors influencing speciation will be similar. It follows that using taxa that inhabit the same region should increase the effectiveness of our supervised approach, much in the same way as comparative phylogeography is a powerful approach for elucidating common underlying regional biogeographic patterns (e.g., [48, 49]).

There are many questions in relation to the choice of training and testing datasets that deserve further attention. How closely related must the taxa be to be considered similar enough for this approach? What is an appropriate divergence date threshold? How ecologically similar (e.g., degree of climatic variable overlap) should they be? Questions relating to similarity and divergence dates can be explored with further data sets, as well as the relative importance of genetic versus ecological similarity between training and testing taxa. This choice will in fact be dependent on the organismal type and may ultimately need to be somewhat subjective, where the experience and organismal knowledge of the taxonomist will be critical in determining suitability of any training data set. Niche overlap can be quantified, however, in the case of our study the differences in local climates and the geographic distance between their respective distributions makes assessing niche overlap difficult. More importantly, the bioclimatic variables used in species distribution modelling do not necessarily capture the similarity in microhabitat "climate" for taxa found underneath the forest surface, living in deep leaf litter and underneath woody debris. Future studies should advance towards quantifiable metrics that determine if a species group(s) is an appropriate training dataset, as well as attempting this approach on taxa that are more directly linked to the bioclimatic variables used in species distribution modelling (e.g., plants).

Incorporating natural history into genetic species delimitation

Genetic species delimitation is driven largely by computational tools, model testing, and a desire for objectivity in analyses. However, this is increasingly at the expense of considering the biological characteristics of organisms. Cryptic species, and those taxa that undergo non-ecological speciation, are one of the biggest challenges in species delimitation, as many species criteria cannot be used in practice. Moving forward, an additional approach to delimiting cryptic species with genetic data can be using information already available, in this case inferring species in difficult unknown systems by using data from robustly known taxa with similar biological characteristics and modes of speciation. Recent integration of machine learning in species delimitation [10, 50, 51] provides algorithmic options which are versatile and customizable. Here we used a system-specific "custom" training dataset in a supervised machine learning framework to delimit cryptic species, where the training data were derived from *Metanonychus*, a previously studied system with robust species supported through multiple data types and with similar biological characteristics to *Theromaster*. These taxa share similar biological and ecological characteristics, most importantly dispersal ability and microhabitat preference, and both undergo the same type of speciation, and as such are expected to have comparable underlying genetic patterns associated with populations and species. The effectiveness of using customized and biologically relevant training data is evidenced by the probability of assigning specimens from

the same geographic location to the same species, which increased dramatically with the "custom" dataset relative to the "general".

The power of applying a supervised machine learning approach derives from the ability to create custom training data sets that are specific to each study system, and to various classes of genetic data (e.g., UCE, RADseq, Sanger), capturing the inherently different characteristics of genetic data types. In this way, our approach combines a computational tool ideally suited for species delimitation, in this case, a supervised machine learning algorithm as a classification tool, with knowledge of the biology and natural history of the focal organisms derived from organismal expertise, leading to more informed, relevant, and reasonable species delimitation decisions when relying on genetic data only. The recently developed program DELINEATE [46] takes a similar approach, using the information from known species to calculate speciation parameters which are then applied to delimit unknown samples. Similarly, a reference-based taxonomic approach was used to delimit putative new species based on genetic distances of known species in a group of closely related and ecologically similar lizards [52].

Conclusions

There are extremely well-studied systems for which genetic species delimitation is largely successful, for example the model organism *Drosophila* [53]. However, for poorly studied groups (perhaps the majority of life) where basic biological and ecological knowledge can be difficult or impossible to acquire, inferring any biological details in an unknown or new taxon commonly relies on generalization from similar or closely related taxa where that information happens to be known. Our approach using known species limits in a supervised machine learning framework to infer unknown limits is a logical analytical extension of this inference process and should be universally applicable to species delimitation in any taxon, particularly when cryptic species are anticipated or prevalent.

Materials and methods

Study system and taxon sampling

The genus *Theromaster* [37] currently includes two described species: the widespread *T. brunneus* [36] (Fig. 1), and the poorly described *T. archeri* [38] from several caves in extreme northeastern Alabama. *Theromaster* are small (body length usually < 3 mm), short-legged, and most often found in sheltered microhabitats under rocks and logs. Because of these natural history characteristics, we anticipate high levels of population genetic structuring and potential cryptic diversification, as seen in many other northern temperate Opiliones with similar biology (e.g., [24–26, 43]). Cryptic diversification is further expected because *T. brunneus* occurs in one of the most topographically complex regions of North America, the southern Appalachian Mountains. The Southern Appalachians are a well-known biodiversity hotspot for animals including vertebrates (e.g., [54–58]) and arthropods (e.g., [26, 59–61]). For arachnids, almost all available molecular datasets for "wide-ranging" taxa in this region indicate in situ phylogeographic diversification, with multiple lineages that likely represent cryptic species (e.g., summarized in [43, 61, 62].

Theromaster brunneus is known from less than 10 literature records [37–39, 63] from western North Carolina, eastern Tennessee, northern Georgia, and northern Alabama. However, our own collections and museum specimens indicate a broader distribution (Fig. 1) that is atypically large for a single species of northern temperate laniatorean Opiliones. We first constructed a distribution map for *T. brunneus*, based on original collections from the Hedin lab and collections of Dr. William Shear (specimens now housed at San Diego State University). Our specimen sample spans the known geographic distribution of the species, including specimens from near the type locality ("valley of Black Mountains", North Carolina). All samples used in this study are identified as or considered *T. brunneus*; specimens morphologically identifiable as the questionable *T. archeri* could not be collected. To attempt to resolve the taxonomic issues associated with *T. archeri* we also examined the holotype specimen held in the American Museum of Natural History. Previous UCE-based phylogenomic analyses of travunioid harvestmen strongly supported a *Theromaster* + *Erebomaster* clade [64]; as such we used *Erebomaster* samples as outgroups for mitochondrial and UCE datasets, and rooted ddRAD phylogenies (without *Erebomaster*) based on the UCE topology.

Morphology

Adult male *T. brunneus* have distinctive cheliceral modifications, making this taxon easily recognizable. Whole specimen and cheliceral segment digital images were captured using a Visionary Digital BK plus system (http://www.visionarydigital.com). Multiple individual images were merged into a composite image using Helicon Focus 6.2.2 software (http://www.heliconsoft.com/heliconfocus.html). For imaging, left chelicerae and pedipalps were dissected from male specimens. Male penises that were not already protruding from the genital operculum were physically extracted using a blunt insect micro pin. Chelicerae, penis, and pedipalps were examined using scanning electron microscopy (SEM). Specimens destined for SEM imaging were mounted onto stubs, critical point

dried, coated with 6 nm platinum, and imaged on the FEI Quanta 450 FEG environmental SEM at the San Diego State University Electron Microscope Facility.

Mitochondrial data collection and analysis

A partial fragment of the mitochondrial COI gene was amplified using PCR primers and conditions as in Hedin and Thomas [26] and Derkarabetian et al. [65] for a total of 39 T. *brunneus* samples from throughout its distribution. PCR products were purified using Millipore plates, Sanger sequenced in both directions at Macrogen USA (Rockville, MD), then edited and aligned manually using Geneious 10.1 (Biomatters Ltd.). Gene trees were reconstructed using RAxML v8 [66], with a GTR GAMMA model applied to separate codon partitions. RAxML was called as follows: -# 500, -n MultipleOriginal, -# 1000, -n MultipleBootstrap. The resulting COI RAxML gene tree was used as input for bPTP species delimitation analyses through the bPTP server (https://species.h-its.org/, [67]). Two replicate analyses were run for 100,000 generations, with thinning at 100, and a burnin of 0.1.

ddRADSeq data collection and analysis

Sixty-one *Theromaster* specimens from 50 distinct geographic locations were included in a "complete" (n = 61 samples) matrix for ddRAD analyses (Fig. 1). Preparation of the ddRAD libraries followed Burns et al. [68] and Derkarabetian et al. [24], adapted from Peterson et al. [33]. DNA was extracted from whole specimens using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA) following the manufacturer's protocol. We used restriction digest enzymes *EcoRI*-HF and *Msp1* (New England Biolabs, Ipswich, MA), and the corresponding adapters from Peterson et al. [33]. Briefly, ~ 500 ng of genomic DNA was digested for 3 h in a 50 µl reaction with 100 units each of the restriction enzymes *EcoRI*-HF and *Msp1* (New England Biolabs, Ipswich, MA), and 1X CutSmart Buffer (New England Biolabs). Samples were purified using Agencourt AMPure XP bead cleanup (Beckman Coulter, Inc., Brea, CA). Adapters were ligated to digested DNA in a 40 µl reaction that consisted of 33 µl digested DNA, 1.05 µM *MspI* P2 adapter, 0.54 µM *EcoRI* P1, 400 units of T4 DNA-ligase, and 1X T4 DNA ligase reaction buffer (New England Biolabs). Ligation reactions were incubated at room temperature for 40 min, heat killed at 65 °C for 10 min, then cooled to room temperature at a rate of 2 °C per 90 s. Samples with different adapters were pooled by column and then purified using AMPure XP bead cleanup. Pooled samples were then size selected to a size range of 400–600 bp with a Pippin Prep automated size-selection instrument (Sage Science, Beverly, MA). Primers with Illumina indices were added to the pooled samples using PCR; 50 µl reactions consisted of 23 µl DNA template, 2 uM PCR Primer P1, 2 uM PCR primer P2 (eight types for second-tier multiplexing, one per pooled sample), and 1X Phusion High Fidelity PCR Mastermix (New England Biolabs). Cycle conditions were 98 °C for 30 s, 12 iterations of 98 °C for 10 s and 72 °C for 20 s (with a 16% ramp to slow cooling), and 72 °C for 10 min. PCR products were purified via AMPure XP bead cleanup and quantified using a Bioanalyzer (Agilent Technologies, Santa Clara, CA). A pool consisting of an equimolar quantity of each library was sequenced as 100 bp SE reads on the Illumina HiSeq2500 platform at the University of California, Riverside's Institute for Integrative Genomics Biology—Genomics Core Facility.

ddRAD data were processed using the *denovo* assembly method of ipyrad v.0.5.15 [69], with the following settings adjusted from default: mindepth_majrule = 6, clust_threshold = 0.9, filter_adapters = 2, filter_min_trim_len = 35, max_Indels_locus = 4, max_shared_Hs_locus = 0.1. For the full 61-sample dataset, we ran min_samples_locus at 31 (50% complete matrix, called 61_31) and 48 (~ 75% complete matrix, called 61_48). Maximum likelihood analyses of 61_31 and 61_48 matrices were run with RAxML v8 [66] using 1000 rapid bootstrap replicates and the GTRGAMMA model. These analyses of 61-sample matrices indicated three divergent and early-diverging lineages (see "Results" section). Given the congruent recovery of three early-diverging lineages, we excluded these early-diverging samples and re-ran min_samples_locus at 45 and 51 for a reduced 51-sample dataset (called 51_45 and 51_51 respectively). These datasets only included "Tennessee Valley" (TNV), Great Smokey Mountain National Park (GSMNP), "northern Trans Asheville" (nTA), and "southern Trans Asheville" (sTA) lineages. This strategy effectively increases the number of loci retained for these derived lineages (see [70]).

Using unlinked SNPs (a single randomly sampled SNP per locus) from the 61_48 matrix, we reconstructed both a "lineage tree" (individuals as OTUs) and "species tree" using SVDquartets [71] in PAUP * v4.0a152 [72] with n = 500 bootstraps. For the species tree, specimens were partitioned into groups following major clades recovered in RAxML and SVDquartets lineage tree analyses. Using the 51_45 unlinked SNPs matrix (n = 1122 unlinked SNPs), we performed k-means clustering of PCA-transformed data using the find.clusters R function ("adegenet" package) [73, 74]. Missing data were replaced by the mean frequency of the haplotype in the sample (scaleGen(data, NA.method = "mean"). The Bayesian information criterion (BIC) was used to compare clustering models with a maximum of K = 20, retaining all principal components, and replicating the analysis 10 times.

We then conducted a discriminant analysis of principal components (DAPC), retaining approximately one-quarter of the principal components and all discriminant functions. Using the same unlinked SNPs matrix, STRUCTURE 2.3.4 [42] runs were conducted using an admixture model with uncorrelated allele frequencies. All other priors were left as default. For individual K values ranging from 2–12, analyses were replicated four times, each run including 200,000 generations with the first 20,000 generations removed as burnin. Data were summarized using CLUMPAK [75], with a best-fit K chosen utilizing the ΔK method of Evanno et al. [41], and the prob (K) method of Pritchard et al. [42].

Based on phylogenetic analysis of the UCE and 61-sample RADSeq datasets, plus STRUCTURE [42] and DAPC [76] analyses of the 51-sample RADSeq dataset (see "Results" section), we chose 18 samples to represent all primary *Theromaster* lineages. For these 18 samples we re-ran ipyrad using settings as above, at 18_12 occupancy. From this, an unlinked SNPs file was created using the Phrynomics package ([77], https://github.com/bbanbury/phrynomics), where nonbinary characters were removed and bases were translated. To further visualize genetic structure and clustering within *Theromaster*, we analyzed this dataset using a Variational Autoencoder (VAE) implemented with a modified version of the sp_deli script (https://github.com/sokrypton/sp_deli) derived from Derkarabetian et al. [40]. The matrix was run through the VAE five times and the analysis with the lowest average loss after removing 50% burnin was considered the optimal output.

UCE data collection and analysis

Studies have shown the utility of UCEs at shallow levels (e.g., [78, 79]), particularly in arthropod taxa (e.g., [40, 80, 81]). The UCEs targeted in the arachnid probe set [82] are exonic in origin with the "core" UCE corresponding to coding region while the "flanking" region are non-coding introns [83]. As such, in taxa with high population structure, flanking non-coding regions of UCEs are informative for population level structure and can be used in phylogenomic species delimitation analyses [40]. Representative samples for UCE sequencing were chosen based on preliminary RAD analyses, subsampling main lineages (see Fig. 2).

Sequence capture of UCEs followed the protocol of Derkarabetian et al. [64]. A subset of 15 samples representing all major ddRAD lineages were used in UCE experiments and were prepared in multiple library preparation and sequencing experiments. Protocols across these experiments were largely identical, differing mainly in sequencing platform. Genomic DNA was extracted from either multiple legs or whole bodies using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA). Extractions were quantified using a Qubit Fluorometer (Life Technologies, Inc.) and quality was assessed via gel electrophoresis on a 1% agarose gel. Up to 500 ng were used in DNA fragmentation procedures, either using a Bioruptor or a Covaris M220 Focused-ultrasonicator, as in Derkarabetian et al. [64]. UCE libraries were prepared using the KAPA Hyper Prep Kit (Kapa Biosystems), using up to 250 ng DNA (i.e., half reaction of manufacturer's protocol) as starting material. Ampure XP beads were used for all cleanup steps. For samples containing < 250 ng total, all DNA was used in library preparation. Target enrichment was performed on pooled libraries using the MYbaits Arachnida 1.1K version 1 kit (Arbor Biosciences, Ann Arbor, MI) following the Target Enrichment of Illumina Libraries v. 1.5 protocol (http://ultraconserved.org/#protocols). Hybridization was conducted at either 60 or 65 °C for 24 h, with a post-hybridization amplification of 18 cycles. Following an additional cleanup, libraries were quantified using a Qubit fluorometer and equimolar mixes were prepared for sequencing either with an Illumina NextSeq (University of California, Riverside Institute for Integrative Genome Biology) with 150 bp PE reads, or an Illumina HiSeq 2500 (Brigham Young University DNA Sequencing Center) with 125 bp PE reads (see Suppl. Material 1).

Raw demultiplexed reads were processed with the Phyluce pipeline [84]. Quality control and adapter removal were conducted with the Illumiprocessor wrapper [85, 86]. Assemblies were created with Velvet [87] at default settings. Contigs were matched to probes using minimum coverage and minimum identity values of 65. UCE loci were aligned with MAFFT [88] and trimmed with Gblocks [89, 90] implemented in the Phyluce pipeline. All individual UCE loci were imported into Geneious 10.1 (Biomatters Ltd.) and manually inspected to check for obvious alignment errors and remove putatively non-homologous sequences (e.g., any sequences more divergent than outgroup taxa).

Concatenated and partitioned phylogenetic analyses were run on two datasets differing in the taxon coverage needed to include a locus in the final dataset: 50% and 70%. Maximum likelihood analyses were run with RAxML v8 [66] using 200 rapid bootstrap replicates and the GTRGAMMA model. Using the 70% concatenated UCE matrix we also reconstructed a lineage tree using SVDquartets [71] with n = 500 bootstraps. Finally, we made a 50% taxon coverage unlinked SNP dataset from alignments with a custom wrapper script using snp-sites [91] to convert alignments to vcf format, randSNPs_from_vcf.pl (https://www.biostars.org/p/313701/) to select a single random SNP from each alignment's vcf file, vcf2phylip.py (https://github.com/edgardomortiz/vcf2p

hylip) to convert vcf files back to phylip, and AMAS [92] to concatenate all randomly selected SNPs into a single phylip file. The Phrynomics R package ([77], https://github.com/bbanbury/phrynomics) was used to select only biallelic SNPs and translate SNPs to integers. The VAE was run on this dataset as done with ddRAD data.

Bayes factor delimitation* analyses

We conducted BFD* [93, 94] species delimitation analyses using SNPs derived from both the ddRAD and UCE data using SNAPP [95] implemented in the BEAST 2.4.5 package [96]. Analyses were run on the 18_12 ddRAD and the 50% taxon coverage UCE datasets. For each SNP dataset we tested multiple alternative species hypotheses. Hypotheses tested were derived from other data types and analyses used in this study including morphological, COI, and phylogenetic and STRUCTURE/DAPC analyses of nuclear data. To test the BFD* approach to its fullest extent in this study system (and hence its potential to delimit populations), we also included a nested set of hypotheses up to the maximum potential number of species, where every specimen was considered a different species. All BFD* analyses were run for 100,000 generations, with 10,000 generations as pre-burnin, 48 steps, and an alpha value of 0.3. Two replicates of each analysis were run to check for convergence. A comparison of marginal likelihoods was conducted using Bayes factors [97], with values above 10 considered to be decisive support.

Supervised machine learning analyses

We analyzed UCE loci with the supervised machine learning species delimitation program CLADES [50]. CLADES is a classification model derived from a type of machine learning algorithm called a support vector machine to classify samples as either "same species" or "different species" using multi-locus data in a two-species model. CLADES computes five summary statistics from the data (both training and testing) and uses these statistics as features to create the model and classify samples: private positions, folded-SFS with *k* bins, pairwise difference ratio, F_{ST}, and longest shared tract (defined in [50]). A training data set, where pairwise comparisons of all samples are defined a priori as either the same or different, is used to build the model and classify a test dataset with unknown species status.

We analyzed *Theromaster* UCE data in CLADES using two training data sets. First, Pei et al. [50] provided a training data set called "All" (which we refer to as "general" here) based on simulated data with varying values of theta (Θ), migration rate, and divergence time under a two species model. This "general" data set is meant to be broadly applicable across taxa as simulated data encompass the broad diversity of genetic patterns across plants and animals [50]. Second, we developed a "custom" training data set derived from the well-known, robust species of *Metanonychus* recently revised in an integrative taxonomic context [25]. All *Metanonychus* species are easily diagnosed based on both somatic and genitalic morphology, with both mitochondrial and nuclear data supporting species status across a broad array of analysis types. Most importantly, *Metanonychus* and *Theromaster* share similar biological and ecological characteristics, including low dispersal ability and microhabitat preferences. Microhabitat preference and ecological similarity are largely based on our experience collecting these taxa over many years. Quantifying biological and ecological similarity, as well as dispersal ability, is difficult in poorly-known taxa with cryptic biology, like those that occupy "hidden" microhabitats (see "Discussion" section for further justification).

For both *Theromaster* and *Metanonychus*, datasets included all UCE loci shared across all samples in each data set (n = 52 for *Theromaster*, n = 12 for *Metanonychus*) (the "spp" dataset of *Metanonychus* in [10]). To create the "custom" *Metanonychus* training dataset, we ran the *Metanonychus* UCE loci through CLADES against the "general" training dataset. As expected, and found in Derkarabetian et al. [40], analyses favored all populations as species, and all output files reflected this. The output files contain pairwise comparisons of all specified populations, those files with pairwise comparisons between populations that belonged to the same species as delimited in Derkarabetian et al. [10] were manually modified to reflect that the samples belonged to the same species (switching + 1 to − 1). All relevant files required for the model (see CLADES documentation) were manually created from these output files. LibSVM [98] was used to create the "*.model" file from the "*.sumstat.scale" file using default parameters, with the addition of training for probability estimates (− b 1).

Following creation of this "custom" training dataset, we ran the *T. brunneus* dataset against it using CLADES. We excluded the Roan sample from these analyses because this specimen (i.e., species) is morphologically distinguishable from the rest of *T. brunneus* and is represented by only a single specimen. In this way our goal was to assess the success of this approach on a dataset consisting only of morphologically similar lineages that are putative cryptic species.

Abbreviations

BFD: Bayes factor delimitation; COI: Cytochrome c oxidase subunit I; ddRAD: Double-digest RADSeq; GSMNP: Great Smoky Mountains National Park; MSC:

Multispecies coalescent; nTA: Northern Trans-Asheville; sTA: Southern Trans-Asheville; TNV: Tennessee Valley; UCE: Ultraconserved elements.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12983-022-00453-0.

Additional file 1: Table S1. Taxon sample information including voucher number, locality information, and morphological and genetic sampling. **Table S2.** Statistics of ddRAD data. **Table S3.** Data matrix statistics. **Table S4.** Models and justification for all species hypotheses tested in *BFD analyses. **Table S5.** Results of *BFD analyses.

Additional file 2: Figures S1–S12. Supplementary figures. See file for further details.

Acknowledgements
We would like to thank Derek Hennen, Matthew Niemiller, Bill Shear, and Kirk Zigler for providing important specimens. For assistance with fieldwork we thank Jason Bond, Fred Coyle, Ryan Faucett, Dalton Hedin, Lars Hedin, Robin Keith, Dan Proud, and Steven Thomas. Erik Ciaccio and Morganne Sigismonti helped with specimen imaging. Ingrid R. Niesman provided SEM support at SDSU. We thank Lorenzo Prendini and Pio Colmenares at the American Museum of Natural History for granting access to the holotype of *Theromaster archeri*. We thank Dr. Diethard Tautz and two anonymous reviewers for helpful comments during review.

Authors' contributions
SD, JS, MH conceived the experiments. SD and JS collected the data. SD, JS, and MH analyzed the data. SD designed and conducted machine learning analyses. All authors contributed to the writing of this manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the National Science Foundation (DEB 1354558 to M.H.). They had no role in the design of the study, collection, analysis, and interpretation of the data, or writing the manuscript.

Availability of data and materials
The data supporting the conclusions of this article is available in the NCBI Short-Read Archive BioProject PRJNA804220 (raw ddRAD reads) and BioProject PRJNA804099 (raw UCE reads), with matrices and tree files available from the Dryad Digital Repository: 10.5061/dryad.79cnp5htb.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, 26 Oxford St., Cambridge, MA 02138, USA. [2]Department of Entomology and Nematology, University of California, Davis, Briggs Hall, Davis, CA 95616-5270, USA. [3]Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-4614, USA.

Received: 19 November 2021 Accepted: 27 January 2022
Published online: 22 February 2022

References

1. da Silva Ribeiro T, Batalha-Filho H, Silveira LF, Miyaki CY, Maldonado-Coelho M. Life history and ecology might explain incongruent population structure in two co-distributed montane bird species of the Atlantic Forest. Mol Phylogenet Evol. 2020;153:106925.
2. Fang F, Chen J, Jiang LY, Chen R, Qiao GX. Biological traits yield divergent phylogeographical patterns between two aphids living on the same host plants. J Biogeogr. 2017;44(2):348–60.
3. Fenker J, Tedeschi LG, Melville J, Moritz C. Predictors of phylogeographic structure among co-distributed taxa across the complex Australian monsoonal tropics. Mol Ecol. 2020. https://doi.org/10.1111/mec.16057.
4. Massatti R, Knowles LL. Microhabitat differences impact phylogeographic concordance of codistributed species: Genomic evidence in montane sedges (*Carex* L.) from the Rocky Mountains. Evolution. 2014;68(10):2833–46.
5. Massatti R, Knowles LL. Contrasting support for alternative models of genomic variation based on microhabitat preference: species-specific effects of climate change in alpine sedges. Mol Ecol. 2016;25(16):3974–86.
6. Gittenberger E. What about non-adaptive radiation? Biol J Linn Soc. 1991;43(4):263–72.
7. Rundell RJ, Price TD. Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation. Trends Ecol Evol. 2009;24(7):394–9.
8. Czekanski-Moir JE, Rundell RJ. The ecology of nonecological speciation and nonadaptive radiations. Trends Ecol Evol. 2019;34(5):400–15.
9. Barley AJ, White J, Diesmos AC, Brown RM. The challenge of species delimitation at the extremes: diversification without morphological change in Philippine sun skinks. Evolution. 2013;67(12):3556–72.
10. Derkarabetian S, Castillo S, Koo PK, Ovchinnikov S, Hedin M. A demonstration of unsupervised machine learning in species delimitation. Mol Phylogenet Evol. 2019;139:106562.
11. Hedin M, Carlson D, Coyle F. Sky island diversification meets the multispecies coalescent–divergence in the spruce-fir moss spider (*Microhexura montivaga*, Araneae, Mygalomorphae) on the highest peaks of southern Appalachia. Mol Ecol. 2015;24(13):3467–84.
12. Niemiller ML, Near TJ, Fitzpatrick BM. Delimiting species using multilocus data: diagnosing cryptic diversity in the southern cavefish, *Typhlichthys subterraneus* (Teleostei: Amblyopsidae). Evolution. 2012;66(3):846–66.
13. Satler JD, Carstens BC, Hedin M. Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, *Aliatypus*). Syst Biol. 2013;62(6):805–23.
14. Yang L, Kong H, Huang JP, Kang M. Different species or genetically divergent populations? Integrative species delimitation of the *Primulina hochiensis* complex from isolated karst habitats. Mol Phylogenet Evol. 2019;132:219–31.
15. Sukumaran J, Knowles LL. Multispecies coalescent delimits structure, not species. Proc Natl Acad Sci U S A. 2017;114(7):1607–12.
16. Leaché AD, Zhu T, Rannala B, Yang Z. The spectre of too many species. Syst Biol. 2019;68(1):168–81.
17. Barley AJ, Brown JM, Thomson RC. Impact of model violations on the inference of species boundaries under the multispecies coalescent. Syst Biol. 2018;67(2):269–84.
18. Dayrat B. Towards integrative taxonomy. Biol J Linn Soc. 2005;85(3):407–17.
19. Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH. Integrative taxonomy: a multisource approach to exploring biodiversity. Annu Rev Entomol. 2010;55:421–38.
20. Jacobs SJ, Kristofferson C, Uribe-Convers S, Latvis M, Tank DC. Incongruence in molecular species delimitation schemes: what to do when adding more data is difficult. Mol Ecol. 2018;27(10):2397–413.
21. Wiens JJ, Graham CH. Niche conservatism: integrating evolution, ecology, and conservation biology. Annu Rev Ecol Evol Syst. 2005;36:519–39.
22. Derkarabetian S, Ledford J, Hedin M. Genetic diversification without obvious genitalic morphological divergence in harvestmen (Opiliones, Laniatores, *Sclerobunus robustus*) from montane sky islands of western North America. Mol Phylogenet Evol. 2011;61(3):844–53.
23. Derkarabetian S, Hedin M. Integrative taxonomy and species delimitation in harvestmen: a revision of the western North American genus *Sclerobunus* (Opiliones: Laniatores: Travunioidea). PLoS ONE. 2014;9(8):e104982.

24. Derkarabetian S, Burns M, Starrett J, Hedin M. Population genomic evidence for multiple Pliocene refugia in a montane-restricted harvestman (Arachnida, Opiliones, *Sclerobunus robustus*) from the southwestern United States. Mol Ecol. 2016;25(18):4611–31.
25. DiDomenico A, Hedin M. New species in the *Sitalcina sura* species group (Opiliones, Laniatores, Phalangodidae), with evidence for a biogeographic link between California desert canyons and Arizona sky islands. ZooKeys. 2016;586:1–36.
26. Hedin M, Thomas SM. Molecular systematics of eastern North American Phalangodidae (Arachnida: Opiliones: Laniatores), demonstrating convergent morphological evolution in caves. Mol Phylogenet Evol. 2010;54(1):107–21.
27. Peres EA, DaSilva MB, Antunes M Jr, Pinto-Da-Rocha R. A short-range endemic species from south-eastern Atlantic Rain Forest shows deep signature of historical events: phylogeography of harvestmen *Acutisoma longipes* (Arachnida: Opiliones). Syst Biodivers. 2018;16(2):171–87.
28. Starrett J, Derkarabetian S, Richart CH, Cabrero A, Hedin M. A new monster from southwest Oregon forests: *Cryptomaster behemoth* sp. n. (Opiliones, Laniatores, Travunioidea). ZooKeys. 2016;555:11.
29. Thomas SM, Hedin M. Multigenic phylogegraphic divergence in the paleoendemic southern Appalachian opilionid *Fumontana deprehendor* Shear (Opiliones, Laniatores, Triaenonychidae). Mol Phylogenet Evol. 2008;46(2):645–58.
30. Boyer SL, Baker JM, Giribet G. Deep genetic divergences in *Aoraki denticulata* (Arachnida, Opiliones, Cyphophthalmi): a widespread 'mite harvestman' defies DNA taxonomy. Mol Ecol. 2007;16(23):4999–5016.
31. Fernández R, Giribet G. Phylogeography and species delimitation in the New Zealand endemic, genetically hypervariable harvestman species, *Aoraki denticulata* (Arachnida, Opiliones, Cyphophthalmi). Invertebr Syst. 2014;28(4):401–14.
32. Chambers EA, Hillis DM. The multispecies coalescent over-splits species in the case of geographically widespread taxa. Syst Biol. 2020;69(1):184–93.
33. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE. 2012;7(5):e37135.
34. Starrett J, Derkarabetian S, Hedin M, Bryson RW Jr, McCormack JE, Faircloth BC. High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. Mol Ecol Resour. 2017;17(4):812–23.
35. Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. Syst Biol. 2012;61(5):717–26.
36. Banks N. A new phalangid from the Black Mountains, NC. J N Y Entomol S. 1902;10(3):142–142.
37. Briggs TS. A new holarctic family of laniatorid phalangids (Opiliones). Pan-Pac Entomol. 1969;45(1):35–50.
38. Goodnight CJ, Goodnight ML. New Phalangodidae (Phalangida) from the United States. Am Mus Novit. 1942;1188:1–18.
39. Kury AB. Annotated catalogue of the Laniatores of the New World: (Arachnida, Opiliones). Rev Ibér Aracnol. 2003;7:5–337.
40. Derkarabetian S, Benavides LR, Giribet G. Sequence capture phylogenomics of historical ethanol-preserved museum specimens: unlocking the rest of the vault. Mol Ecol Resour. 2019;19:1531–44.
41. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005;14(8):2611–20.
42. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.
43. Hedin M, McCormack M. Biogeographical evidence for common vicariance and rare dispersal in a southern Appalachian harvestman (Sabaconidae, *Sabacon cavicolens*). J Biogeogr. 2017;44(7):1665–78.
44. Huang JP. What have been and what can be delimited as species using molecular data under the multi-species coalescent model? A case study using Hercules beetles (*Dynastes*; Dynastidae). Insect Syst Divers. 2018;2(2):3.
45. Mason NA, Fletcher NK, Gill BA, Funk WC, Zamudio KR. Coalescent-based species delimitation is sensitive to geographic sampling and isolation by distance. Syst Biodivers. 2020;18(3):269–80.
46. Sukumaran J, Holder MT, Knowles LL. Incorporating the speciation process into species delimitation. PLoS Comput Biol. 2021;17(5):e1008924.
47. Yang Z, Zhu T. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. Proc Natl Acad Sci U S A. 2018;115(8):1854–9.
48. Carnaval AC, Waltari E, Rodrigues MT, Rosauer D, VanDerWal J, Damasceno R, Prates I, Strangas M, Spanos Z, Rivera D, et al. Prediction of phylogeographic endemism in an environmentally complex biome. Proc R Soc B Biol Sci. 2014;281(1792):20141461.
49. Espíndola A, Ruffley M, Smith L, Carstens BC, Tank DC, Sullivan J. Identifying cryptic diversity with predictive phylogeography. Proc R Soc B Biol Sci. 2016;283(1841):20161529.
50. Pei J, Chu C, Li X, Lu B, Wu Y. CLADES: a classification-based machine learning method for species delimitation from population genetic data. Mol Ecol Resour. 2018;18(5):1144–56.
51. Smith ML, Carstens BC. Process-based species delimitation leads to identification of more biologically relevant species. Evolution. 2020;74(2):216–29.
52. Leaché AD, Davis HR, Singhal S, Fujita MK, Zamudio KR. Phylogenomic assessment of biodiversity using a reference-based taxonomy: an example with Horned Lizards (*Phrynosoma*). Frontiers Ecol Evol. 2021;9:437.
53. Campillo LC, Barley AJ, Thomson RC. Model-based species delimitation: are coalescent species reproductively isolated? Syst Biol. 2020;69(4):708–21.
54. Crespi EJ, Rissler LJ, Browne RA. Testing Pleistocene refugia theory: phylogeographical analysis of *Desmognathus wrighti*, a high-elevation salamander in the southern Appalachians. Mol Ecol. 2003;12(4):969–84.
55. Hedin M. Molecular phylogenetics at the population/species interface in cave spiders of the southern Appalachians (Araneae:Nesticidae:Nesticus). Molecular Biology and Evolution. 1997;14(3):309–24.
56. Kozak KH, Wiens JJ. Niche conservatism drives elevational diversity patterns in Appalachian salamanders. Am Nat. 2010;176(1):40–54.
57. Marek PE. A revision of the Appalachian millipede genus *Brachoria* Chamberlin 1939 (Polydesmida: Xystodesmidae: Apheloriini). Zoological Journal of the Linnean Society. 2010;159(4):817–89.
58. Weisrock DW, Larson A. Testing hypotheses of speciation in the *Plethodon jordani* species complex with allozymes and mitochondrial DNA sequences. Biol J Linn Soc. 2006;89(1):25–51.
59. Caterino MS, Langton-Myers SS. Intraspecific diversity and phylogeography in Southern Appalachian *Dasycerus carolinensis* Horn. Insect Syst Divers. 2019;3(6):1–12.
60. Garrick RC, Newton KE, Worthington RJ. Cryptic diversity in the southern Appalachian Mountains: genetic data reveal that the red centipede, *Scolopocryptops sexspinosus*, is a species complex. J Insect Conserv. 2018;22(5–6):799–805.
61. Keith R, Hedin M. Extreme mitochondrial population subdivision in southern Appalachian paleoendemic spiders (Araneae: Hypochilidae: *Hypochilus*), with implications for species delimitation. J Arachnol. 2012;40(2):167–81.
62. Newton LG, Starrett J, Hendrixson BE, Derkarabetian S, Bond JE. Integrative species delimitation reveals cryptic diversity in the southern Appalachian *Antrodiaetus unicolor* (Araneae: Antrodiaetidae) species complex. Mol Ecol. 2020;29(12):2269–87.
63. Goodnight CJ, Goodnight ML. Speciation among cave opilionids of the United States. Am Midl Nat. 1960;64(1):34–8.
64. Derkarabetian S, Starrett J, Tsurusaki N, Ubick D, Castillo S, Hedin M. A stable phylogenomic classification of Travunioidea (Arachnida, Opiliones, Laniatores) based on sequence capture of ultraconserved elements. ZooKeys. 2018;760:1–36.
65. Derkarabetian S, Steinmann DB, Hedin M. Repeated and time-correlated morphological convergence in cave-dwelling harvestmen (Opiliones, Laniatores) from montane western North America. PLoS ONE. 2010;5(5):e10388.
66. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
67. Zhang J, Kapli P, Pavlidis P, Stamatakis A. A general species delimitation method with applications to phylogenetic placements. Bioinformatics. 2013;29(22):2869–76.
68. Burns M, Starrett J, Derkarabetian S, Richart CH, Cabrero A, Hedin M. Comparative performance of double-digest RAD sequencing across divergent arachnid lineages. Mol Ecol Resour. 2017;17(3):418–30.
69. Eaton DA, Overcast I. ipyrad: interactive assembly and analysis of RADseq datasets. Bioinformatics. 2020;36(8):2592–4.

70. Bryson RW Jr, Savary WE, Zellmer AJ, Bury RB, McCormack JE. Genomic data reveal ancient microendemism in forest scorpions across the California Floristic Province. Mol Ecol. 2016;25(15):3731–51.
71. Chifman J, Kubatko L. Quartet inference from SNP data under the coalescent model. Bioinformatics. 2014;30(23):3317–24.
72. Swofford DL. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland: Sinauer Associates; 2003.
73. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24(11):1403–5.
74. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011;27(21):3070–1.
75. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour. 2015;15(5):1179–91.
76. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010;11(1):94.
77. Leaché AD, Banbury BL, Felsenstein J, De Oca ANM, Stamatakis A. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. Syst Biol. 2015;64(6):1032–47.
78. Zarza E, Connors EM, Maley JM, Tsai WL, Heimes P, Kaplan M, McCormack JE. Combining ultraconserved elements and mtDNA data to uncover lineage diversity in a Mexican highland frog (*Sarcohyla*; Hylidae). PeerJ. 2018;6:e6045.
79. Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst Biol. 2014;63(1):83–95.
80. Blaimer BB, LaPolla JS, Branstetter MG, Lloyd MW, Brady SG. Phylogenomics, biogeography and diversification of obligate mealybug-tending ants in the genus *Acropyga*. Mol Phylogenet Evol. 2016;102:20–9.
81. Hedin M, Derkarabetian S, Blair J, Paquin P. Sequence capture phylogenomics of eyeless *Cicurina* spiders from Texas caves, with emphasis on US federally-endangered species from Bexar County (Araneae, Hahniidae). ZooKeys. 2018;769:49.
82. Faircloth BC. Identifying conserved genomic elements and designing universal bait sets to enrich them. Methods Ecol Evol. 2017;8(9):1103–12.
83. Hedin M, Derkarabetian S, Alfaro A, Ramírez MJ, Bond J. Phylogenomic analysis and revised classification of atypoid mygalomorph spiders (Araneae, Mygalomorphae), with notes on arachnid ultraconserved element loci. PeerJ. 2019;7:
84. Faircloth BC. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics. 2016;32(5):786–8.
85. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
86. Faircloth BC. Illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming. 2013. https://doi.org/10.6079/J9ILL.
87. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9.
88. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
89. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17(4):540–52.
90. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56(4):564–77.
91. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genom. 2016;2(4):e000056.
92. Borowiec ML. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ. 2016;4:e1660.
93. Grummer JA, Bryson RW Jr, Reeder TW. Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae). Syst Biol. 2014;63(2):119–33.
94. Leaché AD, Fujita MK, Minin VN, Bouckaert RR. Species delimitation using genome-wide SNP data. Syst Biol. 2014;63(4):534–42.
95. Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol Biol Evol. 2012;29(8):1917–32.
96. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014;10(4):e1003537.
97. Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc. 1995;90(430):773–95.
98. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Publisher's Note