UNIVERSITY OF CALIFORNIA

Los Angeles

The causes and consequences of venom evolution in cone snails (Family, Conidae)

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Biology

by

Mark Anthony Phuong

2018

ABSTRACT OF THE DISSERTATION

The causes and consequences of venom evolution

in cone snails (Family, Conidae)

by

Mark Anthony Phuong

Doctor of Philosophy in Biology

University of California, Los Angeles, 2018

Professor Michael Edward Alfaro, Chair

The ability to produce venom has evolved multiple times throughout the animal kingdom and consist of complex mixtures of toxic proteins. The causes and consequences of venom evolution frequently remain unclear, potentially due to the difficulty of obtaining comprehensive data on venom composition across multiple species. Here, I present an in-depth analysis of the factors governing venom evolution, and the impact of venom evolution on speciation rates. Chapter 1 focuses on the impact of diet on venom evolution. Using RNAseq data to characterize the venom cocktails of 12 Conidae species, I recovered a positive correlation between venom complexity and dietary breadth, suggesting that species with more generalist diets express a greater number of venom proteins. Chapter 2 investigates the impact of genetics on venom evolution. Using a novel targeted sequencing approach, I describe how positive selection, gene duplication, and expression regulation all contribute to the diversification of venom. Finally, in Chapter 3, I

investigate the impact of venom evolution on speciation rates. Using a targeted sequencing approach, I characterize the venom repertoire of over 300 cone snail species and find no relationship between venom evolution and speciation rates. My results suggest that venom is not the rate-limiting factor in determining speciation events and that other factors, such as ecological opportunity or traits regulating dispersal are more critical in governing diversification patterns.

The dissertation of Mark Anthony Phuong is approved.

Thomas Duda

David K. Jacobs

Michael Edward Alfaro, Committee Chair

University of California, Los Angeles

2018

TABLE OF CONTENTS

ACKNOWLEDGMENTS

Chapter 1 is a reproduction of the following published paper: Phuong, MA, GN Mahardika, ME Alfaro. (2016) Dietary breadth is positively correlated with venom complexity in cone snails. BMC Genomics 17:401. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2755-6. For this work, MAP designed the study, conducted the field and laboratory work, carried out the bioinformatic analyses, and drafted the manuscript. GNM participated in field work. All authors read, reviewed, and approved the final manuscript. This paper is under a Creative Commons Attribution license, which means that it can be re-used in any way, subject to proper attribution.

Chapter 2 is a reproduction of the following published paper: Phuong, MA, GN Mahardika. (2018) Targeted sequencing of venom genes from cone snail genomes improves understanding of conotoxin molecular evolution. Molecular Biology and Evolution 35: 1210–1224. https://doi.org/10.1093/molbev/msy034. This work was published by Oxford University Press.nFor this work, MAP designed the study, conducted the field and laboratory work, carried out the bioinformatic analyses, and drafted the manuscript. GNM participated in field work. All authors read, reviewed, and approved the final manuscript. Under the licensing agreement, authors may use their own material in other publications, provided that Molecular Biology and Evolution is acknowledged as the original place of publication and Oxford University Press is acknowledged as the publisher.

Chapter 3 is based on the following manuscript that is in preparation for publication: Phuong, MA, MA Alfaro, GN Mahardika, RM Marwoto, RE Prabowo, T von Rintelen, PWH Vogt, JR Hendricks, N Puillandre. Speciation rates are decoupled from conotoxin gene diversity in cone snails. *In prep*. For this work, MAP designed the study, conducted the field and laboratory work,

carried out the bioinformatic analyses, and drafted the manuscript. GNM, RMM, REP, TvR, and PWHV participated in field work, JRH provided the fossil calibrations, and NP provided over a majority of the specimens. All authors read, reviewed, and approved the manuscript.

This dissertation was supported by the following grants and fellowships:

1. Chateaubriand Fellowship, Embassy of France in the United States
2. Fulbright U.S. Student Program to Indonesia, Fulbright
3. Graduate Research Fellowship, National Science Foundation
4. Edwin W. Pauley Fellowship, University of California, Los Angeles
5. Five research awards, University of California, Los Angeles, Department of Ecology and Evolutionary Biology
6. Student Research Award, Unitas Malacologica
7. Academic Grant, Conchologists of America
8. Melbourne R. Carriker Student Research Award, American Malacological Society
9. Graduate Research Opportunities Worldwide to France, National Science Foundation
10. Student Research Award, American Society of Naturalists
11. Doctoral Dissertation Improvement Grant, National Science Foundation
12. Grants-in-Aid of Research, Society for Integrative and Comparative Biology
13. Two Grants-in-Aid of Research, Sigma Xi
14. Research award, Society of Systematic Biologists
15. Graduate Research Opportunities Worldwide to Australia, National Science Foundation
16. Lemelson Fellowship, University of California, Los Angeles, Indonesian Studies program
17. Lewis and Clark Fund, American Philosophical Society

18. Young Explorer's Grant, National Geographic Society
19. Travel Award, University of California, Los Angeles Graduate Division
20. Research Grant, American Institute for Indonesian Studies, Council of American Overseas Research Centers
21. Lerner Gray Fund for Marine Research, American Museum of Natural History
22. London Malacological Society Research Grant, London Malacological Society

VITA/BIOGRAPHICAL SKETCH

(a) Professional preparation

University of California, Berkeley	Integrative Biology	B.A., 2012

(b) Appointments

Graduate student, University of California, Los Angeles 2012-present

Lab Technician, University of California, Berkeley May 2012-Sept 2012

(c) Publications

1. Phuong, MA, GN Mahardika. (2018) Targeted sequencing of venom genes from cone snail genomes improves understanding of conotoxin molecular evolution. Molecular Biology and Evolution 35: 1210–1224. https://doi.org/10.1093/molbev/msy034
2. Phuong, MA, K Bi, C Moritz. (2017) Range instability leads to cytonuclear discordance in a morphologically cryptic ground squirrel species complex. Molecular Ecology 26: 4743–4755. https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.14238
3. Phuong, MA, GN Mahardika, ME Alfaro. (2016) Dietary breadth is positively correlated with venom complexity in cone snails. BMC Genomics 17:401. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2755-6
4. Conroy, CJ, JL Patton, MCW Lim, MA Phuong, B Parmenter, S Höhna. (2016) Following the rivers: historical reconstruction of California voles Microtus californicus (Rodentia: Cricetidae) in the deserts of eastern California. Biological Journal of the Linnean Society 119:80-98. http://onlinelibrary.wiley.com/doi/10.1111/bij.12808/full
5. Phuong, MA, MCW Lim, DR Wait, KC Rowe, C Moritz. (2014) Delimiting species in the genus, *Otospermophilus* (Rodentia: Sciuridae) using genetics, ecology, and

morphology. Biological Journal of the Linnean Society 113:1136-1151. http://onlinelibrary.wiley.com/doi/10.1111/bij.12391/abstract

6. Romero, M., MA Phuong, C Bishop, P Krug (2013) Nitric oxide signaling differentially affects habitat choice by two larval morphs of the sea slug *Alderia willowi*: mechanistic insight into evolutionary transitions in dispersal strategies. Journal of Experimental Biology 216: 1114–1125. http://jeb.biologists.org/content/216/6/1114

Chapter 1: Dietary breadth is positively correlated with venom complexity in cone snails

Abstract

Background

Although diet is believed to be a major factor underlying the evolution of venom, few comparative studies examine both venom composition and diet across a radiation of venomous species. Cone snails within the family, Conidae, comprise more than 700 species of carnivorous marine snails that capture their prey by using a cocktail of venomous neurotoxins (conotoxins or conopeptides). Venom composition across species has been previously hypothesized to be shaped by (a) prey taxonomic class (i.e., worms, molluscs, or fish) and (b) dietary breadth. We tested these hypotheses under a comparative phylogenetic framework using ecological data from past studies in conjunction with venom duct transcriptomes sequenced from 12 phylogenetically disparate cone snail species, including 10 vermivores (worm-eating), one molluscivore, and one generalist.

Results

We discovered 2223 unique conotoxin precursor peptides that encoded 1864 unique mature toxins across all species, >90% of which are new to this study. In addition, we identified two novel gene superfamilies and 16 novel cysteine frameworks. Each species exhibited unique venom profiles, with venom composition and expression patterns among species dominated by a restricted set of gene superfamilies and mature toxins. In contrast with the dominant paradigm for interpreting Conidae venom evolution, prey taxonomic class did not predict venom composition patterns among species. We also found a significant positive relationship between dietary breadth and measures of conotoxin complexity.

Conclusions

The poor performance of prey taxonomic class in predicting venom components suggests that cone snails have either evolved species-specific expression patterns likely as a consequence of the rapid evolution of conotoxin genes, or that traditional means of categorizing prey type (i.e., worms, mollusc, or fish) and conotoxins (i.e., by gene superfamily) do not accurately encapsulate evolutionary dynamics between diet and venom composition. We also show that species with more generalized diets tend to have more complex venoms and utilize a greater number of venom genes for prey capture. Whether this increased gene diversity confers an increased capacity for evolutionary change remains to be tested. Overall, our results corroborate the key role of diet in influencing patterns of venom evolution in cone snails and other venomous radiations.

Background

The use of venom for predation has evolved several times across the animal kingdom in organisms such as snakes, snails, and spiders (Fry *et al.* 2009). The majority of venoms consist of complex mixtures of toxic proteins (Rash & Hodgson 2002; Olivera 2002; Fry *et al.* 2009) and extensive variation in venom composition is documented at nearly all biological scales of study ranging from individuals to species (Gutiérrez *et al.* 1990; Daltry *et al.* 1996; da Silva Jr. & Aird 2001; Duda & Remigio 2008). Understanding the forces that shape venom evolution and variation in venom composition among predatory venomous taxa is not only of intrinsic interest to ecological and evolutionary studies (Casewell *et al.* 2013), but has far reaching implications across several biological disciplines, including drug development in pharmacology and understanding protein structure-function relationships in molecular biology (Olivera & Teichert 2007; Vonk *et al.* 2011). Diet is thought to be a major driver of venom composition patterns because venom is intricately linked to a species' ability to capture and apprehend prey (Vonk *et*

al. 2011; Casewell *et al.* 2013). There are currently two major hypotheses that attempt to explain the impact of diet on broad-scale patterns of venom composition across taxa: (1) prey preference should determine venom components and (2), dietary breadth should be positively correlated with venom complexity (Daltry *et al.* 1996; Li *et al.* 2005a; Pahari *et al.* 2007). While both hypotheses are often used interchangeably as evidence for the role of diet in venom evolution (e.g., (Binford 2001; Pahari *et al.* 2007; Brahma *et al.* 2015)), they have separate and distinct predictions on patterns of venom composition among taxa: whereas the former hypothesis predicts the types of venom proteins expected for a given species, the latter hypothesis predicts how many proteins are employed for prey capture.

The idea that prey preference should determine the types of venom proteins employed by a given species is grounded in the logic that natural selection shapes the venom repertoires of species to become more effective at targeting the physiologies of their prey (Daltry *et al.* 1996; Vonk *et al.* 2011). Several studies support this relationship, including correlations between variation in diet and venom components among populations within species (Daltry *et al.* 1996; Creer *et al.* 2003) and functional studies which show that the toxic effects of venoms from different species were maximally effective on their preferred prey (Endean & Rudkin 1963, 1965; da Silva Jr. & Aird 2001; Richards *et al.* 2012). For example, snake venoms from species that preferentially feed on arthropods were more toxic upon injection into scorpions relative to venom extracted from a species that feeds almost exclusively on vertebrates (Barlow *et al.* 2009). However, there are cases where variation in venom composition cannot be attributed to dietary preferences, challenging the generality of this pattern (Williams *et al.* 1988; Gibbs *et al.* 2013). Indeed, gene duplication, positive selection, and protein neofunctionalization are defining features of venom gene evolution (Duda & Palumbi 1999; Fry *et al.* 2003a; Casewell *et al.* 2011;

Chang & Duda 2012) and these forces work in concert to promote divergence in venom composition among taxa. Given the high evolutionary lability of venom toxins, it is unclear that a relationship between dietary preference and venom composition should be expected.

The second hypothesis on dietary breadth and venom complexity seeks to explain why some species employ more venom proteins than others for prey capture (Fry *et al.* 2003b). Under this hypothesis, dietary breadth should be positively correlated with venom complexity because a greater number of venom proteins is necessary to target a wide variety of prey species (Li *et al.* 2005a; Pahari *et al.* 2007). Although rarely invoked in venom studies, this relationship is explicitly predicted by the niche variation hypothesis, which posits that individuals or populations with wider niches should display greater phenotypic variance (Van Valen 1965). To date, nearly all evidence supporting the impact of dietary breadth in shaping patterns of venom complexity are essentially observational. For example, sea snakes, which mostly feed on fish, have less diverse venoms compared to land snakes, which typically feed on arthropods, reptiles, amphibians, birds, and mammals (Pahari *et al.* 2007). In addition, prey specialists tend to have less complex venoms compared to generalists (Pahari *et al.* 2007; Remigio & Duda 2008; Elliger *et al.* 2011). Despite the apparent signal, these observations have yet to be tested in a phylogenetically controlled and rigorous manner.

Although diet is widely accepted as the dominant force governing venom evolution across disparate venomous taxa (Casewell *et al.* 2013), few multi-species comparative studies exist that explicitly examine the impact of diet on venom composition patterns across venomous radiations. The majority of studies implicating the prominent role of diet in venom evolution are based on variation in venom composition among populations within species or among closely related species (Daltry *et al.* 1996; Binford 2001; Barlow *et al.* 2009; Gibbs *et al.* 2013; Chang *et*

al. 2015). In some cases, broad generalizations on the evolutionary trends of venom and diet are made from the analyses of a few individuals from a single species (e.g, (Li *et al.* 2005a; Jin *et al.* 2013)). In addition, knowledge on venom composition is often incomplete – most studies are restricted to commonly known gene families (Li *et al.* 2005b; Kaas *et al.* 2010), challenging the generality of previous results given that a substantial proportion of venomous cocktails potentially go unexamined. Without employing a broad and robust comparative phylogenetic approach in conjunction with comprehensive venom data, it is not possible to determine whether previously reported patterns represent general evolutionary trends in venomous taxa or are idiosyncratic phenomena restricted to the particularities of a given study.

Here, we examine the influence of both dietary preference and dietary breadth on venom evolution in cone snails (Family: Conidae), a hyper diverse group of over 700 predatory marine snails that typically prey on either worms, molluscs, or fish using a cocktail of venomous neuropeptides (known as conotoxins or conopeptides) (Robinson & Norton 2014; Puillandre *et al.* 2014a). Each species' venom repertoire is estimated to contain 50-200 peptides and these peptides can be classified into more than 30 gene superfamilies (e.g, A superfamily, M superfamily, etc.) based on the similarity of the signal region (i.e., a conserved region at the beginning of precursor conotoxins containing ~20 hydrophobic amino acids that directs the peptide into the secretory pathway) (McIntosh *et al.* 1999; Robinson & Norton 2014). To examine the relationship between venom composition and diet in cone snails, we sequenced the mRNA from the venom duct of 12 phylogenetically disparate cone snail species consisting of 10 vermivores (worm-eaters), one molluscivore, and one generalist that feeds on worms, molluscs, and fish (Duda Jr. *et al.* 2001; Puillandre *et al.* 2014a). We analyze ecological data and venom composition patterns under a comparative phylogenetic framework to test two previously

proposed hypotheses that attempt to explain the impact of diet on cone snail venom evolution: (1) traditional prey taxonomic categories (i.e., worms, molluscs, fish) should predict which gene superfamilies are expressed and (2) dietary breadth should be positively correlated with conotoxin complexity (Remigio & Duda 2008; Kaas *et al.* 2010; Elliger *et al.* 2011; Barghi *et al.* 2014).

Results

Transcriptome sequencing and assembly

We extracted RNA from the venom duct of 12 species (1 individual per species): *Californiconus californicus, Conus arenatus, Conus coronatus, Conus ebraeus, Conus imperialis, Conus lividus, Conus marmoreus, Conus quercinus, Conus rattus, Conus sponsalis, Conus varius*, and *Conus virgo* (Table 1-1). Here, we note that *C. sponsalis* refers one lineage of the *C. sponsalis* species complex, where a number of described species are comprised of several, paraphyletic lineages (Duda *et al.* 2008). We use the name *C. sponsalis* to refer to this species complex pending taxonomic revision of this group. We synthesized RNAseq libraries and multiplexed all individuals on a single Illumina HiSeq 2000 lane. We recovered an average of 25.8 million reads per species (Table S1-1) and assembled transcripts using Trinity (Grabherr *et al.* 2011). The number of contigs assembled ranged from 28,878 to 88,052, n50 was 609.25 on average, and the total bases assembled ranged from 15MB to 50MB (Table S1-1).

Conopeptide identification, classification, and diversity

We used a combination of custom Python scripts, BLAST+, ConoSorter (an algorithm used to identify transcripts that code proteins which share similar properties to known

conotoxins), and ConoPrec (a tool used to analyze conopeptide precursors) to identify, filter, and classify conopeptides (Altschup *et al.* 1990; Kaas *et al.* 2012; Lavergne *et al.* 2013). Conotoxins are typically classified into gene superfamilies and the majority of gene superfamily names contain a single letter followed by an Arabic numeral or are named based on their similarity to proteins from other venomous taxa (i.e., conkunitzins and conopressins) (Robinson & Norton 2014). Conotoxins that could not be classified into these categories were generally given a new name with nomenclatural conventions highly dependent on the study organism and the research group (Elliger *et al.* 2011; Lavergne *et al.* 2013; Jin *et al.* 2013). For example, conotoxins from *C. californicus* were given the name "Divergent" to reflect its divergent phylogenetic position relative to the rest of Conidae (Barghi *et al.* 2014). Through the investigation of conotoxin gene superfamily classifications, we noted several cases where changes in the current naming and classification of gene superfamilies were warranted. Based on signal sequence similarity or protein domain similarity, we reclassified the Divergent_MTFLLLLVSV superfamily as conkunitzins and reclassified the Divergent_MSTLGMTLL superfamily as the N superfamily. We observed that the Divergent_M---L-LTVA superfamily contained several conopeptide precursors with unique and divergent signal sequences. We dissolved this gene superfamily and reclassified it along with all other conopeptides that we were not able to assign into known gene superfamilies.

We used a percent signal sequence identify cut off of 70% to cluster unassigned conopeptides. We assign new names to (1) novel groupings of conotoxin gene superfamilies and (2) groups of conopeptides with similarity to previously characterized conotoxins, but were not given a formal classification. In total, we identified 2223 unique conopeptide precursor sequences that ultimately become cleaved and processed into 1864 unique mature proteins, 1685

of which are new to this study (Table 1-1, Table S1-2). A substantial proportion of these conopeptides were never assembled by Trinity (7.2% - 31% per species, Table S1-3), but discovered through read mapping and manual reconstruction. These conopeptides span 58 gene superfamilies, nine of which represent gene superfamilies given new names due to reclassification and two of which are newly described (Table S1-2, Table S1-4, Figure S1-1). Several of these gene superfamilies were recently characterized (e.g. G-like superfamily, SF-mi1 superfamily, etc.), but not given conventional gene superfamily names (i.e. a letter sometimes followed by an Arabic numeral). Although we expanded the membership of these gene superfamilies, we refrained from changing their names pending functional experiments to determine their role in prey apprehension or defense.

Conotoxins are often characterized by their cysteine framework, or the arrangement of cysteine residues often present in mature peptides, which can sometimes provide information on peptide structure and function (Kaas *et al.* 2010). We identified 70 unique cysteine frameworks across all the conotoxins identified from this study, 16 of which display a novel cysteine motif not yet described from conotoxins (Table S1-5). We report 34 new associations between previously identified cysteine frameworks and gene superfamilies (Table S1-5). Of particular note, we identified cysteine-free conotoxins from the A and O2 superfamilies, which is in contrast to the cysteine-containing toxins previously described from these groups (Table S1-5, (Robinson & Norton 2014)).

While comparing conotoxins described in this study with the ConoServer database, we identified several discrepancies concerning the species of origin for particular conopeptides. For example, although we did not detect the Qc23a precursor peptide (originally described from *C. quercinus*) in our *C. quercinus* transcriptome, we found a precursor peptide with 100% identity

in our *C. imperialis* transcriptome. In another case, nearly every protein (i.e., 70/79 proteins) identified from a recent study on *C. flavidus* (Lu *et al.* 2014) had >95 % sequence identity to proteins identified from our *C. lividus* transcriptome. Many of these species mismatches occur between distantly related taxa, where high identity between full precursor peptides is not expected (Olivera *et al.* 1999). We hypothesized that several of these discrepancies are cases of species misidentification because we confirmed species identification in this study morphologically and genetically using mitochondrial DNA sequences from the transcriptome. We note these instances in the supplementary for further inquiry (Table S1-6).

Phylogeny inference

We employed an all-by-all blast approach using the *Lottia gigantea* protein database (GCA_000327385.1, (Simakov *et al.* 2013)) as our reference to identify 821 putatively orthologous loci suitable for phylogenetic analysis. These loci represent a total of 863,132bp and each species had, on average, 88.3% of the total bases possible in the data matrix. We inferred a maximum likelihood phylogeny in RAxML and generated a time tree using the program r8s with two fossil calibrations from previous studies (Figure 1-1, (Kohn 1990; Duda Jr. *et al.* 2001; Sanderson 2003)). The phylogeny was highly resolved with all but two nodes having 100% bootstrap support (Figure 1-1).

Interspecific conopeptide diversity patterns

Across all species examined in this study, the number of unique conotoxin precursors ranged from as low as 70 conopeptides in *C. imperialis* to as high as 401 conopeptides in *C. sponsalis* (Table 1-1). These precursors encoded between 66 to 338 unique mature toxins (also

from *C. imperialis* and *C. sponsalis*, respectively, Table 1-1). We detected only one instance where *C. coronatus* and *C. virgo* expressed the same mature toxin (Co_O2_13, Co_O2_14, and Vi_O2_7, Figure S1-2). In all other cases, each species expressed a unique repertoire of mature toxins with no overlap between species.

Species varied widely in which gene superfamilies were expressed (Table 1-1). On average, each species expressed 28 gene superfamilies, with *C. marmoreus* expressing the lowest number of superfamilies (14 superfamilies, Table 1-1) and *C. arenatus* expressing the highest number of superfamilies (36 superfamilies, Table 1-1). Only four gene superfamilies were expressed by all the species examined in this study (M, N, O1, and O2, Table S1-2). These four superfamilies were also the only superfamilies in common amongst the vermivores. The distribution of conopeptides across gene superfamilies per species tended to be skewed, such that > 50% of the conotoxins originated from three to six gene superfamilies (Table S1-7). The O1 superfamily contained the highest number of mature toxins for seven species, the M superfamily was the most abundant for three species, and the P and con-ikot-ikot superfamilies were each the most abundant for one species (Table 1-1). Interestingly, the O1, M and con-ikot-ikot superfamilies were also amongst the most abundant conotoxins from a recent study from the transcriptomes of *Conus tribblei* and *Conus lenavati* (Barghi *et al.* 2015a). The average number of cysteine frameworks found in each species was 24, with *C. arenatus* expressing the highest number of cysteine motifs (31 frameworks, Table 1-1) and *C. marmoreus* expressing the lowest number of cysteine motifs (16 frameworks, Table 1-1).

Interspecific conotoxin expression patterns

We used the RSEM algorithm to generate Transcript Per Million (TPM) values to compare venom duct expression levels between species (Li & Dewey 2011; Wagner *et al.* 2012). Total conotoxin expression, or the summed TPM values of conotoxin genes divided by total TPM of all transcripts, averaged 53% among species and ranged from as low as 26% in *C. californicus* to as high as 70.7% in *C. coronatus* (Table 1-2). The most highly expressed gene superfamily was the M superfamily for four species, the T superfamily for two species, the O1 superfamily for four species, and the Divergent_MRFYIGLMAA and L superfamilies each being the most abundant for one species (Table 1-2). On average, the most abundantly expressed gene superfamily represented 28.0% of total conotoxin transcripts (Table 1-2). In *C. ebraeus*, *C. marmoreus*, *C. sponsalis*, and *C. virgo*, the most abundantly expressed gene superfamily did not contain the most highly expressed mature conotoxin (Table 1-2). For example, while the most abundant gene superfamily was the M superfamily for *C. ebraeus*, the most highly expressed transcript was Eb_SF-mi2_2, a conotoxin from the SF-mi2 superfamily (Table 1-2). The average contribution of the highest expressed mature toxin from each species to overall conotoxin expression was 16.1% (Table 1-2). Conotoxin expression patterns tended to be dominated by a few gene superfamilies and mature conotoxins, such that 2-5 gene superfamilies and 2-23 mature toxins represented more than half of each species' conotoxin expression levels (Table 1-2, Table S1-8). Amongst the most highly expressed gene superfamilies (i.e., representing > 50% of conotoxin expression levels), we did not identify a single superfamily that was shared across all species (Table S1-8). We identified 14 gene superfamilies with expression levels contributing to at least 10% of overall conotoxin expression in at least one of the species examined in this study, whereas 33 superfamilies never constituted more than 5% of total conotoxin expression in any of the species (Figure 1-1, Table S1-9).

Pseudogene expression

We report a single instance where a premature stop codon interrupts the coding region of an O1 conotoxin expressed by *C. sponsalis* (Sp_O1_79, Figure S1-3). The stop codon appears within the signal region and the predicted mature conotoxin is identical to another conotoxin expressed by *C. sponsalis* (Sp_O1_87, Figure S1-3). The pseudogenized copy, Sp_O1_79, is more highly expressed than the functional copy, Sp_O1_87, by two orders of magnitude (TPM = 1242.59 and TPM = 11.66, respectively, Figure S1-3).

Diet and conotoxin composition

We employed the similarity statistic Schoener's D, a value commonly used to measure niche overlap in diet and/or microhabitat, to quantify the degree of overlap between conotoxin composition among cone snail species with different diets (Schoener 1968). D values can range from 0 (no overlap) to 1 (complete overlap) (Schoener 1968). To quantify venom composition similarity, we calculated the D statistic for (a) the percentage of mature toxins belonging to each gene superfamily (referred to as D_{mature}) and (b) the percent expression of gene superfamilies (referred to as $D_{expression}$) between all possible pairwise species comparisons. D_{mature} (avg = 0.48, range = 0.28 – 0.7) values on average, were higher than $D_{expression}$ (avg = 0.37, range = 0.09 to 0.68) values (Table S1-10). To control for phylogenetic signal, we generated residuals from a linear model between both values of D and pairwise phylogenetic distances from the time calibrated phylogeny. We used the residuals in an analysis of variance (ANOVA) to determine whether the distribution of conotoxin overlap values differed depending on whether or not the pairwise species comparison consisted of (a) a generalist and a vermivore, (b) a molluscivore and

a vermivore, or (c) two vermivores. ANOVA results revealed no significant differences between these categories in both D_{mature} (ANOVA, $F = 1.69$, $p > 0.05$) and $D_{expression}$ (ANOVA, $F = 2.26$, $p > 0.05$, Figure 1-2).

Dietary breadth and conotoxin complexity

To quantify dietary breadth, we retrieved Shannon diversity index values (H′) representing the diversity of prey species consumed by 10 of the cone snail species in this study that was available in the literature (Table S1-11, (Kohn 1959a, 1966, 1968, 1981; Marsh 1971; Kohn & Nybakken 1975; Chang *et al.* 2015)). To account for phylogenetic non-independence in regression analyses between conotoxin complexity and dietary breadth, we used a phylogenetic generalized least-squares (PGLS) analysis implemented in the caper package within R (Orme 2013). We found a significant positive relationship between averaged H′ derived from the literature and three measures of conotoxin complexity: the number of mature toxins (PGLS, $\lambda = 1$, $p < 0.001$), the number of gene superfamilies (PGLS, $\lambda = 0.861$, $p < 0.05$), and the number of cysteine frameworks (PGLS, $\lambda = 1$, $p < 0.05$) (Table S1-12, Figure 1-3). The inclusion of *C. californicus* in the PGLS analysis may bias the results because *C. californicus* is often regarded as an atypical member of Conidae due to its extremely broad diet and its distant phylogenetic relationship to the rest of Conidae (Elliger *et al.* 2011; Puillandre *et al.* 2014a). When removed, relationships remained significant between dietary breadth and the number of mature toxins (PGLS, $\lambda = 0$, $p < 0.001$), the number of gene superfamilies (PGLS, $\lambda = 0$, $p < 0.05$), but not the number of cysteine frameworks (PGLS, $\lambda = 0$, $p > 0.05$, Table S1-12).

Discussion

Broad scale patterns of cone snail venoms

Through analysis of venom duct transcriptomes of 12 species spanning a broad phylogenetic distribution within Conidae, we were able to provide a snapshot of species level variation in conotoxin expression. For eight species within this study, the total number of predicted mature toxins was within the range of decades-old estimates of 50-200 conotoxins expressed per species (Table 1-1, (McIntosh *et al.* 1999)). The number of mature toxins for the four remaining species was above 200 (Table 1-1). Our species-level estimates of conotoxin diversity conflict with a recent study that documented the existence of 3303 conotoxin precursor peptides from the venom duct transcriptome of *Conus episcopatus* (Lavergne *et al.* 2015). Even when considering the total number of unique conotoxin precursors identified in this study across 12 species (2223 precursors, Table 1-1, Table S1-2), this value is still significantly less than what was reported from *C. episcopatus* (Lavergne *et al.* 2015). Additionally, estimates of conopeptide diversity from our *C. marmoreus* transcriptome (81 unique conotoxin precursors, Table 1-1) is much lower than estimates from a previous study on the same species (263 unique conotoxin precursors, (Lavergne *et al.* 2013; Dutertre *et al.* 2013)). We hypothesized that these large differences occurred because of the somewhat common practice among cone snail transcriptome studies to identify conopeptides directly from read depth and subsequently to not verify each conotoxin through read mapping (Lavergne *et al.* 2013, 2015; Dutertre *et al.* 2013; Jin *et al.* 2013). In many cases, unique conopeptide precursors were only supported by a single read (Dutertre *et al.* 2013; Jin *et al.* 2013). These practices produce over-estimates of conopeptide diversity and can lead to erroneous insights on conotoxin variation among species by confounding biological variation with sequencing errors produced by next-generation sequencing

platforms (Gilles *et al.* 2011; Minoche *et al.* 2011). Indeed, recent studies have invoked molecular mechanisms such as 'mRNA messiness' [31] and 'RNA editing' [51] to explain the unexpected abundance of lowly expressed transcript variants likely caused by sequencing errors. Our results echo the sentiments of a previous study emphasizing the importance of carefully and rigorously examining conotoxin sequences generated by new sequencing technologies (Robinson *et al.* 2014).

Our results confirmed that conotoxin compositions are dominated by a few gene superfamilies and by a few number of transcripts (Conticello *et al.* 2001; Liu *et al.* 2012; Dutertre *et al.* 2013). This pattern was evident whether we examined conopeptide counts (Table S1-7) or expression values (Table 1-2, Table S1-8). Interestingly, this pattern where venom compositions are dominated by a minority of toxins is also evident in some species of spiders (Haney *et al.* 2014) and snakes (Pahari *et al.* 2007; Casewell *et al.* 2009; Rokyta *et al.* 2013). Conticello *et al*. (2001) hypothesized that lowly expressed transcripts, which represent the majority of the conotoxins in cone snail venoms, may not be critical for prey capture and are thus subject to weaker selection pressures that allow for functional divergence over time. These transcripts may provide the substrate for phenotypic novelty as adaptive regimes change, whereby lowly expressed transcripts that have gained new functions become upregulated in response to the environment (Conticello *et al.* 2001). It is unclear whether these processes are occurring in cone snail venoms because it is difficult to determine levels of gene expression that are biologically relevant. Further, the strong relationship between dietary breadth and conotoxin diversity documented in this study suggests that the entire complement of toxins is necessary for species to apprehend prey in their environment, indicating that lowly expressed transcripts likely have a significant role in prey capture (Figure 1-3).

Our estimates of total conotoxin expression in cone snails species from this study were within the range of values estimated from previous work (Table 1-2, (Remigio & Duda 2008; Casewell *et al.* 2009; Zhang *et al.* 2010; Hu *et al.* 2011; Rokyta *et al.* 2012)). Due to low sample sizes per species and a lack of variation in treatment conditions (e.g., age, stress, etc.), we cannot determine whether conotoxin genes are always expressed at similar magnitudes in the venom duct, or expression levels change due to demand at different time points in the lifetime of these organisms. However, if the level of transcription devoted to conotoxin production is constant among adult individuals, total conotoxin expression magnitude may be related to the frequency in which venom is utilized for prey capture, as implicated in (Remigio & Duda 2008). Several laboratory observations report that some cone snail species swallow their prey whole without first injecting venom, though the extent to which this occurs in nature across Conidae is not well documented (Kohn 1959a). Interestingly, *C. californicus* has the lowest total conotoxin expression level estimated in this study and is known to scavenge on dead prey (Saunders & Wolfson 1961), providing some evidence for this hypothesis.

Similar to reports from the *de novo* assembly of the *C. episcopatus* venom duct transcriptome (Lavergne *et al.* 2015), Trinity performed poorly in reconstructing all conopeptide precursors reported in this study (Table S1-3). This may be attributed to the conservation of the signal sequence in precursor peptides, which essentially function as abundant repetitive regions in venom duct RNAseq data. We hypothesize that algorithms devoted to assembly of repetitive regions (e.g., (Bresler *et al.* 2012; Prjibelski *et al.* 2014)) may streamline efforts to characterize Conidae venom duct transcriptomes, which will be the subject of future research efforts.

Conotoxin composition and diet

Although few broad comparative studies exist that examine variation in venom composition patterns across cone snail taxa, prey taxonomic class has remained the dominant framework by which cone snail venom evolution is studied (Olivera *et al.* 1990; Kaas *et al.* 2010) and forms the basis for categorization of conotoxins on ConoServer, a molecular database of known conotoxins (Kaas *et al.* 2010). Contrary to common wisdom derived from numerous previous studies of cone snail venom, our results show unequivocally that prey class performs poorly in predicting conotoxin composition patterns among cone snail species. We did not detect a single gene superfamily that separated vermivores from the molluscivore or the generalist (Table S1-2); rather, the defining feature across cone snail venoms was that every species, regardless of diet category, expressed a unique repertoire of conopeptides and gene superfamilies at different magnitudes (Table S1-2, Table 1-2, Figure 1-1). The uniqueness of each species' conotoxin composition is underscored by moderate to low D values (Table S1-10) and no apparent differences in the distribution of D values among species that do or do not share the same diet class (Figure 1-2). Unfortunately, we were not able to obtain venom duct RNAseq data for several molluscivores and piscivores, but our results are in agreement with a meta-analysis indicating that there is no gene superfamily that there is exclusive to any of the traditionally recognized diet types (Puillandre *et al.* 2012).

The poor predictive power of diet class on venom composition patterns may be due to several factors. Estimated rates of gene duplication and nonsynonymous substitution rates for conotoxin genes are the highest across metazoans (Duda & Palumbi 2000; Chang & Duda 2012) and these extraordinary rates of molecular evolution may be more likely to promote divergence rather than convergence in venom composition, as documented in this study. Alternatively, the lack of predictive power may be in part, due to how categories are constructed for venom

components and diet classes. Prey specialization exists at the level of protein function, but it is known that conotoxin gene superfamilies generally do not provide predictive information on protein function and conotoxins targeting similar neurological targets can evolve convergently in several gene superfamilies (Kaas *et al.* 2010; Robinson & Norton 2014). Further, a past meta-analysis showed that the distribution of signal sequence identities (which are used to distinguish between gene superfamilies) between conotoxins within a gene superfamily and between gene superfamilies are largely overlapping (Puillandre *et al.* 2012), suggesting that the gene superfamilies are somewhat arbitrary constructions in and of themselves. Therefore, characterizing conotoxin composition patterns by gene superfamilies does not fully measure functional similarities and differences in conotoxins between species. Prey classes may not explain conotoxin composition patterns potentially due to an over generalization of the diversity present within the vermivorous category. Vermivory, as traditionally used in Conidae studies, is broadly defined and includes a wide variety of taxa that represent hemichordates, echiurans, and several polychaete families that diverged ≥ 400 million years ago (Duda Jr. *et al.* 2001; Parry *et al.* 2014). Future studies that account for the taxonomic breadth of worms and the functional diversity of conotoxins may better predict patterns of venom composition among species.

We found strong support for a highly significant, positive relationship between venom composition complexity and dietary breadth across cone snails (Figure 1-3), corroborating hypotheses that were made from species that represent the extremes of the dietary breadth spectrum in Conidae (Remigio & Duda 2008; Elliger *et al.* 2011). The total number of mature toxins provided the strongest predictor of this relationship, possibly because mature toxin diversity better encapsulates functional diversity in cone snail venoms. The relatively weaker correlations documented for gene superfamilies and cysteine frameworks support the notion that

these traditional means of classifying conotoxins are not simple or direct correlates of conotoxin function (Figure 1-3, Table S1-12 (Millard *et al.* 2004; Robinson *et al.* 2014)). The positive relationship between dietary niche breadth and venom composition complexity corroborates the niche variation hypothesis, suggesting that diverse venoms are required to subdue diverse prey. Although the majority of studies invoking this hypothesis focus on population level trait variance within species (e.g, (Bolnick *et al.* 2007)), our results extend the generality of this hypothesis to species-level patterns of conotoxin complexity across cone snails. These results also align with observations on venom complexity in snakes (Li *et al.* 2005a; b; Pahari *et al.* 2007), suggesting that dietary breadth may explain evolutionary trends in venom complexity across radiations of venomous taxa.

At the population level, cone snails follow the predictions supported by the niche variation hypothesis, such that snail populations with a larger dietary breadth show greater population-level allelic diversity in some conotoxin loci (Duda & Lee 2009; Chang *et al.* 2015). How these population-level patterns translate into species-level properties remains unclear. Although conotoxin allelic diversity is higher in populations with greater dietary breadth, each individual within the population possesses only a small subset of the available alleles per locus because these alleles belong to individual loci and are not separate genes (Duda & Lee 2009; Chang *et al.* 2015). Presumably, gene products encoded by these loci are effective at paralyzing diverse prey and toxins encoded by variants at these loci may be more effective if expressed in tandem (Duda & Lee 2009). Given the exceptional rates of gene duplication estimated within this group (Chang & Duda 2012), gene duplication presents a mechanism by which allelic variants previously restricted to a single conotoxin locus can ultimately evolve to be expressed

simultaneously through duplication, potentially leading to species-level patterns of venom composition complexity and dietary breadth.

What are the evolutionary consequences of increased dietary breadth? Our results imply that dietary breadth plays a role in determining how many conotoxins are utilized for prey capture. Consequently, this also determines the number of genes available for forces such as mutation, selection, and drift to generate novel functions for adaptation. Higher gene diversity is thought to provide increased opportunities for novel phenotypes to arise (Crow & Wagner 2006), potentially shaping a lineage's evolvability, or a lineage's capacity to evolve in response to their environment (Kirschner & Gerhart 1998). Therefore, greater dietary breadth (leading to higher gene diversity) may signify a larger potential for lineages to diversify, potentially influencing patterns of species diversification in cone snails. Although venom is viewed as a key innovation and thought to play a major role in the evolutionary diversification of venomous taxa (Fry *et al.* 2009; Pyron & Burbrink 2011), the interplay between diet and venom on patterns of lineage diversification is rarely tested explicitly. High resolution molecular phylogenies of venomous taxa and comprehensive venom composition data that can now be rapidly obtained using new sequencing technologies will provide the necessary datasets to facilitate an examination of the evolutionary dynamics of diet, venom, and speciation over long evolutionary time-scales.

A recent study showed that cone snail species may inject separate suites of conopeptides for predation and defense, and that defensive venoms are produced in the proximal region of the venom duct while predatory venoms are produced in the distal region (Dutertre *et al.* 2014). Because we generated our venom duct RNAseq data from the entire length of the organ, our results may be confounded between these two distinct ecological roles that venom performs. However, it is unclear how broad this pattern is throughout cone snails. In many cases, functional

work has shown that venoms extracted from different regions of the venom duct were able to successfully paralyze prey (Endean & Rudkin 1963, 1965). In addition, the functional roles of conotoxins that are relevant to each species' ecology are poorly understood. Conotoxin function is typically determined by assays in mice or vertebrate neuronal cells which are not representative of the intended targets of conotoxins (Robinson & Norton 2014); as previously asserted in snake systems (Barlow *et al.* 2009; Richards *et al.* 2012) over-interpretation of these results can lead to misleading or conflicting inferences. For example, a δ-conotoxin (a type of conotoxin thought to be critical for fish-hunting) isolated from the vermivore, *Conus susturatus*, was implicated as a defensive toxin against fish predators due to its effects on vertebrate sodium ion channels and its proximal expression in the venom duct, where defensive toxins are thought to be synthesized (Jin *et al.* 2015). However, behavioral observations from *Conus tessulatus*, a vermivore in the same subgenus that also expresses a similar δ-conotoxin, was shown to prey on fish on occasion through venom injection (Aman *et al.* 2015). These results emphasize the necessity of examining conotoxins in the context of each species' ecology to accurately understand the natural history of cone snails.

Conclusion

In contrast to the most widely accepted hypothesis of cone snail venom evolution, diet class did not predict patterns of venom composition among cone snails. These results suggest either (a) the fast rates of venom evolution drive rapid divergence of conotoxin composition that bear no relationship to prey taxonomic class, or (b) current ways of categorizing both prey species (i.e., worm, mollusc, fish) and conotoxins (i.e., gene superfamily) fail to accurately reflect evolutionary interactions between dietary specialization and venom function. Therefore,

future studies placing more emphasis on the taxonomic breadth of cone snail prey and conotoxin function on prey capture may better encapsulate the impact of diet on cone snail venom evolution. In addition, our results highlight the importance of dietary breadth in shaping species-level venom complexity patterns among cone snails. To our knowledge, this relationship is rarely tested quantitatively across venomous radiations despite its potential to explain variation in venom complexity as demonstrated here. While our results show that species with broad diets tend to have more diverse venoms, the evolutionary consequences of this tendency remains unclear. What is certain is that selective pressures driven by diet plays a major role in shaping evolutionary patterns in venom across cone snails and other venomous taxa.

Methods

Sampling and sequencing

We collected one individual from 11 species of *Conus* from Bali, Indonesia (*C. arenatus, C. coronatus, C. ebraeus, C. imperialis, C. lividus, C. marmoreus, C. quercinus, C. rattus, C. sponsalis, C. varius, C. virgo*) and WF Gilly provided 1 *C. californicus* species from Monterey Bay, California. We immediately placed dissected venom ducts from live snails in RNALater and stored samples in a 4°C refrigerator until they could be placed in a -20°C freezer within 2 weeks of collection. All snails were adults and were starved for at least 24 hours prior to the dissections. We isolated RNA using TRIzol reagent (Invitrogen, USA) and purified the sample using a Qiagen RNeasy Mini Kit. We extracted RNA from the entire venom duct, or along sections of the venom duct if it was particularly long because venom composition is known to change along the length of the duct in some species (Hu *et al.* 2012). We used Bioanalyzer traces to assess total RNA quality and to determine suitability for sequencing. We constructed cDNA

libraries by using the TruSeq RNA Sample Prep Kit to recover mRNA via Poly-A selection, synthesize cDNA, ligate adapters, and index samples. We sequenced all 12 samples on a single Illumina HiSeq 2000 lane with 100-bp paired-end reads.

Transcriptome assembly

During initial attempts to assemble transcripts in Trinity, we were not able to assemble known transcripts present in the sequencing data, potentially due to the repetitiveness and high sequence complexity of venom transcripts. To circumvent this issue, we employed an iterative assembly approach. For each iteration, we trimmed adapters and low quality bases using Trimmomatic (Bolger *et al.* 2014), merged reads using FLASh (Magoč & Salzberg 2011), and assembled transcripts using Trinity (Grabherr *et al.* 2011). During the first assembly iteration, we assembled a 0.1% random subset of the total reads for each sample. Then, we used blastx to identify transcripts with similarity (evalue = 1-e10) to known conotoxin genes listed on ConoServer. We used bowtie2 (Langmead & Salzberg 2012) to align and identify reads that matched to those putative venom transcripts. For the second iteration, we assembled reads from the 0.1% subset that did not align to venom transcripts identified from the first iteration. Then, we identified additional putative venom transcripts from the contigs generated. For the final iteration, we assembled reads from the full dataset that did not align to venom transcripts identified from the first two iterations and identified additional contigs that shared similarity to conotoxins.

Conopeptide identification

We used Conosorter to identify novel venom transcripts and took a conservative approach towards accepting venom transcripts because ConoSorter has a tendency to over-classify sequences. For example, the recently discovered Y2 superfamily identified through ConoSorter is actually molluscan insulin (Safavi-Hemami *et al.* 2015). We used ConoSorter to analyze transcripts with TPM values > 1000 and retained conotoxins that (1) had all three conotoxin regions (i.e., signal region, propetide region, and mature toxin coding region) and (2) a precursor protein length > 38 and < 200 (boundaries were generated based from the empirical length distribution of conotoxin proteins identified from this study). We used blastx to query the novel venom transcripts for similar sequences in every transcriptome in our dataset and also against the ConoServer database. We retained sequences if similar signal regions could be found in the transcriptomes of other species or in ConoServer. We removed transcripts if they produced erroneous blast results (e.g., best-scoring transcript in a different species' transcriptome produced a protein with several stop codons), suggesting that the novel conotoxin identified by ConoSorter may have been in an incorrect reading frame.

To generate a venom gene reference for each species, we combined all venom transcripts from each assembly iteration and transcripts identified through ConoSorter. We used Python scripts to remove transcripts with redundant proteins and transcripts belonging to the incorrect species. We identified several cases where highly expressed transcripts in one species could be found at low representation in some, or in all of the other species sequenced. We note that cross-contamination across every single sample is unlikely, given that these samples were prepared in different sets and were pooled just before sequencing. We hypothesized that this phenomenon occurred due to cluster misidentification during sequencing, potentially due to high sequence similarity of conotoxin transcript signal sequences. We used blastp to identify and remove

transcripts from species that had high identity (>95%) in the protein coding region to another species. Here, we do not expect >95% identity across the entire conotoxin precursor protein across the taxa in this study, given the exceptionally high nonsynonymous substitution rates estimated in venom genes from Conidae [37]. We chose the transcript from the species that had the highest coverage (estimated using bowtie2) to be the true transcript.

To reassemble transcripts that were incomplete (missing start or stop codon), we used an approach called Assembly by Reduced Complexity (ARC, https://github.com/ibest/ARC). ARC is a pipeline that allows for *de novo* assembly of specific targets by only assembling reads that map to reference targets. We removed venom transcripts that could not be reassembled, or were not full length (included a start and stop codon) after a maximum of three ARC iterations. Then, we used bowtie2 to map reads to all venom genes to verify the nucleotide sequence of each putative conotoxin. We removed sequences that did not have reads aligning to the entire length of the transcript.

Through mapping, we identified sequence polymorphisms in the conotoxin transcripts. In some cases, these polymorphisms represented allelic differences and we generated a separate conotoxin sequence if the sequence translated into a unique precursor peptide. In other cases, the polymorphisms represented completely distinct conotoxin transcripts that were never assembled, but partially mapped to the existing reference. We assembled these conotoxin transcripts by manually aligning representative reads in Geneious (Biomatters, Auckland, New Zealand) and verified each sequence through additional read mapping. To check for chimeric sequences, we generated 80bp fragments every 20bp along the length of each transcript and searched for the existence of these fragments directly from read depth for sequences that had > 30X coverage. We manually examined sequences flagged by this filter and removed sequences if necessary. We

used ConoPrec to remove sequences that did not have a clearly defined signal sequence. Finally, we manually inspected all venom genes to identify any unusual conotoxin transcripts.

Conopeptide classification

We employed several approaches to classify conotoxins into gene superfamilies. First, we compared conotoxin transcripts to sequences from the ConoServer database using a blastx search and assigned transcripts to gene superfamilies using the best-scoring hit. For transcripts that did not have a blast hit, we compared signal sequences using blastp against the ConoServer database and classified transcripts to gene superfamilies that had a percent signal sequence similarity > 76%, a threshold used in a previous study (Barghi *et al.* 2014). We noted that the Divergent_M---L-LTVA superfamily was composed of several transcripts with unique signal sequences. With members of this superfamily along with all other unclassified transcripts, we aligned signal sequences and generated a pairwise distance matrix in Geneious. Then, we used a custom Python script to cluster conopeptides that shared a percent signal sequence similarity > 70%. We derived this threshold empirically to minimize the number of clusters, yet still represent salient differences among clusters (e.g., similar cysteine frameworks). We provided names for novel, reclassified, and unclassified superfamilies with five letters representing the first five amino acids that the majority of their constituent sequences shared.

To provide names for conotoxin precursors identified in this study, we followed the naming conventions similar to (Barghi *et al.* 2015a). Briefly, we named each conotoxin with the following: two letters to denote the species, the gene superfamily name, and a number denoting the order of discovery within the gene superfamily for that species. These fields are separated

with an underscore. We did not provide new names for previously identified conotoxins unless there was evidence of species misidentification in previous work.

Phylogeny inference

For all transcripts not classified as conotoxins, we used CAP3 (Huang & Madan 1999) to reduce redundancy and annotated the transcriptomes using blastx against the *Lottia gignatea* (owl limpet) protein database (Simakov *et al.* 2013). To identify putatively orthologous loci for phylogenetic reconstruction, we employed a reciprocal blast approach via blastx and tblastx between each species' transcriptome and the *L. gigantea* database. We retained loci that had at least 10 species represented. For each of these loci, we also considered other contigs within each species' transcriptome that was annotated with the same protein, but spanned a non-overlapping portion because transcriptomes are often fragmentary. We created alignments using MAFFT (Katoh *et al.* 2005) and manually inspected each locus in Geneious. We used a custom Python script to calculate uncorrected patristic distances between all possible pairwise comparisons of the taxa in this study. For each comparison, we removed loci with patristic distances greater than two standard deviations away from the mean to remove potential paralogous sequences. We concatenated all loci and inferred a phylogeny using RAxML under a GTRGAMMA model of sequence evolution with 100 bootstrap replicates [35], rooting our tree using *C. californicus* based on previous phylogenetic hypotheses [21].

We dated the maximum likelihood phylogeny generated from RAxML using the program r8s with two fossil calibrations: a fixed rate of 55 my (million years) representing the origin of cone snails in the fossil record at the root of the tree (Kohn 1990), and a minimum constrained age of 11my (the earliest date showing fossil evidence of both *C. lividus* and *C. quercinus*) at the

node representing the ancestor of *C. lividius* and *C. quercinus* (Duda Jr. *et al.* 2001). The inclusion of *C. californicus* in this study allowed us to place the fossil calibration at the root because this node possibly represents the ancestor to all Conidae (Puillandre *et al.* 2014a).

Conotoxin expression

We removed transcripts with significant homology to the mtDNA genome of *Conus consors* and the *L. gigantea* non-coding RNA database via blastn to remove potential biases associated with quantifying venom expression. To normalize read counts, we used the RSEM algorithm to map reads with bowtie2 and generate TPM values. We only included the conotoxin coding regions in the mapping reference when calculating conotoxin expression levels.

Diet and conotoxin composition

To calculate overlap in conotoxin composition among species, we employed Schoener's D statistic (Schoener 1968):

$$D(p_x, p_y) = 1 - \frac{1}{2}\sum_i | p_{x,i} - p_{y,i} |$$

where p_x and p_y represent the frequencies for species x and species y for the ith category. D ranges from 0 (no niche overlap) to 1 (niches are identical). We calculated D by using (a) the percentage of mature conotoxins belonging to each gene superfamily (D_{mature}) and (b) the percentage of overall conotoxin expression levels for each gene superfamily ($D_{expression}$) for all possible pairwise species comparisons. We categorized each comparison as either occurring between (a) a generalist and a vermivore, (b) a molluscivore and a vermivore, or (c) between two vermivores. To control for phylogenetic signal, we generated a pairwise phylogenetic distance

matrix from the time-calibrated phylogeny using the R package picante (Kembel *et al.* 2010). Then, we calculated residuals from a linear regression model between phylogenetic distance and both D values. We used the residuals in an ANOVA to determine whether the distribution of overlap values between the species comparisons were significantly different based on diet.

Dietary breadth and venom complexity

We obtained dietary breadth measurements estimated by Shannon's diversity index values (H′) for 10 species from primary literature (Kohn 1959a, 1966, 1968, 1981; Marsh 1971; Kohn & Nybakken 1975; Chang *et al.* 2015). We retrieved H′ values directly from these past studies or calculated them if necessary. When calculating H′, we ignored categories that appeared to represent an amalgamation of several unidentified taxa. We only included H′ values that were calculated from at least 5 individuals where the prey item could be identified to the genus level. We generated average H′ for each species rather than recalculate H′ values for the total number of prey taxa that each species can consume, because each cone snail species preyed on a different set of taxa depending on the geographic locality and what congeners were present (Kohn 1959a, 1966, 1968; Marsh 1971; Kohn & Nybakken 1975).

We quantified conotoxin composition diversity as either the (a) number of mature toxins, (b) number of gene superfamilies, or (c) number of cysteine frameworks. We correlated these values with averaged values of H′ in a PGLS analysis implemented in the R package caper. When performing regression analyses, the PGLS function in caper incorporates a covariance matrix by using branch lengths from an ultrametric phylogeny and assuming a Brownian motion model of trait evolution (Orme 2013). We executed these analyses with and without *C. californicus*.

Data availability

Raw read data are available at the National Center for Biotechnology Information Sequence Read Archive (Accession Numbers SRX1323883-SRX1323894). All conotoxin transcript sequences will be uploaded to ConoServer. All scripts and final datasets used for analyses will be uploaded onto Dryad.

Acknowledgements

We thank AJ Kohn and TF Duda Jr. for advice on working with cone snails in the field; DST Hariyanto, MBAP Putra, MKAA Putra, and the staff at the Indonesian Biodiversity Research Center in Denpasar, Bali for assistance in the field; the community at the Evolutionary Genetics Lab at UC Berkeley for laboratory support; and J Chang, MCW Lim, E McCartney-Melstad, WF Gilly, the J McGuire lab group, the UCLA next-generation sequencing working group, and two anonymous reviewers for insightful comments on earlier versions of this manuscript. This work was supported by a research grant from the London Malacological Society, a Grants-in-Aid of research from Sigma Xi, a research grant from the Ecology and Evolutionary Biology department at UCLA, the Lerner Gray Fund for Marine Research from the American Museum of Natural History, an Edwin W. Pauley Fellowship from UCLA, a Fulbright Fellowship, and a NSF Graduate Research Fellowship awarded to MAP. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. We thank the Indonesian Ministry of State for Research and Technology (RISTEK, permit number 277/SIP/FRP/SM/VIII/2013) for providing permission to conduct fieldwork in Bali. The *C. californicus* specimen was collected under a

California Department of Fish and Wildlife collecting permit granted to WF Gilly (SC-6426). We thank C. Brown for illustrating the images in Figure 1-1.

List of Figures

Figure 1-1. Conotoxin expression in a phylogenetic context. Time-calibrated maximum likelihood phylogeny of Conidae species sequenced in this study generated from 821 loci. Values at nodes represent bootstrap support and • indicates bootstrap support = 100. Tree is rooted with *Californiconus californicus*. Taxa are colored by diet (green = generalist, black = vermivore, orange = molluscivore). Heat map shows relative contribution (measured as percentage of total conotoxin TPM per species) of gene superfamilies that contributed to at least 10% of overall conotoxin expression in at least one species.

Figure 1-2. Conotoxin composition overlap and dietary preference. Boxplots showing the distribution of conotoxin overlap values (D) categorized by whether the species comparison occurred between a generalist and a vermivore, a molluscivore and a vermivore, or two vermivores. Values were calculated by the percentage of mature toxins belonging to each gene superfamily (D_{mature}) and the percent expression of gene superfamilies ($D_{expression}$).

Figure 1-3. Dietary breadth and conotoxin complexity. Correlations between dietary breadth (Averaged H′) and measures of conotoxin complexity: number of mature toxins, number of gene superfamilies, and number of cysteine frameworks. Graphs are labelled with correlation coefficients. *denotes significant correlation from a PGLS analysis.

List of Tables

Figures

Figure 1-1.

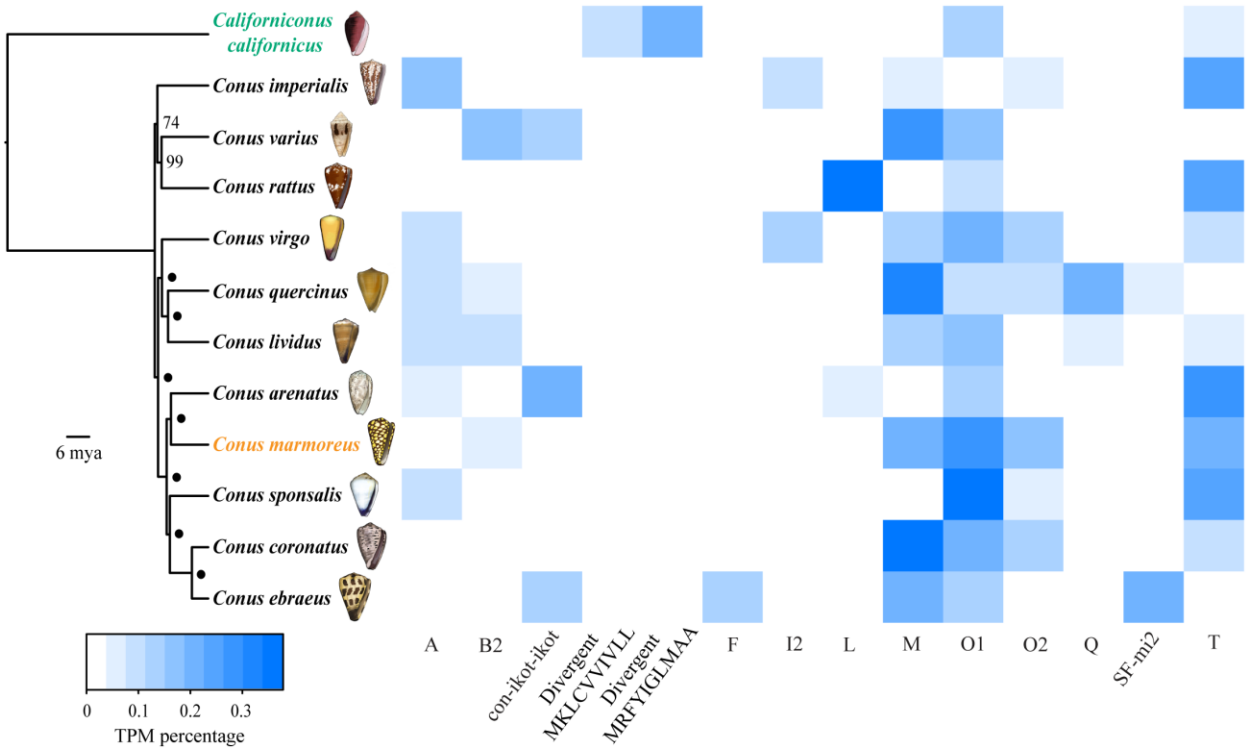


Figure 1-2.

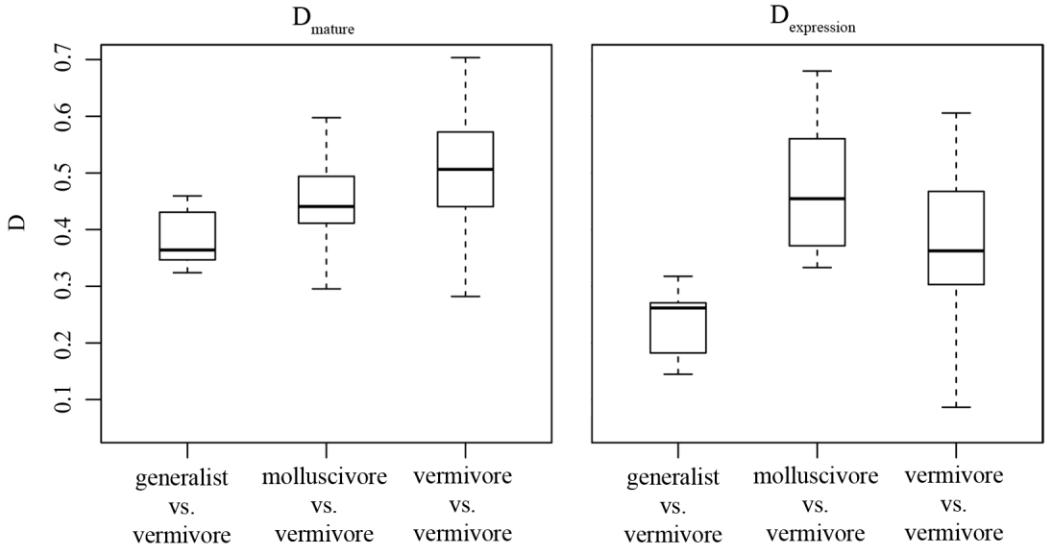
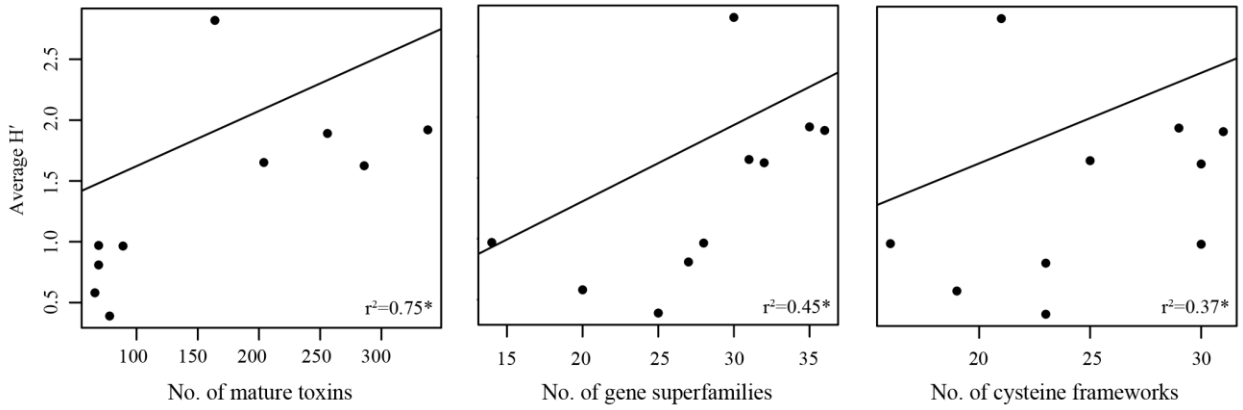


Figure 1-3.


Average H'
0.5
1.0
1.5
2.0
2.5
r²=0.75*
100
150
200
250
300
No. of mature toxins
r²=0.45*
15
20
25
30
35
No. of gene superfamilies
r²=0.37*
20
25
30
No. of cysteine frameworks

Tables

Table 1-1.

Genus/subgenus	Species	No. of unique conotoxin precursors	No. of unique mature toxins	No. of gene superfamilies	No. of cysteine frameworks	Most abundant gene superfamily and frequency*	Main diet summarized from [29, 42]
Puncticulis	*arenatus*	326	256	36	31	O1, (20.3%)	eunicids, nereids, capitellids
Californiconus	*californicus*	185	164	30	21	O1, (20.1%)	molluscs, polychaetes, fish
Virroconus	*coronatus*	331	286	32	30	O1, (19.6%)	eunicids, capitellids
Virroconus	*ebraeus*	75	69	27	23	M, (31.9%)	eunicids, nereids
Stephanoconus	*imperialis*	70	66	20	19	P, (16.7%)	amphinomids
Lividoconus	*lividus*	244	204	31	25	O1, (10.7%)	enteropneusts, terebellids
Conus	*marmoreus*	81	69	14	16	M, (26.1%)	gastropods
Lividoconus	*quercinus*	97	78	25	23	O1, (15.4%)	enteropneusts, sabellids
Rhizoconus	*rattus*	102	89	28	30	con-ikot-ikot, (18%)	eunicids
Harmoniconus	*sponsalis*	401	338	35	29	O1, (28.1%)	eunicids, nereids
Strategoconus	*varius*	198	168	29	24	M, (10.7%)	polychaetes
Virgiconus	*Virgo*	113	78	25	21	O1, (24.4%)	terebellids

*calculated from no. of unique mature toxins

Table 1-2

Species	Total conotoxin expression	Most highly expressed gene superfamily (frequency)	No. of superfamilies representing > 50% TPM values	Most highly expressed mature toxin (superfamily, frequency)	No. of mature toxins representing > 50% TPM values
arenatus	57.6%	T (30.2%)	2	Ar_T_9 (T, 10.0%)	11
californicus	26.0%	Divergent_MRFYIGLMAA (20.5%)	4	Cl_DivMRFYIGLMAA_6 (Divergent_MRFYIGLMAA, 16.3%)	10
coronatus	70.7%	M (37.9%)	2	Co_M_18 (M, 13.5%)	11
ebraeus	45.4%	M (21.9%)	3	Eb_SF-mi2_2 (SF-mi2, 15.9%)	5

imperialis	64.3%	T (26.1%)	3	Im5.4 (T, 23.2%)	5
lividus	56.1%	O1 (17.2%)	5	Li_O1_25 (O1, 5.2%)	18
marmoreus	67.9%	O1 (27.3%)	3	MaI51 (O2, 17.2%)	6
quercinus	49.5%	M (32.9%)	2	Qc_M_13 (M, 30.5%)	4
rattus	35.5%	L (36.7%)	2	Rt_L_3 (L, 29.3%)	2
sponsalis	55.7%	O1 (36.7%)	2	Sp_A_4 (A, 6.0%)	23
varius	38.5%	M (26.6%)	3	Vr3-SP02 (M, 16.8%)	5
Virgo	68.9%	O1 (21.3%)	4	Vi_M_2 (M, 8.9%)	11

List of supplementary material

Figure S1-1. New conotoxin gene superfamilies described in this study. Signal sequences are underlined, mature toxin regions are bolded, and cysteines within the mature toxin region are highlighted. Ar = *C. arenatus*, Co = *C. coronatus*, Im = *C. imperialis*, Li = *C. lividus*, Qc = *C. quercinus*, Rt = *C. rattus*, Sp = *C. sponsalis*, Vi = *C. virgo*.

Figure S1-2. Identical mature toxin from two different species. Full conopeptide precursor sequences of identical mature toxins expressed in more than one species. Signal sequences are underlined, mature toxin regions are bolded, and cysteines within the mature toxin region are highlighted. Co = C. coronatus, Vi = C. virgo.

Figure S1-3. Pseudogene identified from *C. sponsalis*. Full precursor peptide sequences from the functional and pseudogenized copy of a mature toxin identified from *C. sponsalis* (Sp = *C. sponsalis*). TPM values are shown, signal sequences are underlined, mature toxin regions are bolded, and cysteines within the mature toxin region are highlighted.

Table S1-1. Sequencing and assembly statistics for each species. Values were calculated using all transcripts assembled from all iterations of Trinity.

Table S1-2. Conotoxin diversity by gene family for each species sequenced in this study. Each value represents the number of unique predicted mature toxins. Values in parentheses represent the number of unique precursor peptides.

Table S1-3. Comparison between final conotoxin dataset and conotoxins assembled through three iterations of Trinity. Number and percent of unique precursor peptides from the final dataset that had at least 95% sequence identity to transcripts assembled by trinity and that matched to (a) 90%, (b) 95%, or (c) 99% of the total precursor sequence length.

Table S1-4. Conopeptide gene superfamilies that were reclassified and provided new names in this study.

Table S1-5. Cysteine frameworks identified in this study for each gene superfamily. * indicates novel cysteine framework discovered in this study. [a] indicates a cysteine framework found in cone snail venoms, but not yet described from this gene superfamily.

Table S1-6. Conopeptides sequenced in this study with 100% protein sequence identity to a different species on ConoServer. Matches ≥95% are shown for conopeptides described *Conus flavidus* to highlight similarities with the *C. lividus* transcriptome from this study.

Table S1-7. Gene superfamilies that represent > 50% of the conopeptides identified per species.

Table S1-8. Gene superfamilies and conotoxins representing major components of each species' venom expression levels. Divergent is abbreviated as Div.

Table S1-9. Relative contribution of each gene superfamily to conotoxin expression. Values calculated as % conotoxin TPM (superfamily TPM/total conotoxin TPM)

Table S1-10. Conotoxin composition overlap values (D). Values calculated by (a) the frequency of conopeptide membership to different gene superfamilies (b) the percent expression level of each gene superfamily.

Table S1-11. Dietary breadth, measured by Shannon-Wiener's index (H′), for 10 cone snail species sequenced in this study.

Table S1-12. Results from Phylogenetic Generalized Least Squares (PGLS) analysis assessing the relationship between prey breadth (H') and conotoxin complexity.

Chapter 2: Targeted sequencing of venom genes from cone snail genomes improves understanding of conotoxin molecular evolution

Abstract

To expand our capacity to discover venom sequences from the genomes of venomous organisms, we applied targeted sequencing techniques to selectively recover venom gene superfamilies and non-toxin loci from the genomes of 32 cone snail species (family, Conidae), a diverse group of marine gastropods that capture their prey using a cocktail of neurotoxic peptides (conotoxins). We were able to successfully recover conotoxin gene superfamilies across all species with high confidence (> 100X coverage) and used these data to provide new insights into conotoxin evolution. First, we found that conotoxin gene superfamilies are composed of 1-6 exons and are typically short in length (mean = ~85bp). Second, we expanded our understanding of the following genetic features of conotoxin evolution: (a) *positive selection*, where exons coding the mature toxin region were often three times more divergent than their adjacent noncoding regions, (b) *expression regulation*, with comparisons to transcriptome data showing that cone snails only express a fraction of the genes available in their genome (24%-63%), and (c) *extensive gene turnover*, where Conidae species varied from 120-859 conotoxin gene copies. Finally, using comparative phylogenetic methods, we found that while diet specificity did not predict patterns of conotoxin evolution, dietary breadth was positively correlated with total conotoxin gene diversity. Overall, the targeted sequencing technique demonstrated here has the potential to radically increase the pace at which venom gene families are sequenced and studied, reshaping our ability to understand the impact of genetic changes on ecologically relevant phenotypes and subsequent diversification.

Introduction

Understanding the molecular basis for adaptation and speciation is a central goal in evolutionary biology. Past studies have described several genetic characteristics that seem to be associated with rapidly radiating clades or the evolution of novel phenotypes, including evidence for diversifying selection, gene gains and losses, and accelerated rates of sequence evolution (Henrissat *et al.* 2012; Guillén *et al.* 2014; Brawand *et al.* 2014; Cornetti *et al.* 2015; Malmstrøm *et al.* 2016; Pease *et al.* 2016). Although large-scale comparative genomic studies have vastly increased our knowledge of the genetic changes associated with diversification, the link between genotype and ecologically relevant phenotypes frequently remains unclear. Often, the functional consequences of genetic patterns such as an excess of gene duplicates or regions under positive selection are unknown (Brawand *et al.* 2014; Cornetti *et al.* 2015; Pease *et al.* 2016), limiting our ability to understand how genetic changes shape the evolutionary trajectory of species.

Animal venoms provide an excellent opportunity to study the interplay between genetics and adaptation because of the relatively simple relationship between genotype, phenotype, and ecology. Venoms have evolved multiple times throughout the tree of life (e.g., spiders, snakes, and snails) and play a direct role in prey capture and survival (Barlow *et al.* 2009; Casewell *et al.* 2013). Venoms are composed of mixtures of toxic proteins and peptides that are usually encoded directly by a handful of known gene families (Kordis & Gubensek 2000; Fry *et al.* 2009; Casewell *et al.* 2013). Exceptionally high estimated rates of gene duplication and diversifying selection across these venom genes families are thought to contribute to the evolution of novel proteins and thus changes in venom composition (Duda & Palumbi 1999; Gibbs & Rossiter 2008; Chang & Duda 2012), allowing venomous taxa to specialize and adapt onto different prey species (Kohn 1959b; Daltry *et al.* 1996; Li *et al.* 2005a; Barlow *et al.* 2009; Chang & Duda

2016; Phuong *et al.* 2016). Therefore, the study of venomous taxa can facilitate understanding of the genetic contributions to ecologically relevant traits and subsequent diversification.

A fundamental challenge associated with the study of venom evolution is the inability to rapidly obtain sequences from venomous multi-gene families. Traditionally, venom genes were sequenced through cDNA cloning techniques, which can be labor intensive and time-consuming (Gibbs & Rossiter 2008; Chang *et al.* 2015). Although transcriptome sequencing has greatly increased the pace of venom gene sequencing and the discovery of previously undescribed gene families (e.g., Casewell et al. 2009; Phuong et al. 2016), transcriptome sequencing still requires fresh RNA extracts from venom organs, which may be difficult to obtain for rare and/or dangerous species. Targeted sequencing approaches have vastly improved the capacity to obtain thousands of markers across populations and species for ecological and evolutionary studies (Faircloth *et al.* 2012; Bi *et al.* 2012). Until now, these approaches have not been applied to selectively sequence venomous genomic regions. This may be in part, due to the extraordinary levels of sequence divergence exhibited by venom loci (Gibbs & Rossiter 2008; Chang & Duda 2012), potentially rendering probes designed from a single sequence from one gene family unable to recover any other sequences in the same family (Fig. 2-1). However, past studies have shown that noncoding regions (i.e., introns, untranslated regions [UTRs]) adjacent to hypervariable mature toxin exons are conserved between species (Nakashima *et al.* 1993, 1995; Gibbs & Rossiter 2008; Wu *et al.* 2013), suggesting that these conserved regions can be used for probe design to potentially recover all venom genes across clades of venomous taxa.

Here, we used a targeted sequencing approach to recover venom genes and study the evolution of venom gene families across 32 species of cone snails from the family, Conidae. Cone snails are a hyper diverse group of carnivorous marine gastropods (> 700 spp.) that capture

their prey using a cocktail of venomous neurotoxins (Puillandre *et al.* 2014a). Cone snail venom precursor peptides (conotoxins) are typically composed of three regions: the signal region that directs the protein into the secretory pathway, the prepro region that is cleaved during protein maturation, and the mature region that ultimately becomes the mature peptide (Robinson & Norton 2014). In some instances, there exists a 'post' region of the peptide following the mature region that is also cleaved during protein processing (Robinson & Norton 2014). Conotoxins are classified into > 40 gene superfamilies (e.g., A superfamily, O1 superfamily, etc.) based on signal sequence identity, though some gene superfamilies were identified based on domain similarities to proteins from other venomous taxa (Robinson & Norton 2014). To examine the evolution of conotoxin gene superfamilies from genomic DNA, we designed probes targeting over 800 non-conotoxin genes for phylogenetic analyses and conotoxins from 12 previously sequenced Conidae transcriptomes (Phuong *et al.* 2016). With the recovered conotoxin loci, we describe several features of conotoxin genes, including its genetic architecture, molecular evolution, expression patterns, and changes in gene superfamily size. Finally, we use comparative phylogenetic methods to test whether diet specificity or dietary breadth can explain patterns of gene superfamily size evolution.

Results

Exon capture results

We used custom designed 120bp baits (custom MYbaits-1 kit, 20,000 bait sequences; Arbor Biosciences, Ann Arbor, Michigan, USA) to selectively target phylogenetic markers and conotoxin genes from 32 Conidae species (Table S2-0). We sequenced all samples on a single

Illumina HiSeq2000 lane, producing an average of 12.8 million reads per sample (Table S2-0). After redefining exon boundaries for the phylogenetic markers, we generated a reference that consisted of 5883 loci. We recovered an average of 5335 loci (90.7%) across all samples representing ~0.66 Mb (Megabases) on average (Table S2-0). For the conotoxin loci, given that conotoxin introns can be several kilobases in length (Wu *et al.* 2013) and the average insert size of the libraries was ~350bp, we were only able to assemble conotoxin exon fragments (conotoxin exons with any adjacent noncoding regions). The number of sequences we assembled containing a conotoxin exon ranged from 281 fragments in *Conus papilliferus* to 2278 fragments in *C. atristophanes* (Table S2-0). Approximately 48.8% of the reads mapped to both the phylogenetic markers and venom genes with 52.3% of these reads being marked as duplicates (Table S2-0). Average coverage across the phylogenetic markers was 95.9X, while the average coverage for the conotoxin exons was 149.6X (Table S2-0).

We recovered representative exons from all 49 conotoxin gene superfamilies targeted, plus exons from the Q gene superfamily which we did not explicitly target (Fig. S2-1). Of the 49 targeted gene superfamilies, 'capture success' (defined in Materials and Methods) was 80% or above for 34 gene superfamilies, even though we did not explicitly target every single transcript (Table S2-1). For example, we only targeted 1 sequence of the A gene superfamily from *C. varius*, but we recovered sequences that showed high identity to every single transcript from the A gene superfamily discovered in the *C. varius* transcriptome (Table S2-1). We assessed the ability of targeted sequencing to recover conotoxins from species that were not explicitly targeted in the bait sequences by calculating the number of previously sequenced conotoxins (obtained via Genbank and Conoserver (Kaas *et al.* 2010)) recovered in our dataset (Table S2-2). We recovered a higher percentage of previously sequenced conotoxins if the species was

included in the bait design (52.35%, Table S2-2) compared to species not included in the bait design (39.5%, Table S2-2).

Conotoxin genetic architecture

Through analyses of conotoxin genetic structure across species, we found that the number of exons that comprise a conotoxin transcript ranged from 1 to 6 exons and exon size ranged from 5bp to 444bp, with an average length of 85.2 bp (Fig. 2-2, Table S2-3). Whether or not UTRs were adjacent to terminal exons was dependent on the gene superfamily, with some gene superfamilies always having both 5' and 3' UTRs adjacent to terminal exons and some where the 5' or 3' UTRs cannot be found directly adjacent to the terminal exons (Table S2-3). Regions in conotoxin transcripts identified as the signal region, the mature region, or the post region were most often confined to a single exon (Fig. 2-3). In contrast, the prepro region was more frequently distributed across more than one exon (Fig. 2-3).

Conotoxin molecular evolution

To determine if there are differences in divergence depending on what conotoxin precursor peptide region each exon contains, we calculated uncorrected pairwise differences to quantify the level of sequence divergence between exons and immediately adjacent noncoding regions. Exons containing the signal region were more conserved than their adjacent noncoding regions (average relative ratio < 1, Table S2-4, Fig. 2-4, S2-2). In contrast, all other exon classifications generally showed the opposite pattern, where the exons were typically more divergent relative to their adjacent noncoding regions (average relative ratio > 1, Table S2-4, Fig. 2-4, S2-2). The largest contrast in divergence between exons and adjacent noncoding regions came from exons

containing the mature region, where the coding region was on average 2.9 times more divergent than regions surrounding the exon (Table S2-4, Fig. 2-4, S2-2). For comparison, exons from non-conotoxin genes were more conserved than their adjacent noncoding regions (average relative ratio < 1, Fig. 2-4).

Conotoxin expression

To examine expression regulation across gene superfamilies and species, we compared transcriptomes we previously sequenced (Phuong *et al.* 2016) to the targeted sequencing data. The proportion of conotoxin genes expressed per gene superfamily was highly variable (Table S2-5, Fig. S2-3) and the exact proportion depended on the gene superfamily and the species. In several cases, all gene copies of a gene superfamily were not expressed in the transcriptome (e.g., *Conus ebraeus,* A gene superfamily, 0/9 copies expressed, Table S2-5, Fig. S2-3), and in other cases, all copies were expressed in the transcriptome (e.g., *C. californicus,* O3 gene superfamily, 3/3 copies expressed, Table S2-5, Fig. S2-3). The average proportion of gene copies expressed per gene superfamily per species was 45% (range: 24% – 63%, Table S2-5).

Conotoxin gene superfamily size evolution

With a concatenated alignment of 4441 exons representing 573854bp, we recovered a highly supported phylogeny with all but 4 nodes having ≤ 95% bootstrap support (Fig. 2-5). Total conotoxin gene diversity ranged from as low as 120 in *C. papilliferus* to as high as 859 in *C. coronatus* (Fig. 2-5). 25 gene superfamilies showed evidence of phylogenetic signal in gene superfamily size, such that closely related species tended to have similar gene superfamily sizes (Table S2-6). For example, a clade consisting of *C. coronatus, C. aristophanes,* and *C. miliaris*

contains nearly 5 times more gene copies of the O1 superfamily than their immediate sister clade. (Fig. 2-5). CAFE v3.1 (Han *et al.* 2013) estimates of net gene gains and losses showed that species-specific net conotoxin expansions and contractions are scattered throughout the phylogeny (Fig. 2-5, S2-4).

Diet and conotoxin gene superfamily evolution

We used comparative phylogenetic methods and extensive prey information from the literature to examine the impact of diet specificity (i.e. what prey a cone snail feeds upon) and dietary breadth (i.e., how many prey species a cone snail feeds upon) on two measures of conotoxin composition: (a) gene superfamily size and (b) total conotoxin diversity. Neither diet specificity nor dietary breadth was correlated in changes with gene superfamily size (D-PGLS [distance-based phylogenetic generalized least squares], $p > 0.05$). While diet specificity did not predict changes in total conotoxin diversity (PGLS, $p > 0.05$), we found a significant positive relationship between dietary breadth and total conotoxin diversity in both the full conotoxin dataset (PGLS, $p < 0.05$, Fig. 2-6) and the conotoxin dataset containing gene superfamilies that had > 80% capture success (PGLS, $p < 0.001$).

Discussion

Targeted sequencing and conotoxin discovery

Through targeted sequencing of conotoxins in cone snails, we demonstrate the potential to rapidly obtain venom sequences at high coverage (> 100X, Table S2-0) from species for which no venom information is available and without the need of RNA from the venom duct. This is remarkable, given that alignments in amino acid sequences between mature regions of a single

gene superfamily within a single individual can be incomprehensible (Fig. 2-1) due to the rapid evolution of the mature region (Duda & Palumbi 1999). Effective capture of conotoxin gene superfamilies was possible in part because conotoxin exons were often directly adjacent to conserved UTRs, which were targeted in the design (Table S2-3). While it has been recognized for decades that cone snails collectively harbor tens of thousands of biologically relevant proteins for fields such as molecular biology and pharmacology in their venoms (Olivera & Teichert 2007; Lewis 2009), traditional techniques for conotoxin sequencing (e.g., cDNA cloning) have barely begun to uncover and characterize the full breadth of conotoxin diversity. The targeted sequencing technique presented in this study adds an additional tool to increase the speed at which conotoxins are discovered. Specifically, this technique will allow the rapid toxin sequencing of genetic samples housed in museum collections, which often contain a large proportion of the species diversity for venomous taxonomic groups that have been amassed through several decades of intensive field expeditions. In addition, targeted sequencing approaches are not limited by the venom transcripts that are expressed at any one particular time point. As demonstrated here, on average, over half of the conotoxin genes available in the genome are not expressed in adult individuals (Table S2-5, Fig. S2-3). Therefore, sequencing conotoxins from genomic samples effectively doubles the number of conotoxin sequences that can be recovered.

Overall, the proportion of reads that mapped to our targeted sequences (mean = 48.8%, Table S2-0) is on par with studies that employed similar techniques in Anuran frogs (mean = 60.2% (Portik *et al.* 2016)) and Salamanders (mean = 18.21% (McCartney-Melstad *et al.* 2016)). The proportion of reads marked as duplicates were higher than previous studies (mean = 24.5% (McCartney-Melstad *et al.* 2016), mean = 17.5% (Portik *et al.* 2016)). The high duplication

levels in our dataset may have been a function of our small target size (~0.8Mb) or a sign that we over-amplified our post-hybridization product. To reduce the duplication levels in the future, we may reduce the number of post-hybridization PCR cycles. Given the high coverage on both the phylogenetic markers and conotoxin sequences (average cov > 95X, Table S2-0), future sequencing experiments should be able to multiplex more than 32 samples on a single lane.

Although most of the gene superfamilies had high capture success, some gene superfamilies performed poorly (Table S2-1). Variation in overall capture success can be attributed to several factors: first, a lack of diversity in bait sequences for a particular gene superfamily may have impeded effective capture. For example, we only had bait sequences designed from two species for the divMKFPLLFISL gene superfamily and we were unable to recover full sequences from any of the other species (Table S2-1). Second, the genetic organization of gene superfamilies my hinder capture success. For example, the mature toxin exon for the T gene superfamily is not readily recoverable because it is not adjacent to a conserved UTR that is discoverable through transcriptome sequencing (Table S2-1, S2-3, Fig. S2-1). Finally, conotoxin sequence properties may hinder capture, as it has been documented that high or low GC content values can depress capture efficiency statistics (Gnirke *et al.* 2009). To increase capture success of gene superfamilies in the future, we recommend including a large diversity of sequences from several species for gene superfamilies that had low capture success. In addition, bait sequences should be redesigned for gene superfamilies in which the prepro region or the mature region were not immediately adjacent to the conserved UTRs. We recovered intron sequences in this study that can be used in future bait designs to effectively recover the entire coding region because adjacent noncoding regions are often evolving at a

much slower rate than the coding region containing the prepro or mature region (Table S2-4, Fig. 2-4, S2-2).

Although targeted sequencing can increase the speed at which conotoxins are sequenced, we note several limitations with this approach. First, venom sequencing is limited to the sequences used in the bait design – only known venom gene families can be recovered. Therefore, approaches such as RNAseq are still necessary to identify undiscovered venom gene families. However, broad-scale discovery of venom gene superfamilies through transcriptome sequencing can be performed prior to applying target capture techniques (as done in this study) to obtain target sequences for most of the major venom components. Second, for some gene families, annotation of the exact mature toxin coding region may be difficult if the mature toxin is separated across multiple exons and if there is no closely related reference to accurately define the mature protein. Thus, expressed data is still necessary in some cases to study the mature toxin. Finally, depending on the level of sample multiplexing, targeted sequencing approaches can be more expensive than using RNAseq for venom discovery ($230.65 per RNAseq sample vs. $285.62 per targeted sequencing sample, Table S2-7). Therefore, the choice between two sequencing strategies will depend on the overall goal of the research project, as no one next generation sequencing method is suited for all research applications (Jones & Good 2016).

When compared to conotoxin sequences available on Genbank and ConoServer, we found that we were able to recover a larger proportion of previously sequenced conotoxins if species were explicitly targeted with the baits (Table S2-2). Although this comparison is biased by unequal conotoxin discovery effort across species, nearly half of previously sequenced conotoxins were not recovered in this study. We performed a coarse investigation of database conotoxins and determined potential reasons for why we were not able to recover certain

previously sequenced conotoxins. These reasons include: (a) the species in the database was misidentified, which was extensively documented in (Phuong *et al.* 2016), (b) the database conotoxin had no reliable reference in the literature, (c) the conotoxin was present in our species, but we could not recover it with the current bait design or the conotoxin was filtered during the bioinformatics processing of the data, and (d) the conotoxin was recoverable (i.e., high sequence similarity to bait sequences designed in this study), but the gene was not present in the genome. Future work integrating both population level RNAseq and targeted sequencing data may account for the large proportion of unrecovered conotoxins.

Conotoxin genetic architecture

Conotoxin exon length (range = 5bp – 444bp, Table S2-3) and the number of exons per gene (range = 1 – 6 exons, Table S2-3) are not unusual and fall within the range of variation seen in the genomes of other organisms (Deutsch & Long 1999; Sakharkar *et al.* 2004). In addition, the number of exons per gene within a gene superfamily align with results previous studies based on a relatively smaller number of sequences (Wu *et al.* 2013; Barghi *et al.* 2015b). For example, (Barghi *et al.* 2015b) found that the J superfamily consisted of a single exon and (Wu *et al.* 2013) found several gene superfamilies (i.e., I1, I2, M, etc.) consisting of 3 exons, which are identical to the results presented here.

A previous study suggested that rate variation among conotoxin functional regions (i.e., signal, prepro, mature) may be partially enabled by separation onto distinct exons in the genome (Olivera *et al.* 1999). Our results partially support this earlier hypothesis, given that the signal region and mature region were often confined to single exons (Fig. 2-3). However, we found that the prepro region was distributed across multiple exons, conflicting with earlier hypotheses.

Although not explicitly quantified, these results are also seen in earlier work examining the genomic architecture from conotoxin genes (Wu *et al.* 2013; Barghi *et al.* 2015b).

Conotoxin molecular evolution

A previous analysis of patterns of conotoxin divergence suggested that introns within contoxoin gene superfamilies were similar across species within a gene superfamily (Wu *et al.* 2013). Our results partially corroborate this suggestion, as the ratio of exon to noncoding divergence depended on what conotoxin region was encoded by the exon. Specifically, the exon containing the signal region was conserved and evolved much more slowly than adjacent noncoding regions (Table S2-4, Fig. 2-4, S2-2). This is similar to the pattern found in non-conotoxin exons (Fig. 2-4), indicative of purifying selection removing deleterious mutations from coding regions of critically important proteins (Hughes & Yeager 1997). In contrast, the exon diverges faster than the noncoding regions in all other exons, with the clearest difference between exon and noncoding region divergence seen in the exon containing most or all of the mature toxin region. This pattern is indicative of positive selection and is the same pattern is also seen in other genes under positive selection, such as PLA2 genes in snakes (Nakashima *et al.* 1993, 1995; Gibbs & Rossiter 2008) and fertilization genes in abalone (Metz *et al.* 1998). Overall, the patterns reported in this study aligns with previous work characterizing rate variation in snake venom proteins (Nakashima *et al.* 1993, 1995; Gibbs & Rossiter 2008). Although we did not use traditional methods to test for positive selection (e.g., MK tests, etc.), positive selection is well documented in cone snails (Duda & Remigio 2008; Duda 2008; Puillandre *et al.* 2010) and is therefore inferred to shape patterns of increased divergence in coding regions relative to noncoding regions. In addition, this genomic divergence pattern is consistent with a

recent analysis suggesting that the rapid evolution of conotoxin mature regions is due to positive selection (Roy 2016).

We found that on average, cone snails only express a fraction of the conotoxin genes available in their genomes (Table S2-5, Fig. S2-3), concurring with similar reports from smaller sets of gene superfamilies (Chang & Duda 2012, 2014; Barghi *et al.* 2015b). Several reasons could lead to this pattern. First, it is known that expression changes throughout an individual's lifetime in cone snails (Barghi *et al.* 2015a; Chang & Duda 2016), suggesting that the complement of genes expressed in the transcriptomes from Phuong *et al*. 2016 represent the adult conotoxins, and genes not discovered in the transcriptome but recovered from the genome are genes that are expressed in other life stages. Second, prey taxa available to cone snail species change with geography and so do the conspecifics it must compete against (Kohn 1959b, 1978; Kohn & Nybakken 1975; Duda & Lee 2009; Chang *et al.* 2015); therefore, different genes may be turned on or off in different geographic localities depending on the prey resources available and the composition of competitors in an individual's environment. Third, some of the conotoxin genes in the genome may not be expressed because they are no longer functional and have become psuedogenized. Finally, conotoxin gene expression may be regulated by defensive strategies against predators, as cone snails have been documented to release different conotoxins based on differing external stimuli (presentation of a prey item vs. physical agitation through poking, Dutertre *et al.* 2014). However, this hypothesis remains to be tested as there exists no ecological information to suggest that cone snails use their venom for defense – observations in the literature show that often, cone snails will hide in their shell or become completely devoured when confronted with an aggressor (Kohn 1959a). Future work comparing patterns of expression

relative to genomic availability will be able to disentangle the impact of conotoxin expression on changes to the venom phenotype.

We detected evidence for phylogenetic signal in the membership size of 25 gene superfamilies (Table S2-6, Fig. 2-5), suggesting that history plays a role in shaping gene gains and losses in cone snails. We note that uncovering evidence for phylogenetic signal in gene superfamily size does not imply that natural selection has not played a role in the evolution of venom as implied in (Gibbs *et al.* 2013). As described in Revell *et al.* (2008), evolutionary processes should not be inferred from patterns of phylogenetic signal because several contrasting models of trait evolution can lead to similar amounts of phylogenetic signal. Through CAFE v3.1 analyses, we also showed that venom composition is shaped by both net gains and losses in the entire genomic content of conotoxins (Fig. 2-5, S2-4). This result is in line with past studies showing that gene turnover is a fundamental characteristic shaping species' genic venom content (Duda & Palumbi 1999; Chang & Duda 2012; Dowell *et al.* 2016). We note a few limitations to the data used to examine gene turnover. First, total conotoxin gene diversity may be underestimated if large undiscovered gene superfamilies are present in specific clades of cone snails. Second, we only used one sample per species and technical variability in sequence capture and sample quality may have impacted our total conotoxin gene diversity estimates.

Diet and venom evolution

Why do cone snails vary in conotoxin gene superfamily size? Contrary to the popular assumption that particular gene superfamilies are associated with certain prey items (e.g, Kaas et al. 2010; Jin et al. 2013), prey families did not predict changes in gene superfamily size or total conotoxin diversity. This result aligns with a growing body of literature suggesting that the specific prey a

species feeds upon may not accurately predict conotoxin gene superfamily composition (Puillandre *et al.* 2012; Chang *et al.* 2015; Phuong *et al.* 2016). Although this study, along with previous studies, did not find a correlation between prey families and measures of conotoxin composition, this does not imply a lack of a relationship between diet specificity and conotoxin evolution for the following reasons. Characterizing the functional aspects of conotoxins is critical to understanding the relationship between diet specificity and conotoxin evolution because prey specialization exists at the level of protein function. However, it is known that conotoxin gene superfamilies are poor predictors of protein function and conotoxins with similar functions can convergently evolve in different gene superfamilies (Kaas *et al.* 2010; Puillandre *et al.* 2012; Robinson & Norton 2014). Therefore, if the functional aspects of cone snail venom repertoires are examined, a correlation between diet specificity and conotoxin composition may appear, such as in the case with cone snail insulins (Safavi-Hemami *et al.* 2016). We also acknowledge that our sampling of dietary diversity is not comprehensive (mollusc-hunting and fish-hunting species are undersampled) and this may limit our ability to fully examine diet specificity and conotoxin composition.

While dietary breadth also did not predict changes in gene superfamily size, we found a significant positive relationship with total conotoxin diversity (Fig. 2-6), aligning with several studies showing a coupling between dietary breadth and venom gene diversity in cone snails at nearly all biological scales of organization (Duda & Lee 2009; Chang *et al.* 2015; Chang & Duda 2016; Phuong *et al.* 2016). The correlation coefficient in this study between dietary breadth and total conotoxin diversity was weaker (r^2=0.25) than in our previous study (r^2=0.75, (Phuong *et al.* 2016)), possibly due to examining all of the conotoxin genes in the genome rather than just expressed transcripts. Conotoxin expression is known to change throughout an individual's

lifetime (Barghi *et al.* 2015a; Chang & Duda 2016) and these changes in expression have been shown to track dietary breadth (Chang & Duda 2016). Therefore, the weaker relationship could possibly be explained by using dietary breadth values measured from adult populations and examining its relationship to the total conotoxin repertoire an individual may draw from throughout its lifetime. In addition, we note that we were not able to distinguish between pseudogenes and functional genes, and this may contribute to the weaker relationship between total conotoxin diversity and dietary breadth. Another limitation of these analyses is that the number of individuals sampled per H index calculation was uneven and H index values may be biased towards species that had populations that were more extensively sampled.

The importance of dietary breadth shaping venom evolution remains underappreciated and untested in other venomous systems despite signals across several studies in cone snails. Future work examining the role of dietary breadth in shaping the evolution of venom in other venomous taxa will greatly advance our understanding between the interplay between diet and venom. The lack of a relationship between dietary breadth and changes in conotoxin gene superfamily size suggests that venom should be characterized as an aggregate trait rather than decomposed into individual parts to fully assess the impact of dietary breadth on conotoxin evolution. Further, studies have documented synergistic and complementary effects of conotoxins on prey species, suggesting that selection may act on the entire cocktail rather than individual components (Olivera 1997).

Conclusions

Through targeted sequencing of conotoxin genes, we provided comprehensive analyses of the gene structure of conotoxin gene superfamilies. In addition, we improved understanding of

conotoxin molecular evolution, including examining how positive selection impacts patterns of genomic divergence, how expression regulation of gene superfamilies varies across species, and how total conotoxin diversity changes through time. In addition, we found that variation in conotoxin diversity tracks changes in dietary breadth, suggesting that species with more generalist diets contain a greater number of conotoxin genes in their genome. Given that increased gene diversity is thought to confer an increased capacity for evolutionary change and species diversification (Kirschner & Gerhart 1998; Yang 2001; Malmstrøm *et al.* 2016), generalist species may speciate at faster rates than species with specialist diets. The targeted sequencing technique presented in this paper provides the necessary methodological advancement to rapidly sequence toxin genes across diverse clades of species, allowing tests of the relationship between ecology, toxin gene diversity, and higher order biodiversity patterns to be realized in future work.

Materials and Methods

Bait design and data collection

To recover markers for phylogenetic analyses, we targeted 886 protein coding genes representing 728,860bp. 482 of these genes were identified to be orthologous in Pulmonate gastropods (Teasdale *et al.* 2016) and we identified the remaining 404 genes using a reciprocal blast approach with 12 Conidae transcriptomes from (Phuong *et al.* 2016). For each gene, we chose the longest sequence from one of the 12 Conidae transcriptomes as the target sequence. For 421 of these genes, we used the entire length of the sequence as the target sequence, while for the remaining genes, we sliced the target sequences into smaller components based on

exon/intron boundaries inferred with EXONERATE v2.2.0 (Slater & Birney 2005) using the *Lottia gigantea* genome as our reference. EXONERATE v2.2.0 was run under default parameters and under the est2genome model. We chose to use the *L. gigantea* genome as our reference because it is highly contiguous (scaffold N50 = 1.87 Mb) and well annotated (Simakov *et al.* 2013), as a Conidae genome of comparable quality was not available at the time of the bait design. Often, exon/intron boundaries are conserved across fairly divergent taxa and can be used to define exon/intron boundaries (Bi *et al.* 2012); therefore, the *L. gigantea* genome was an appropriate choice to define exon/intron boundaries here. If exons were below 120bp in length (i.e., our desired bait length), but longer than 50bp, we generated chimeric target sequences by fusing immediately adjacent exons. We tiled bait sequences every 60bp across each target sequence. We note that the split bait design was due to an internal communication error and was not for a pre-specified purpose. For the conotoxin genes, we targeted 1147 conotoxins discovered from an early analysis of the 12 transcriptomes described in (Phuong *et al.* 2016). These sequences represent regions targeting 49 gene superfamilies and we included the full protein coding region plus 100bp of the 5' and 3' untranslated regions in our bait design when possible (Table S2-1). We tiled bait sequences every 40 bp across each conotoxin target sequence.

We obtained tissue samples preserved in 95% ethanol for 32 Conidae species through field collections in Australia and Indonesia and from the collections at the Australian Museum in Sydney, Australia (Table S2-0). We verified species identities using shell morphology and by sequencing CO1 prior to any next-generation sequencing laboratory work. We extracted genomic DNA from foot tissue using an EZNA Mollusc DNA kit (Omega Bio-Tek, Doraville, GA, USA) and used 1500ng of total DNA to prepare index-specific libraries following the Meyer and

Kircher (2010) protocol. To increase the probability of obtaining sequence information beyond the targeted regions, we performed 1X bead purifications after all enzymatic steps to remove fragments below 250bp. Fragment lengths of DNA samples ranged from 300-1000bp, with an average length of 450bp prior to hybridizations. We performed capture reactions following the MYbaits (v2) with the following specifications:

(1) We pooled 8 samples at a total concentration of 1.6μg DNA per capture reaction and allowed the baits to hybridize with the DNA for ~24 hours.

(2) We substituted the universal blockers provided with the MYbaits kit with xGEN blockers (Integrated DNA Technologies).

(3) We executed the 'stringent wash' protocol during the recovery of the captured targets.

After hybridization, we sequenced all 32 samples on a single lane HiSeq2000 lane with 100bp paired-end reads. Fragment lengths prior to hybridization were identical to the fragment length distributions post-hybridization that were submitted for sequencing.

Data assembly, processing, and filtration

We trimmed reads for quality and adapter contamination using Trimmomatic v0.33 (Bolger *et al.* 2014) under the following conditions: (a) we used the ILLUMINACLIP option to trim adapters with a seed mismatch threshold of 2, a palindrome clip threshold of 40, and a simple clip threshold of 15, (b) we performed quality trimming used the SLIDINGWINDOW option with a window size of 4 and a quality threshold of 20, (c) we removed reads below 36bp by setting the MINLEN option to 36, and (d) we removed leading and trailing bases under a quality threshold of 15. We merged reads using FLASH v1.2.8 (Magoč & Salzberg 2011) with a min overlap parameter of 5, a max overlap parameter of 100, and a mismatch ratio of 0.05. We generated

assemblies for each sample using SPAdes v3.1.0 (Bankevich *et al.* 2012) under default parameters. We reduced redundancy in the assemblies with cap3 (Huang & Madan 1999) under default parameters and cd-hit v4.6 (Li & Godzik 2006) using a sequence identity threshold of 99%.

For the phylogenetic markers, we used BLAST+ v2.2.31 (Altschup *et al.* 1990) with an evalue threshold of 1×10^{-10} and a word size value of 11 to associate assembled contigs with the target sequences. We used EXONERATE v2.2.0 under default parameters and used the est2genome model to redefine exon/intron boundaries because either (a) exon/intron boundaries were never denoted or (b) previously defined exons were actually composed of smaller exons. For each sample, we used bowtie2 v2.2.6 (Langmead & Salzberg 2012) using the very sensitive local alignment option and not allowing for discordant pair mapping (unexpected paired read orientation during mapping) to map reads to a reference containing only the contigs associated with the original target sequences. We marked duplicates using picard-tools v2.0.1 (http://broadinstitute.github.io/picard) using default parameters. We masked all positions that were below 5X coverage and removed the entire sequence if > 30% of the sequence was masked. To filter potential paralogous sequences in each species, we calculated heterozygosity (number of heterozygous sites/total number of sites) for each locus by identifying heterozygous positions using samtools v1.3 using default parameters and bcftools v1.3 (Li *et al.* 2009) using the call command and removed loci that were at least two standard deviations away from the mean heterozygosity.

For the conotoxin sequences, it is known that traditional assemblers perform poorly in reconstructing all potential conotoxin gene copies (Lavergne *et al.* 2015; Phuong *et al.* 2016). To ameliorate this issue, we reassembled conotoxin genes using the assembler PRICE v1.2 (Ruby *et*

al. 2013), which employs iterative mapping and extension using paired read information to build out contigs from initial seed sequences. To identify potential seed sequences for contig extension, we first mapped reads to the entire assembly outputted by SPAdes using bowtie2 v2.2.6 with previously stated parameters for each program. Then, we identified all sequence regions that locally aligned to any part of the original conotoxin target sequences via blastn v2.2.31 using previously stated parameters; these regions represented our preliminary seed sequences. We kept all preliminary seed sequences that were at least 100 bp (read length of samples in this study) and extended these seeds to 100bp if the alignable region was below that threshold. When extending these initial regions, we used Tandem Repeats Finder v4.09 (Benson 1999) to identify simple repeats and minimize the presence of these genomic elements in the preliminary seed sequences. We executed Tandem Repeats Finder v4.09 with default parameters except for the Minscore parameter, which we set at 12, and the Maxperiod parameter, which we set at 2. Often, only a subset of conotoxins are fully assembled with traditional assemblers (Phuong *et al.* 2016). However, when reads are mapped to these assemblies, unique conotoxin loci are similar enough to each other that relaxed mapping parameters will allow multiple copies to map to the contigs that were assembled. Therefore, multiple conotoxin copies will often map to each preliminary seed sequence. To generate seed sequences for all unique conotoxin loci, we used the python module pysam (https://github.com/pysam-developers/pysam) to pull all reads that mapped to regions of contigs representing the preliminary seed sequence and we reconstructed contigs from these reads using cd-hit v4.6 (percent identity = 98%) and cap3 (overlap percent identity cutoff =99%). From these reconstructed contigs, we used blastn v2.2.31 using previously described parameters to identify >100bp regions that matched the original preliminary seed and used these hits as our final seeds. We merged all final seeds that were

100% identical using cd-hit v4.6, mapped reads to these seeds using bowtie2 v2.2.6 with previously described parameters, and used PRICE v1.2 to re-assemble and extend each seed sequence under 5 minimum percentage identity (MPI) values (90%, 92%, 94%, 96%, 98%) with only the set of reads that mapped to that initial seed. Sequences were assembled using a minimum overlap length value of 40 and a threshold value of 20 for scaling overlap for contig-edge assemblies. A sequence was successfully reassembled if it shared ≥ 90% identity to the original seed sequence and if the final sequence was longer than the initial seed. For each seed sequence, we only retained the longest sequence out of the 5 MPI iterations for downstream filtering. We illustrated and described this workflow in Figure S2-5.

In order to generate a conotoxin reference database containing sequences that included both exons and adjacent noncoding regions, we used blastn v2.2.31 and EXONERATE v2.2.0 (using parameters described above) on species that were used in the bait design to (a) perform species-specific searches between our reassembled contigs and reference conotoxin sequences from (Phuong *et al.* 2016) and (b) define exon/intron boundaries on our reassembled contigs. We chose to constrain our blast searches to species-specific searches in order to improve accuracy and decrease the complexity of the data processing. In cases where a predicted terminal exon (i.e., the first or last exon of a conotoxin) was short (< 40 bp) and did not blast to any reassembled contig in our exon capture dataset, we replaced the reference conotoxin from Phuong *et al.* 2016 with the identical conotoxin containing the adjacent UTR regions to aid in the sequence searches. We generated conotoxins with UTR regions using the PRICE v1.2 algorithm as described above because the reference conotoxins from the final dataset in Phuong *et al*. 2016 did not include the UTR regions. We concatenated all annotated sequences into a single file to

create the final conotoxin reference, which consisted of sequences with exons and introns defined from all species that were initially used in the bait design.

With the final conotoxin reference, we used blastn v2.2.31 to associate contigs with this reference in every species and used EXONERATE v2.2.0, blastx v2.2.31, and tblastn v2.2.31 to define exon/intron boundaries. For the BLAST+ v2.2.31 software, we used parameters previously described above and for EXONERATE v2.2.0, we used the protein2genome model and an alignment score threshold value of 50. When exon/intron boundaries could not be defined through these methods, we guessed the boundaries by aligning the assembled contig to the reference sequence using MAFFT v7.305b (Katoh *et al.* 2005) and denoted the boundaries across the region of overlap with the exon in the reference sequence. For each sample, we mapped reads using bowtie2 v2.2.6, accounted for duplicates using picard-tools v2.0.1, and retained sequences that had at least 10x coverage across the exons defined within each contig. We increased mapping stringency for bowtie2; in addition to using the parameters previously described, we modified the alignment score threshold (--score-min L,70,1). We masked regions below 10x coverage and used cd-hit v4.6 to merge contigs that were ≥ 98% similar, generating our final conotoxin gene models. Finally, we used HapCUT v0.7 (Bansal & Bafna 2008) under default parameters to generate all unique haplotypes across coding regions. We note that our final conotoxin dataset consists of exon fragments (conotoxin exons with any assembled adjacent noncoding regions), rather than full conotoxin genes. As described in (Wu *et al.* 2013), conotoxin introns can often be several kilobases in length, which is much longer than the average insert size of our sequencing experiment (~350bp).

To assess the overall effectiveness of our targeted sequencing experiment, we calculated (a) percent of reads aligned to intended targets, (b) percent duplicates, and (c) average coverage

across targeted regions. To assess capture success of conotoxins, we divided the number of conotoxin transcripts successfully recovered in the exon capture dataset by the number of conotoxin transcripts discovered in Phuong *et al*. 2016 for each gene superfamily. We defined a conotoxin transcript to be successfully sequenced if > 80% of the transcript was recovered in the exon capture experiment with > 95% identity. To assess the ability of targeted sequencing to recover gene superfamily sequences from species that were not explicitly targeted in the bait sequences, we calculated the number of previously sequenced conotoxins that match contigs recovered in our dataset. We gathered conotoxin sequences from Genbank and ConoServer (Kaas *et al.* 2010) with species names that correspond to species in this study, merged sequences with 98% identity using cd-hit v4.6, and used blastn v2.2.31 under previously stated parameters to perform species-specific searches. We defined a conotoxin as successfully sequenced if the hypervariable mature toxin coding region aligned with ≥ 95% identity to a sequence in our dataset.

Conotoxin genetic architecture

To characterize conotoxin genetic architecture, we quantified the following values: (a) the number of exons comprising a conotoxin transcript, (b) average length of each exon, and (c) the size range of exon length. We also determined the proportion of terminal exons adjacent to UTRs by conducting sequence searches (via blastn v2.2.31 under previously stated parameters) between contigs containing terminal exons against a database of conotoxins from Phuong *et al.* 2016 that were reassembled to contain the UTRs. To determine how traditional conotoxin precursor peptide regions are distributed among exons, we calculated the average proportion of each conotoxin region found on each exon in every gene superfamily. We defined regions of the

Phuong *et al.* (2016) transcripts using ConoPrec (Kaas *et al.* 2012). We restricted these conotoxin genetic structure analyses to transcripts from Phuong *et al*. (2016) that were successfully recovered in the exon capture dataset and that were retained after clustering with cd-hit v4.6 (similarity threshold = 98%). We performed clustering to avoid over-inflating estimates because unique transcripts from Phuong *et al.* (2016) may have originated from the same gene.

Conotoxin molecular evolution

We first classified all exons into conotoxin precursor peptide regions. For species with transcriptome data, we first labeled exons as either the signal region or the mature region by identifying the exons containing the largest proportion of these separate regions. Then, exons between the signal and mature exon were labeled as the prepro exon(s) and exons after the mature region were labeled as the post exons. Gene superfamilies containing only a single exon were denoted as such. We then used blastn v2.2.31 under previously stated parameters to classify sequences without transcriptome data into these conotoxin precursor peptide regions. For each functional category within each gene superfamily, we calculated uncorrected pairwise distances between all possible pairwise comparisons. To avoid spurious alignments, we only considered comparisons within clusters that clustered with cd-hit v4.6 at an 80% threshold and we excluded comparisons if (a) the alignment length of the two exons was 20% greater than the longer exon, (b) the align-able nocoding region was below 50bp, or (c) the shorter exon's length was less than 70% of the length of the longer exon. We calculated separate pairwise distance estimates for regions of the alignment that contained the exon and regions of the alignment that contained the noncoding DNA. We excluded region-labeled exons within superfamilies from this analysis that had less than 50 possible comparisons. For comparison, we also calculated pairwise distances

between exons and noncoding regions across our phylogenetic markers which represent non-conotoxin exons, filtered with similar criteria described above.

Conotoxin expression

To characterize variation in expression patterns among species per gene superfamily, we calculated the number of conotoxin genes expressed in species with transcriptome data divided by the number of genes available in the genome. We restricted these analyses to instances where 90% of the unique mature toxins were recovered for a gene superfamily within a species. To estimate gene superfamily size, we used the exon labeled as containing most or all of the mature region. We used the mature region because it is unique between sequences discovered from the transcriptome. We could not, for example, use a signal region sequence, as they may map to several sequences from the target capture data given that they are highly conserved within a gene superfamily. We defined a conotoxin gene as expressed if we retained a blast hit with 95% identity to a unique mature toxin sequence found in the transcriptome.

Gene superfamily size change estimation

To compare and contrast gene superfamily size changes between species, we used the total number of exons containing most or all of the signal region as our estimate of gene superfamily size because exons containing the signal regions are relatively conserved across species (Robinson & Norton 2014) and thus have the highest confidence of being recovered through exon capture techniques. To quantify and test the amount of phylogenetic signal in conotoxin gene diversity, we estimated Pagel's lambda (Pagel 1997) in the R package phytools (Revell 2012). Lambda values range from 0 (phylogenetic independence) to 1 (phylogenetic

signal) and p-values < 0.05 represent significant departure from a model of random trait distribution across species with respect to phylogeny. To estimate conotoxin gene superfamily gains and losses along every branch, we used the program CAFEv3.1 (Han *et al.* 2013), which uses a stochastic gene birth-death process to model the evolution of gene family size. As input, we used a time-calibrated phylogeny and estimates of gene superfamily size for 37 superfamilies that were present in at least 2 taxa. To estimate a time-calibrated phylogeny, we aligned loci that had at least 26 species using MAFFT v7.305b under default parameters and used a concatenated alignment to build a phylogeny in RAxML under a GTRGAMMA model of sequence evolution (Stamatakis 2006). We performed a maximum likelihood search of the phylogenetic tree and the rapid bootstrapping analyses with 100 replicates. We time-calibrated the phylogeny with the program r8s (Sanderson 2003) under default parameters and using two previous fossil calibrations described in cone snails (Duda Jr. *et al.* 2001). We excluded *Californiconus californicus* from the CAFE v3.1 analysis due to optimization failures.

Diet and conotoxin gene superfamily size evolution

To examine the role of diet specificity and dietary breadth on conotoxin gene superfamily size evolution and total conotoxin diversity, we retrieved prey information from the literature (Kohn 1959a; b, 1966, 1968, 1978, 1981, 2001, 2003, 2015; Marsh 1971; Kohn & Nybakken 1975; Taylor 1978, 1984, 1986; Taylor & Reid 1984; Nybakken & Perron 1988; Kohn & Almasi 1993; Reichelt & Kohn 1995; Kohn *et al.* 2005; Nybakken 2009; Chang *et al.* 2015). For diet specificity, we classified prey items into one of 27 different prey categories (Table S2-8). For dietary breadth, we retrieved estimates of the Shannon's diversity index (H') or calculated it if there were at least 5 prey items classified to genus with a unique species identifier. When

multiple H' values were obtained for a species, we averaged them because species will consume different sets of prey taxa depending on geography. Raw data are available in Table S2-8. To examine the impact of prey group and dietary breadth on changes in gene superfamily size, we used D-PGLS (distance-based phylogenetic generalized least squares), a phylogenetic regression method capable of assessing patterns in high-dimensional datasets (Adams 2014). To reduce redundancy among prey group variables, we removed variables that were 80% correlated with each other using the redun function in the R package Hmisc. We used the total number of exons containing the signal region as our estimate of gene superfamily size. To convert gene superfamily size counts into continuous variables, we transformed the data into chi-squared distances between species in 'conotoxin gene superfamily space' using the deostand function in the R package vegan (Oksanen *et al.* 2016). To examine the impact of diet specificity and dietary breadth on total conotoxin diversity, we used a PGLS analysis implemented in the caper package within R (Orme 2013). We ln-transformed total conotoxin diversity for the PGLS analysis. We performed all analyses with the full dataset and a subset of the data that only included gene superfamilies with > 80% capture success. We did not perform any analyses with *C. californicus* because it is regarded as an outlier species amongst the cone snails due to its atypical diet and its deep relationship with the rest of Conidae (Kohn 1966; Puillandre *et al.* 2014a).

Acknowledgements

We thank DST Hariyanto, MBAP Putra, MKAA Putra, and the staff at the Indonesian Biodiversity Research Center in Denpasar, Bali for assistance in the field in Indonesia; F Criscione, F Köhler, A Moussalli, A Hogget, and L Vail for logistical assistance for fieldwork at the Lizard Island Research Station in Australia; M Reed, A Hallan, and J Waterhouse for access

to specimens at the Australian Museum in Sydney, Australia; J Finn, M Mackenzie, and M Winterhoff for access to specimens at the Museum Victoria in Melbourne, Australia; WF Gilly for access to the *C. californicus* specimen, K Bi, L Smith, and A Moussalli for advice on bait design; A. Devault and MYcroarray for great service and technical support for bait synthesis; EM McCartney-Melstad, the B Shaffer lab, and the Evolutionary Genetics Lab at UC Berkeley for laboratory support; A. Kohn for providing cone snail diet data from several publications; MCW Lim and J Chang for thoughtful advice and discussions; N Puillandre, S Prost, S Robinson and six anonymous reviewers for insightful comments on earlier versions of this manuscript. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575. This work was supported by a Grants-in-Aid of research from Sigma Xi, a Grants-in-Aid of Research from the Society for Integrative and Comparative Biology, a research grant from the Society of Systematic Biologists, a Student Research Award from the American Society of Naturalists, a National Science Foundation Graduate Research Opportunities Worldwide to Australia, the Lerner Gray Fund for Marine Research from the American Museum of Natural History, research grants from the Department of Ecology and Evolutionary Biology at UCLA, a small award from the B Shaffer Lab, a National Science Foundation Graduate Research Fellowship, an Edwin W. Pauley fellowship, and a Chateaubriand fellowship awarded to MAP. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. We thank the Indonesian Ministry of State for Research and Technology (RISTEK, permit number 277/SIP/FRP/SM/VIII/ 2013) for providing permission to conduct fieldwork in Bali. The C. californicus specimen was collected

under a California Department of Fish and Wildlife collecting permit granted to WF Gilly (SC-6426).

List of Figures

(blue = signal region, black= prepro region, orange = mature region, light blue = single exon, grey = non-conotoxin exon). x = y line is shown.

Figure 2-5. Diet and conotoxin evolution in a phylogenetic context. Time-calibrated maximum likelihood phylogeny of 32 Conidae species generated from concatenated alignment of 4441 exons. Phylogeny is rooted with *Californiconus californicus*. Branches are colored based on net gains or losses in total conotoxin diversity based on CAFE v3.1 analyses. Recognized subgenera are alternately colored pink. Total conotoxin diversity, the number of expressed precursors for species with transcriptomes, size estimates for commonly studied gene superfamilies, and dietary breadth displayed next to tip names. Recorded observations of each species preying on each of the 27 represented prey families shown in the matrix adjacent to the phylogeny, with cells colored based on whether or not a species has been observed to feed on that prey family (grey = no, blue = yes). Phylum level classifications are shown at the top of the diet matrix and family level classifications are shown at the bottom of the diet matrix.

Figure 2-6. Scatterplot of total conotoxin gene diversity and dietary breadth. Each point represents a unique species. Graphs are labeled with a regression line and Pearson's correlation coefficient generated from a Phylogenetic Generalized Least Squares (PGLS) analysis. * denotes significant correlation from the PGLS analysis.

Figures

Figure 2-1.

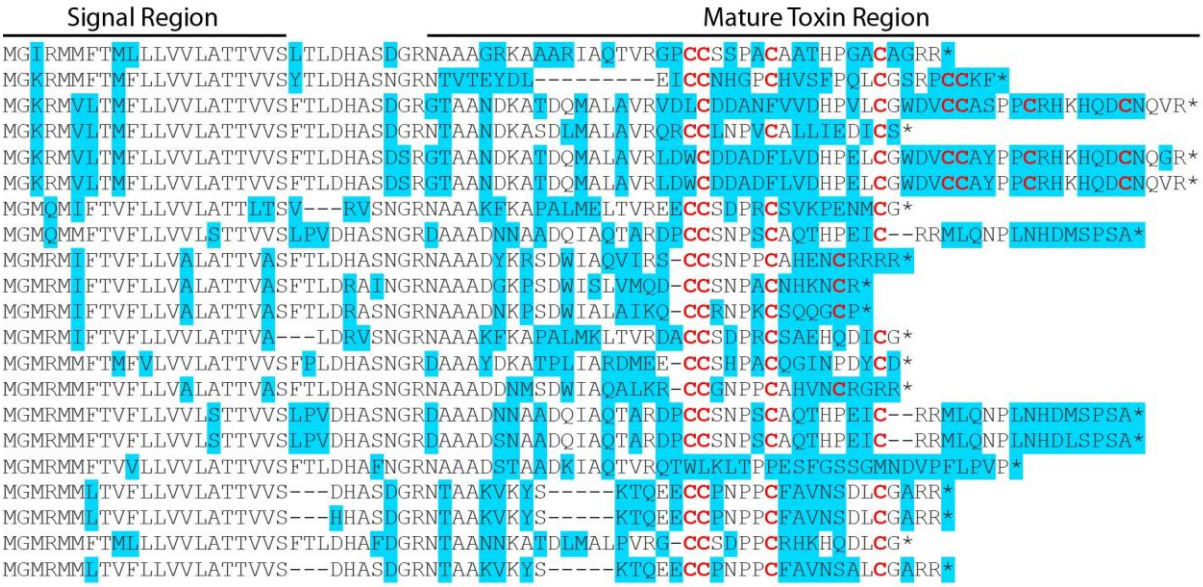


Figure 2-2.

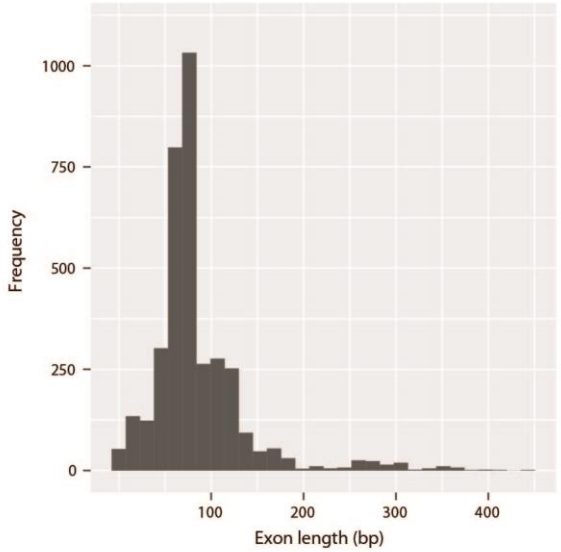


Figure 2-3.

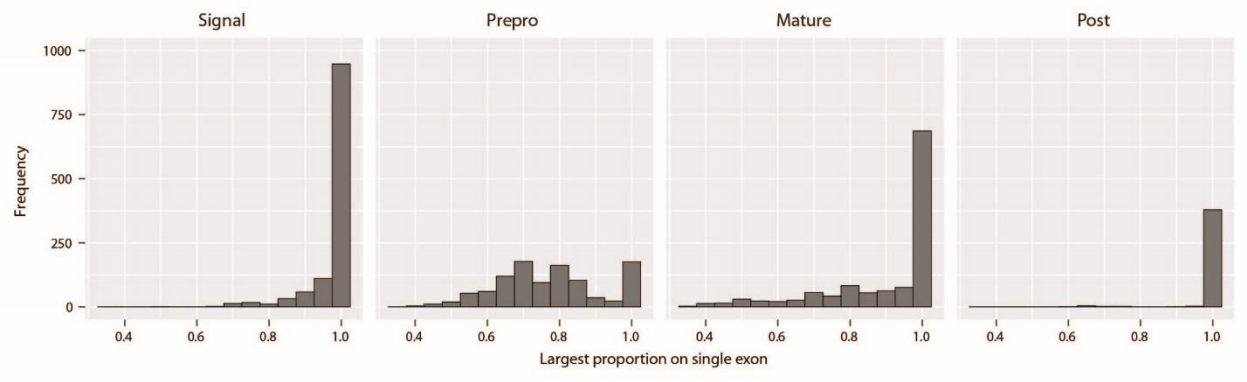


Figure 2-4.

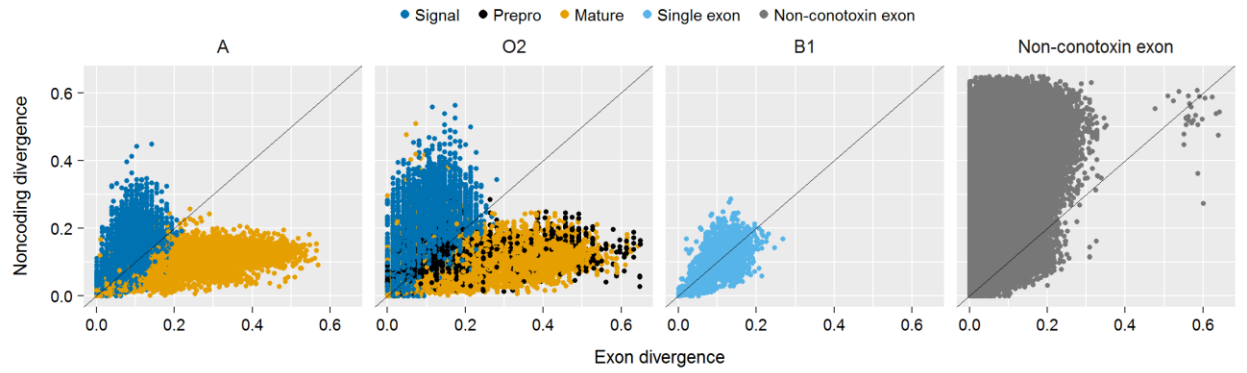


Figure 2-5.

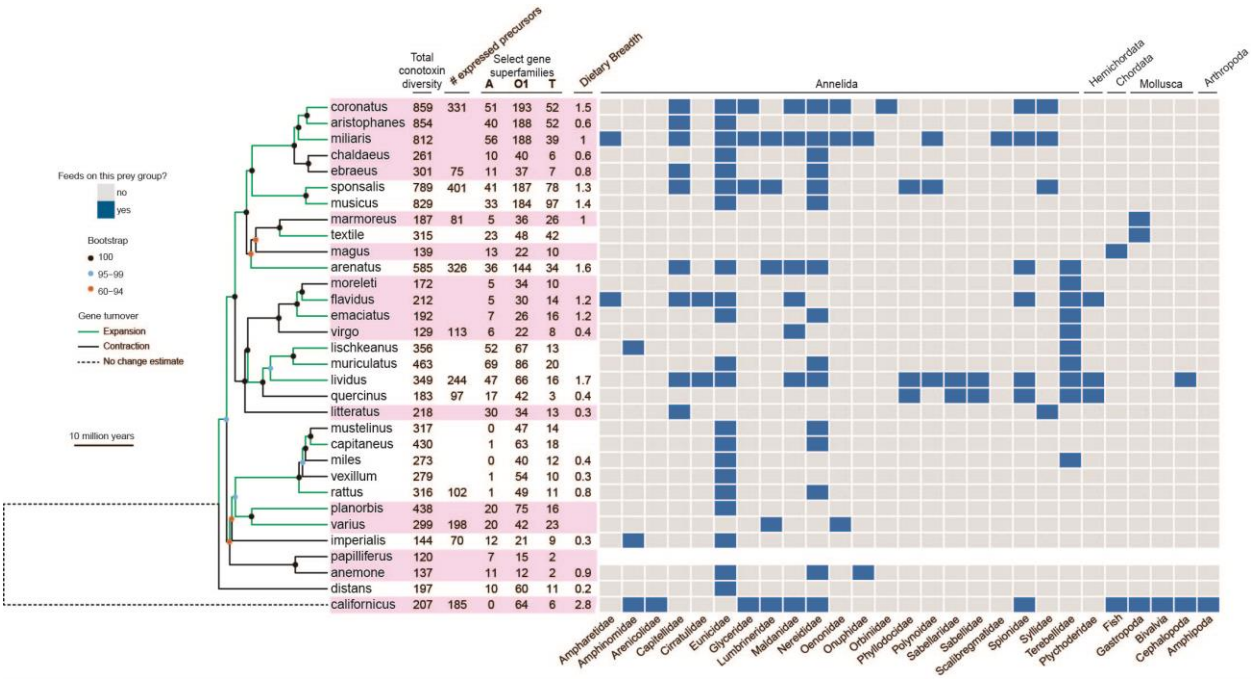
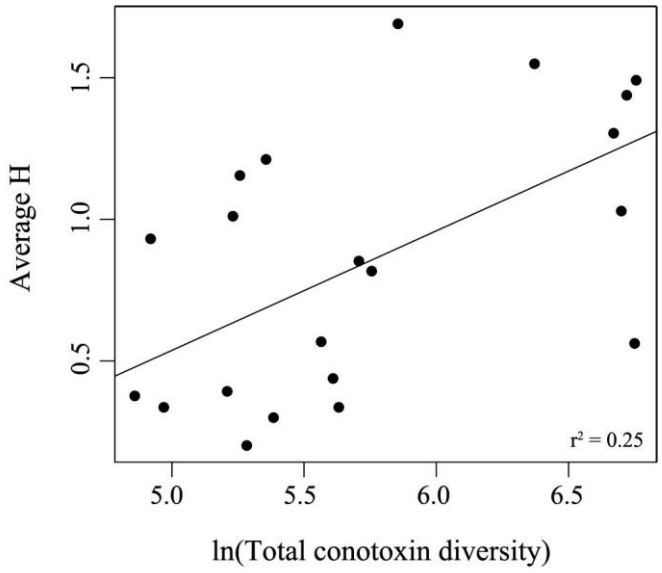


Figure 2-6.



List of supplementary material

Figure S2-1. Conotoxin diversity per gene family per conotoxin functional region in a phylogenetic context.

Figure S2-2. Scatterplots of all gene superfamilies showing the relationship between exon divergence and noncoding divergence. Divergence was estimated by calculating uncorrected pairwise distances. Each point on the graph represents a unique pairwise comparison and points are colored by conotoxin functional region (blue = signal region, black= prepro region, orange = mature region, light blue = single exon, grey = non-conotoxin exon). x = y line is shown.

Figure S2-3. Conotoxin expression regulation. % genes expressed relative to the total number of genes in the genome. Percentages range from 100% in dark blue to 0% in light grey. A white cell indicates no data. Only species with transcriptome data are represented. In addition, values were only calculated if 90% of the transcripts per gene superfamily per species were recovered in the target capture data. Pruned time-calibrated maximum likelihood tree is shown and tree is rooted with *C. californicus*.

Figure S2-4. CAFE v3.1 net gene gains and losses plotted across a time-calibrated phylogeny of cone snails. Values on branches represent the number of conotoxin genes gained or loss along that branch. Total conotoxin diversity listed next to species names.

Figure S2-5. Bioinformatic workflow for conotoxin assembly. We outline the overall bioinformatic work flow and provide complementary figures to illustrate the bioformatic processing of the data.

Table S2-0. Sample information and exon capture statistics.

Table S2-1. Conotoxin sequences targetted and recovered. In each cell, the first value represents the number of sequences targetted by probes, the second value (left of the forward slash) represents the number of sequences matching at least 80% of sequence expressed in the transcriptome (with a 90% similarity threshold), and the last value (right of forward slash) represents the number of unique transcripts discovered in the Phuong et al. (2016) transcriptomes. Capture success represents the number of sequences recoverd (summed 2nd values across a gene superfamily) divided by the number of known transcripts (summed 3rd values across a gene superfamily).

Table S2-2. Capture success of previously sequenced conotoxins. # sequences on Genbank/ConoServer represent values after merging with cd-hit (percent identity threshold = 98%).

Table S2-3. Conotoxin genomic architecture listed by gene superfamily. First six columns show (1) gene superfamily, (2) the number of exons composing a transcript, (3) the number of transcripts with that particular exon configuration, (4) exon number, (5) avg. exon length, and (6) minimum and maximum exon length. First four percentages (% Signal, % Prepro, % Mature, %

Post) represent the average proportion of each conotoxin region found on each exon. Last two columns represent the proportion of transcripts composed of terminal exons that are adjacent to either a 5' or 3' UTR (untranslated region).

Table S2-4. Relative exon divergence to adjacent region divergence by gene superfamily.

Table S2-5. Proportion of conotoxin genes expressed per gene superfamily per species. In each cell, the first value (left of the forward slash) represents the number of genes expressed in the transcriptome, the second value (right of the forward slash) represents the total number of genes in the genome, and the proportion of genes expessed is noted in the parentheses.

Table S2-6. Phylogenetic signal of conotoxin gene superfamilies. P-value < 0.05 are bolded indicates significant difference from model with no phylogenetic signal.

Table S2-7. Estimated costs for transcriptome and targeted sequencing experiments. These cost estimates only include main expenses and does not include consumables, such as gloves, pipette filter tips, eppendorf tubes, etc., which are expected to increase in cost with increases in sample size. Prices are based at the time the materials were purchased. Price of sequencing based on UC rates provided by the Vincent J. Coates Genomics Sequencing Laboratory.

Table S2-8. Raw diet data

Chapter 3: Speciation rates are decoupled from venom gene diversity in cone snails

Abstract

Understanding why some groups of organisms are more diverse than others is a central goal in macroevolution. Evolvability, or lineages' intrinsic capacity for evolutionary change, is thought to influence disparities in species diversity across taxa. Over macroevolutionary time scales, clades that exhibit high evolvability are expected to have higher speciation rates. Cone snails (family: Conidae, >900 spp.) provide a unique opportunity to test this prediction because their venom genes can be used to characterize differences in evolvability between clades. Cone snails are carnivorous, use prey-specific venom (conotoxins) to capture prey, and the genes that encode venom are known and diversify through gene duplication. Theory predicts that higher gene diversity confers a greater potential to generate novel phenotypes for specialization and adaptation. Therefore, if conotoxin gene diversity gives rise to varying levels of evolvability, conotoxin gene diversity should be coupled with macroevolutionary speciation rates. We applied exon capture techniques to recover phylogenetic markers and conotoxin loci across 314 species, the largest venom discovery effort in a single study. We paired a reconstructed timetree using 13 fossil calibrations with species-specific estimates of conotoxin gene diversity and used trait-dependent diversification methods (BiSSE, STRAPP) to test the impact of evolvability on diversification patterns. Surprisingly, we found no relationship between conotoxin gene diversity and speciation rates, suggesting that venom evolution is not the rate-limiting factor controlling diversification dynamics in Conidae. Comparative analyses detected a shift in diversification rates leading to the subgenus *Lautoconus*, a previously documented radiation of mainly Cape Verde island endemics thought to be the result of geographic isolation and the presence of non-dispersing larvae. Our results suggest that the rapid evolution of Conidae venom may cause other

factors to become more critical to diversification, such as ecological opportunity or traits that promote isolation among lineages.

Introduction

Why are some taxa more diverse than others? Species richness and phenotypic diversity are not distributed evenly across the tree of life (Rabosky *et al.* 2013). For example, there exists over 10,000 species of birds, but their closest relatives (crocodiles and alligators) comprise only of 23 species. Differences in evolvability, or lineages' intrinsic capacity to adapt and diversify, is one reason commonly used to explain these disparities (Wagner & Altenberg 1996; Yang 2001; Jones *et al.* 2007; Pigliucci 2008; Losos 2010). Evolvability is thought to be determined by the underlying genetic architecture of organisms – some genomes of organisms have a greater propensity to generate variation that may be adaptive in the future (Wagner & Altenberg 1996; Jones *et al.* 2007; Pigliucci 2008). For example, gene duplication increases evolvability – the copied gene is free from the selective pressures of the original gene (Crow & Wagner 2006). Mutation, selection, and drift can act on the copied gene, facilitating the possibility of new phenotypes to arise; this shapes the extent that taxa can diversify and exploit resources (Crow & Wagner 2006). Over long evolutionary time scales, clades that exhibit higher evolvability are predicted to have increased species richness and diversification rates (Yang 2001).

Despite the ubiquity of this concept in macroevolutionary theory, few studies explicitly test these predictions; this is possibly due to the difficulty of identifying genes responsible for phenotype (Hoekstra & Coyne 2007). Past studies that have attempted to test the impact of evolvability on diversification have produced mixed results (Santini *et al.* 2009; Soltis *et al.* 2009; Mayrose *et al.* 2011; Rabosky *et al.* 2013; Zhan *et al.* 2014; Tank *et al.* 2015; Malmstrøm

et al. 2016). For example, whole genome duplication events, which are hypothesized to increase the genomic potential of organisms, have been documented to increase (Santini *et al.* 2009; Soltis *et al.* 2009; Tank *et al.* 2015), decrease (Mayrose *et al.* 2011), and have no impact (Zhan *et al.* 2014) on the long-term evolutionary success of clades. In another case, a positive correlation between evolvability and speciation rates exist when measuring evolvability through morphological proxies (Rabosky *et al.* 2013). One limitation of past research on this hypothesis is the inability to tie genomic changes with ecological factors driving diversification patterns (Robertson *et al.* 2017). Although gene duplication and whole genome duplication events can increase the evolutionary capacity of organisms, genes that are ecologically relevant for adaptation may not be readily available for selection to drive divergence.

Here, we study the relationship between evolvability and diversification in cone snails (family, Conidae), a diverse group (> 900 spp.) of predatory marine gastropods. These snails feed on either worms, molluscs, or fish by paralyzing their prey with a cocktail of venomous neurotoxins (conotoxins, Duda & Palumbi 1999). Cone snail provides a unique opportunity to test predictions of evolvability and diversification for the following reasons: first, cone snail species share an ecologically relevant trait, venom. Conidae species are globally distributed in tropical and subtropical regions, where >30 species can co-occur within the same habitat (Kohn 2001). High numbers of species hypothesized to be able to co-occur because species have diversified to specialize on different prey using prey-specific conotoxins (Duda & Palumbi 1999). Second, venom genes are known and diversify through gene duplication (Duda & Palumbi 2000; Kaas *et al.* 2010, 2012; Chang & Duda 2012). Diet specialization is thought to be enabled by the rapid evolution of the genes that underlie conotoxins – estimated rates of gene duplication and nonsynonymous substitutions rates for conotoxin genes are the highest across

metazoans (Duda & Palumbi 2000; Chang & Duda 2012). Therefore, conotoxin genes provide a natural way to characterize differences in evolvability between clades.

We employ a sequence capture technique previously used in cone snails (Phuong & Mahardika 2017) to recover phylogenetic markers and conotoxin genes from 314 described species. We use the phylogenetic markers to reconstruct a time-calibrated phylogeny and perform trait-dependent diversification analyses to test the impact of evolvability on diversification patterns. We predict that clades with a greater number of conotoxin gene copies should have higher speciation rates.

Methods

Bait design

We used a targeted sequencing approach to recover markers for phylogenetic inference and obtain an estimate of conotoxin gene diversity from Conidae species. For the phylogenetic markers, we identified loci using a previous Conidae targeted sequencing dataset (Phuong & Mahardika 2017) and the Conidae transcriptome data from (Phuong *et al.* 2016). In the Conidae targeted sequencing dataset, the authors generated a phylogeny using 5883 loci across 32 species (Phuong & Mahardika 2017). For our sequencing experiment, we only retained loci that were >180bp and were present in at least 26 out of 32 taxa with at least 10X coverage. We chose to only include longer loci to increase confidence in identifying orthologous fragments in other Conidae species. To identify additional phylogenetic markers from the transcriptome data (Phuong *et al.* 2016), which consisted of venom duct transcriptomes from 12 species, we performed the following:

(1) identified reciprocal best blast hits between the assembled transcriptome and the *Lottia gigantea* protein reference (Simakov *et al.* 2013) using BLAST+ v2.2.31 (evalue = 1e-10). We also considered fragments that had their best hit to the protein reference, but to a non-overlapping portion (<20% overlapping).

(2) mapped reads using bowtie2 v2.2.7 (Langmead & Salzberg 2012)

(3) removed duplicates using picard-tools v.2.1.1 (http://broadinstitute.github.io/picard)

(4) fixed assembly errors by calling single nucleotide polymorphisms (SNPs) using samtools v1.3 and bcftools v1.3 (Li *et al.* 2009)

(5) aligned sequences per locus using mafft v7.222 (Katoh *et al.* 2005)

(6) calculated uncorrected pairwise distances within each locus for all possible pairwise comparisons

(7) removed sequences if the uncorrected pairwise distance was greater than the 90th percentile for those pair of species

(8) denoted exon boundaries by comparing the transcriptome sequences to the *Lottia gigantea* genome reference (Simakov *et al.* 2013), retaining exons >180bp

For all retained phylogenetic markers, we also performed the following: (1) we generated an ancestral sequence using FastML v3.1 (Ashkenazy *et al.* 2012) between a *Californiconus californicus* sequence and another Conidae sequence that had the highest amount of overlap with the *C. californicus* sequence (we generated these ancestral sequences to decrease the genetic distances between the target sequence and the orthologous sequence from any Conidae species), (2) removed sequences that had a GC content < 30% or > 70% because extreme GC contents can reduce capture efficiency (Bi *et al.* 2012), (3) removed loci that contained repeats identified through the RepeatMasker v4.0.6 web server (Smit *et al.* 2015), and (4) performed a self-blast

with the target sequences via blastn v2.2.31 (evalue = 1e-10) and removed loci that did not blast to itself with sequence identity >90%. The final set of target loci for phylogenetic inference included 1749 loci, with a total length of 470,435 bp.

To recover conotoxin loci, we targeted sequences generated from both the previous targeted sequencing dataset (Phuong & Mahardika 2017) and the transcriptome dataset (Phuong *et al.* 2016). For conotoxin sequences discovered from the targeted sequencing dataset (Phuong & Mahardika 2017), we performed the following to generate our target sequences: (1) we trimmed each sequence to only retain the coding region and included 100bp flanking the exon, (2) merged sequences using cd-hit v4.6.4 (Li & Godzik 2006) at 95% sequence similarity to reduce redundancy among conotoxin loci (3) masked repeats using the RepeatMasker v4.0.6 web server (Smit *et al.* 2015), and (4) retained loci >120bp to ensure that the locus was longer than our desired bait sequence length. We concatenated all sequences below 120bp to create a single, chimeric sequence for capture. The final set of target sequences from the previous targeted sequencing dataset consisted of 12,652 unique loci totaling 3,113,904 bp and a single concatenated sequence representing 351 merged loci with a total length of 37,936 bp. We also targeted conotoxin loci from the transcriptomes described in (Phuong *et al.* 2016) to obtain conotoxin loci from gene superfamilies that were not targeted in (Phuong & Mahardika 2017) or performed poorly. We performed the following to generate a set of conotoxin loci from the transcriptome data: (1) we trimmed sequences from (Phuong *et al.* 2016) to only include the coding region and 100bp of the untranslated regions (UTRs), (2) merged sequences using cd-hit v4.6.4 (Li & Godzik 2006) at 97% sequence similarity to reduce redundancy among conotoxin loci, and (3) masked repeats using the RepeatMasker v4.0.6 web server (Smit *et al.* 2015). This filtered dataset contained 395 conotoxin loci with a total length of 171,317 bp.

We submitted the following datasets to MYcroarray (Ann Arbor, Michigan, USA) for bait synthesis: (1) 1749 loci for phylogenetic inference, (2) 12652 conotoxin loci using data from (Phuong & Mahardika 2017), (3) a single concatenated sequence using data from (Phuong & Mahardika 2017), and (4) 395 additional conotoxin loci using transcriptome data from (Phuong *et al.* 2016). We chose to synthesize a MYbaits-3 kit, which included 60,000 bait sequences to accommodate all the targeted loci. Because our aim was to recover sequences from species throughout Conidae, each bait sequence was 120bp in length, which increases the efficiency of recovering divergent fragments. We used a 2X tiling density strategy (a new probe every 60bp) across the sequences from datasets (1) and (2) and used a 4X tiling density strategy (a new probe every 30bp) across datasets (3) and (4). We chose to increase the tiling density for datasets (3) and (4) because the boundaries between exons were not denoted and we wanted to ensure effective capture of the conotoxin loci. The set of probe sequences will be made available on DRYAD following publication.

Genetic samples, library preparation, hybridization, and sequencing

We performed the targeted sequencing experiment across 362 samples representing both described Conidae species and unique lineages/potential new species identified during routine species verification using the mitochondrial locus (results not shown), CO1 (Table S3-1, Folmer *et al.* 1994). We also sequenced *Bathyoma* sp. as an outgroup based on a recent molecular phylogeny of the Conoideans, a clade of gastropods that includes Conidae (Table S3-1, Puillandre *et al.* 2011). We obtained these genetic samples from two field expeditions in Indonesia and Australia and from five museum collections (Table S3-1). We extracted DNA from tissue using the EZNA Mollusc DNA kit (Omega Bio-Tek, Doraville, GA, USA). There

was slight variation in tissue preservation strategy among samples, with most tissues preserved directly in 95% ethanol (Table S3-1). For 10 samples, tissue was not available but DNA was available from a previous extraction. For these samples, we ran the DNA through the EZNA Mollusc DNA kit to purify the DNA prior to library preparation. We extracted a minimum of 2000 ng per sample prior to library preparation, when possible. We sheared DNA using a Biorupter UCD-200 (Diagenode) when necessary and used a 1X bead purification protocol to ensure that the DNA fragments per sample ranged from 250-600bp, centered on ~350bp. We aimed to generate libraries with longer fragment sizes to ensure that we could recover exons containing the mature toxin region, which are often only recoverable because they are flanking conserved regions that are targeted by our bait design (Phuong & Mahardika 2017).

We prepared libraries for following the (Meyer & Kircher 2010) protocol with the following modifications: (1) we started library preparation with at least 2000ng, rather than the 500ng suggested by the protocol to increase downstream capture efficiency, (2) we performed 1X bead clean-up for all enzymatic reactions and (3) generated dual-indexed libraries by incorporating adapters with unique 7bp barcodes. We were able to re-use libraries for the 32 species sequenced in (Phuong & Mahardika 2017) and incorporated new indexes for these samples.

We generated equimolar pools of 8 samples and hybridized probes with 2000ng of the pooled DNA for ~24 hours. We substituted the adapter blocking oligonucleotides provided by MYcroarray with custome xGen blocking oligonucleotides (Integrated DNA technologies). We performed 3 independent post-capture amplifications using 12 PCR cycles and pooled these products. We sequenced all samples across 5 Illumina HiSeq 4000 lanes with 100bp paired-end reads. We multiplexed 80 samples per lane for the first 4 lanes and multiplexed the remaining 43

samples on the last lane. Sequencing was carried out at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley. We note that our third lane containing 80 samples was contaminated, with 65% of the reads belonging to corn DNA. We were able to resequence this entire lane, resulting in overall increased sequencing effort for samples belonging to our third lane (Table S3-1).

Data filtration and initial assembly

We filtered the raw read data as follows:

(1) we trimmed reads using Trimmomatic v0.36 under the following conditions: (a) we used the ILLUMINACLIP option to trim adapters with a seed mismatch threshold of 2, a palindrome clip threshold of 40, and a simple clip threshold of 15, (b) we performed quality trimming used the SLIDINGWINDOW option with a window size of 4 and a quality threshold of 20, (c) we removed reads below 36bp by setting the MINLEN option to 36, and (d) we removed leading and trailing bases under a quality threshold of 15.

(2) we merged reads using FLASH v1.2.11 (Magoč & Salzberg 2011) with a min overlap parameter of 5, a max overlap parameter of 100, and a mismatch ratio of 0.05.

(3) we removed low complexity reads using prinseq v0.20.4 (Schmieder & Edwards 2011) using the entropy method with a conservative threshold of 60.

We assembled the filtered read data using SPAdes v3.8.1 using default parameters and reduced redundancy in the resultant assemblies with cap3 (Huang & Madan 1999) under default parameters and cd-hit v4.6 (Li & Godzik 2006, sequence identity threshold = 99%).

Phylogenetic data processing and filtering

To associate assembled contigs with the target sequences for phylogenetic inference, we used blastn v2.2.31 (word size = 11, evalue = 1e-10). For the set of target sequences that originated from the transcriptome dataset, we redefined exon/intron boundaries using EXONERATE v2.2.0 (Slater & Birney 2005) using the est2genome model because we found that several predicted exons actually consisted of several smaller exons. For each sample, we mapped reads using bowtie2 (very sensitive local and no discordant options enabled) to a reference that contained only sequences associated with the targeted phylogenetic markers. We marked duplicates using picard-tools v2.0.1 and masked all regions below 4X coverage and removed the entire sequence if more than 30% of the sequence was below 4X coverage. We called SNPs using samtools v1.3 and bcftools v1.3 and estimated average heterozygosity across all contigs within a sample. We removed sequences if a contig had a heterozygosity value greater than two standard deviations away from the mean.

Conotoxin assembly, processing, and filtering

Commonly used assembly programs are known to poorly reconstruct all copies of multilocus gene families (Lavergne *et al.* 2015; Phuong *et al.* 2016). To address this issue, we followed the conotoxin assembly workflow outlined in (Phuong & Mahardika 2017). Briefly, we first mapped reads back to our assembled contigs using the 'very sensitive local' and no discordant' options. Then, we identified conotoxins within our dataset by using blastn v2.2.31 (word size = 11, evalue = 1e-10) to associate our assembled contigs (from SPAdes) with conotoxins we targeted in the bait design. We generated a set of unique conotoxin 'seed sequences' (a short stretch [~100bp] of conotoxin-blasted sequence) using a combination of of the pysam module (https://github.com/pysam-developers/pysam), cd-hit v4.6 (percent identity =

98%), cap3 (overlap percent identity cutoff = 99%), blastn v2.2.31 (word size = 11, evalue=1e-10), and Tandem Repeats Finder v4.09 (Benson 1999, minscore = 12, maxperiod = 2). We mapped reads to these seed sequences using bowtie2 v2.2.6 (very sensitive local and no discordant options enabled) and built out the conotoxin sequences using the PRICE v1.2 algorithm, which uses an iterative mapping and extension strategy to build out contigs from initial seed sequences (Ruby *et al.* 2013). We ran price on each seed sequence at 5 minimum percentage identity (MPI) values (90%, 92%, 94%, 96%, 98%) with a minimum overlap length value of 40 and a threshold value of 20 for scaling overlap for contig-edge assemblies. A reassembled sequence was retained if it shared 90% identity with the original seed sequence and we reduced redundancy by only retaining the longest sequence per seed sequence out of the 5 MPI assembly iterations. This approach is described in further detail in (Phuong & Mahardika 2017). We note that the final conotoxin sequences per sample consisted of exon fragments, where each sequence represents a single conotoxin exon flanked by any adjacent noncoding region.

We updated our conotoxin reference database because we targeted additional conotoxin transcripts from (Phuong *et al.* 2016).We used blastn v2.2.31 (word size = 11, evalue =1e-10) and EXONERATE v2.2.0 to define exon/intron boundaries for these additional conotoxin transcripts and added them to our conotoxin reference database. The final conotoxin reference database consisted of conotoxin sequences with the coding regions denoted and gene superfamily annotated. We also annotated the conotoxin sequences for functional region (e.g., signal, pre, mature, post) using blastn v2.2.31 (word size = 11, evalue = 1e-10) with a conotoxin reference database that was previously categorized by functional region (Phuong & Mahardika 2017).

With the final conotoxin reference database, we performed blastn v2.2.31 (word size = 11, evalue = 1e-10) searches between the conotoxin reference and every sample's re-assembled conotoxin sequences. We retained sequences if they could align across the entire coding region of the reference sequence. We guessed the coding region for each retained sequence by aligning the query sequence with the reference conotoxin using mafft v7.222 and denoting the coding region as the region of overlap with the exon in the reference conotoxin. We fixed misassemblies by mapping reads with bowtie2 (very sensitive local and no discordant options enabled, score min = L, 70, 1) back to each conotoxin assembly and marked duplicates using picard-tools v2.0.1. We masked regions below 5X coverage and discarded sequences if coverage was below 5X across the entire predicted coding region. To generate the final set of conotoxin sequences per sample, we merged sequences using cd-hit v4.6.4 (percent identity = 98%, use local sequence identity, alignment coverage of longer sequence = 10%, alignment coverage of short sequence = 50%).

Targeted sequencing experiment evaluation

We generated the following statistics to evaluate the overall efficiency of the capture experiment: (1) we calculated the % reads mapped to our targets by mapping reads to a reference containing all targets (both phylogenetic markers and conotoxin sequences) using bowtie2 v2.2.7 (very sensitive local and no discordant options enabled, score min = L, 70, 1), (2) we calculated the % duplicates that were identified through the picard-tools, and (3) we calculated average coverage across the phylogenetic markers and conotoxin sequences. We also evaluated the effect of tissue quality (measured by the maximum fragment length of the extracted DNA sample via gel electrophoresis) and genus (only on *Conus*, *Profundiconus*, and *Conasprella*, the three genera

with more than 1 sample included in this study) on these capture efficiency metrics using an Analysis of Variance (ANOVA). To assess the effectiveness of conotoxin sequence recovery, we compared our capture results with conotoxin diversity estimates from (Phuong & Mahardika 2017) and calculated the average change in those estimates.

Phylogenetic inference

In addition to the 362 samples that we sequenced in this study, we obtained sequences for 10 other species (Table S3-1). For two of these species, we used data from another targeted sequencing study (Abdelkrim *et al.* unpublished). We used blastn (word size = 11, evalue = 1e-10) to identify loci that were present in our phylogenetic marker reference. These sequences were filtered under conditions similar to the filtering strategy applied to the phylogenetic markers in this study. For the other eight species, we used data from venom duct transcriptomes (Safavi-Hemami *et al.* unpublished). With these transcriptomes, we trimmed data using trimmomatic v0.36 and merged reads using flash using parameters previously described above. We assembled each transcriptome using Trinity v2.1.1 (Grabherr *et al.* 2011) reduced redundancy in these transcriptomes with cap3 and cd-hit (percent identity = 99%). We used blastn (word size = 11, evalue=1e-10) to associate contigs with the phylogenetic markers present in our dataset. We used bowtie2 v2.2.7 (very sensitive local and no discordant enabled), samtools v1.3, and bcftools 1.3 to map reads and call SNPs. We removed sequences if they were below 4X coverage for > 30% of the sequence and masked bases if they were below 4X coverage. We also removed sequences if they had a heterozygosity value two standard deviations away from the mean heterozygosity within a sample. We used to mafft v7.222 to align loci across a total of 373 samples.

We inferred phylogenies under both maximum likelihood (Stamatakis 2006) and coalescent-based methods (Mirarab & Warnow 2015). We used RAxML v8.2.9 (Stamatakis 2006) to generate a maximum likelihood phylogeny using a concatenated alignment under a GTRGAMMA model of sequence evolution and estimated nodal support via bootstrapping. We generated the coalescent-based phylogeny using ASTRAL-II v5.5.9 (Mirarab & Warnow 2015) with individual locus trees generated under default parameters in RAxML v8.2.9. We estimated local posterior probabilities as a measure of branch support (Sayyari & Mirarab 2017). Due to the underperformance of the capture experiment, we ran both phylogenetic analyses with loci that had 80% of the taxa, 50% of the taxa, and 20% of the taxa. For each iteration, we removed taxa that had > 90% missing data.

Time calibration

We estimated divergence times using a Bayesian approach with MCMCTree implemented in PAML v4.9g (Yang 2007). Given the size of our alignments, we first estimated branch lengths using baseml and then estimated divergence times using Markov chain Monte Carlo (MCMC). We used a HKY85 + Γ substitution model and used an independent rates clock model. We left all other settings on default. We performed two independent runs of the analysis and checked for convergence among the runs. To account for uncertainty in branching order in our phylogeny, we executed dating analyses across all trees generated from RAxML.

For time calibration, we applied a maximum constraint of 55 million years at the root of Conidae, which corresponds with the first confident appearance of Conidae in the fossil record (Kohn 1990). We assigned 12 additional fossils (Table S3-2, Fig. S3-1 (Duda Jr. *et al.* 2001; Hendricks 2009, 2015, 2018)) to nodes throughout the phylogeny as minimum age constraints,

which MCMCtree treats as soft bounds on the minimum age (Yang 2007). Further information on fossil placement on nodes can be found in the Supplementary.

Characterizing diversification patterns

To visualize lineage accumulation patterns, we generated a log-lineage through time plot using the R package APE (Paradis *et al.* 2018). We estimated diversification rates and identified rate shifts using BAMM (Bayesian Analysis of Macroevolutionary Mixtures) (Rabosky 2014), which uses reversible jump Markov chain Mone Carlo to explore potential lineage diversification models. To account for non-randomness in species sampling across Conidae genera, we applied generic-specific sampling fractions. Using the number of valid Conidae names on WoRMS as estimates of total species diversity in each genus (Worms Editorial Board 2017), we applied a sampling fraction of 32.1% to *Profundiconus*, 50% to *Lilliconus*, 100% to *Californiconus*, 16.7% to *Pygmaeconus*, 28% to *Conasprella*, and 33.7% to *Conus*. We ran BAMM for 100 000 000 generations and assessed convergence by calculating ESS values. We analyzed and visualized results using the R package BAMMtools (Rabosky *et al.* 2014).

Trait dependent diversification

We tested for the impact of evolvability (measured as conotoxin gene diversity) on diversification patterns using two trait dependent diversification methods, focusing on the genus *Conus*. We focused our hypothesis testing on *Conus* because conotoxin diversity is well-characterized in this group (Phuong *et al.* 2016) and the sequence capture approach used in this study likely represents uniform sampling in conotoxin gene diversity across the genus. This is in contrast to other genera in Conidae, such as *Conasprella* or *Profundiconus*, where low conotoxin

diversity values are likely the result of poor knowledge of the venom repertoire of these genera (Fig. S3-2)

First, we used BiSSE (binary state speciation and extinction, (Maddison *et al.* 2007)) implemeneted in the R package diversitree (FitzJohn 2012), which employs a maximum likelihood approach to estimate the impact of a binary trait on speciation, extinction, and transition rates between character states. We coded the conotoxin gene diversity data as 'low' or 'high' across several thresholds (i.e., 250, 300, 350, 400, 500, 550, or 600 estimated conotoxin genes per species) and compared BiSSE models where diversification parameters were allowed to vary or remain equal between traits. We applied a sampling fraction of 33.7%, taking the maximum number of *Conus* species to be the number of valid names on WoRMS (World Register of Marine Species, (Worms Editorial Board 2017)). We determined the best-fitting model using Akaike Information Criterion (AIC).

We also used STRAPP (Structured Rate Permutations on Phylogenies, (Rabosky & Huang 2016) implemented in the R package BAMMtools (Rabosky *et al.* 2014). STRAPP is a semi-parametric approach that tests for trait dependent diversification by comparing a test statistic with a null distribution generated by permutations of speciation rates across the tips of the phylogeny (Rabosky & Huang 2016). We generated the empirical correlation (method = Spearman's rank correlation) between speciation rates and conotoxin gene diversity and compared this test statistic with the null distribution of correlations generated by permutations of evolutionary rates across the tree. We performed a two-tailed test with the alternative hypothesis that there is a correlation between speciation rates and total conotoxin gene diversity.

Results

Targeted sequencing data

We sequenced an average of 9,548,342 reads (range: 1,693,918 – 29,888,444) across the 363 samples (Table S3-1). After redefining exon/intron boundaries in the phylogenetic marker reference, we ultimately targeted 2210 loci. On average, we recovered 1388 of these loci per sample (range: 30 – 1849, Table S3-1) at an average coverage of 12.39X (range: 3.08X – 27.87X, Table S3-1). For the conotoxin dataset, each sequence we re-assembled contained a single conotoxin exon with any associated noncoding regions (referred to here as 'conotoxin fragments'). We recovered on average 3416 conotoxin fragments per sample (range: 74 – 11535 fragments, Table S3-1) at an average coverage of 32.3X (range: 5.06X – 65.77X, Table S3-1). When mapped to a reference containing both the phylogenetic markers and conotoxin genes, the % reads mapped to our targets was on average 14.86% (range: 0.7% - 38.07%, Table S3-1) and the average level of duplication was 47.47% (range: 22.89% - 89.06%, Table S3-1).

We found that genus had an impact on % mapped and % duplication, where non-*Conus* genera had lower % mapping and lower % duplication (Fig. S3-2). These differences likely occurred because conotoxin fragments were not easily recovered in these genera (ANOVA, $p < 0.0001$, Fig. S3-2). Genus did not have an impact on coverage or the number of phylogenetic markers recovered (ANOVA, $p > 0.05$, Fig. S3-2). We found that tissue quality, measured by the maximum fragment length visualized via gel electrophoresis, had a significant impact on the capture efficiency metrics (ANOVA, $p < 0.0001$, Fig. S3-3). DNA samples with strong genomic bands at the top of the gel tended to have higher % mapping, less % duplication, higher coverage, and a greater number of targets recovered (Fig. S3-3).

Our final conotoxin sequence dataset consists of exon fragments and we do not have information on exon coherence (which exons pair together on the same gene). We were unable to assemble full conotoxin genes because conotoxin introns are long (>1 kilobases, (Wu *et al.* 2013)) and exceed the average insert size of our sequencing experiment (~350bp). We recovered fragments from all 58 gene superfamilies we targeted and obtained 159,670 sequences containing some or all of the mature toxin region (Table S3-3). Total conotoxin gene diversity per species (estimated by summing across all signal region exon fragments and sequences containing the entire coding region) ranged from 5 to 1280 copies in *Conus*, 31 to 88 copies in *Profundiconus*, and 7 to 164 in *Conasprella* (Table S3-1). Total conotoxin diversity was 311 copies for *Californiconus californicus*, 12 copies for *Pygmaeconus tralli*, and 30 copies for the outgroup taxon, *Bathyoma sp* (Table S3-1). When compared to samples in (Phuong & Mahardika 2017), the average change (increase or decrease) in total conotoxin gene diversity was ~90 gene copies (Table S3-4). If samples performed poorly in the number of phylogenetic markers recovered, conotoxin gene diversity estimates tended to be lower in this study than in (Phuong & Mahardika 2017) and vice versa (Fig. S3-4). The average absolute change in the number of fragments recovered per gene superfamily by region was 3.7 for sequences containing the signal region, 12.2 for the prepro region, 9.6 for the mature region, 48.9 for the post region, and 3.4 for sequences containing the entire coding region (Table S3-5, Fig. S3-5). We note several key outliers: the average absolute change in the number of fragments was 104.3 for the T gene superfamily containing the prepro region, 210.4 for the O1 gene superfamily prepro region, 57.4 for the O1 gene superfamily mature region, 219.9 for the O2 gene superfamily mature region, and 1417 for the T gene superfamily post region (Table S3-5, Fig. S3-5).

Phylogeny

The amount of missing data from the alignments was 15.4% when a minimum of 80% of the taxa were present in each locus, 26.8% when 50% of the taxa were present, and 38.6% when 20% of the taxa were present. The number of loci retained in the alignment was 387 (107,011 bp) when a minimum of 80% of the taxa were present in each locus, 976 (237,027 bp) when 50% of the taxa were present, and 1476 loci (336,557 bp) when 20% of the loci were present. Across all methods and datasets, we recovered phylogenies with a moderate level of resolution (average number of nodes resolved = 71.1%, range = 61.4 - 79.2%, Table S3-6). In general, as increased amounts of sequence data was given to the phylogenetic programs, more nodes became resolved (Table S3-6). While we recovered all 6 genera within Conidae with high confidence, relationships among subgenera were less supported (bootstrap and PP = 100%, Fig. S3-1, Fig. S3-6, S3-7, S3-8).

Divergence time estimation

We found evidence for three major branching events during the Eocene: (1) a branching event leading to *Profundiconus* (54.7 mya, CI = 47.8 – 64 mya, Fig. S3-1, S3-9), (2) a branching event leading to *Conus* (53.2 mya, CI = 46.2 – 61.7 mya, Fig. S3-1, S3-9), and (3) a branching event separating *Conasprella* and *Californiconus*, *Lilliconus*, and *Pygmaeconus* (44.5 mya, CI = 52.8 – 36.5 mya, Fig. S3-1, S3-9). The branching event leading to *Californiconus* occurred during the Oligocene (25.2 mya, CI = 18.2 – 32.2 mya, Fig. S3-1, S3-9) and the split between *Lilliconus* and *Pygmaeconus* occurred during the Miocene (16.7 mya, CI = 11.7 – 22.1 mya, Fig. S3-1, S3-9).

Diversification patterns

We found that most branching events within each genus began to occur in the Miocene and continued until the present (Fig. S3-1). We found support for diversification rate heterogeneity, where BAMM identified at least one rate shift across Conidae (Fig. S3-1, S3-10). Across the 95% credible set of distinct shift configurations, BAMM detected an increase in diversification rates on the branch leading to *Lautoconus*, a clade consisting mainly of species endemic to the Cape Verde islands (Fig. S3-1, S3-10).

Trait dependent diversification

Across all thresholds for the BiSSE analysis, we found that diversification rates were not influenced by conotoxin gene diversity. In all cases, the null model was either preferred (delta AIC > 2, Table S3-7) or was indistinguishable from a model where speciation and extinction were allowed to vary (delta AIC < 2, Table S3-7). STRAPP analyses revealed that speciation rates were not correlated with conotoxin gene diversity (Spearman's correlation coefficient (r) = 0.01, p = 0.93, Fig. S3-11).

Discussion

Capture results

Our targeted sequencing experiment underperformed initial testing of this sequencing method on cone snails (Phuong & Mahardika 2017). Although tissue quality impacted capture metrics (Fig. S3-3), the % of reads mapping to our targets for even our best samples was ~30% lower than expected (Phuong & Mahardika 2017). While it is difficult to determine the exact cause of this depression in our capture statistics, we hypothesized that changes made in the bait

design between this study and (Phuong & Mahardika 2017) may have led to poorer capture results. For example, we recovered an overabundance of conotoxin sequences containing the post region from the T gene superfamily that has no clear co-variation pattern with phylogenetic relatedness (Fig. S3-12), which likely indicates a large amount of non-specific binding due to conotoxin misclassification. In the future, we suggest re-designing the baits to only include sequences from only the most critical regions (signal region and mature region) to avoid non-specific binding. Although overall capture efficiency statistics were low, the absolute change in conotoxin diversity estimates per gene superfamily was generally minor (Table S3-5). Therefore, we do not believe that total conotoxin diversity metrics were severely biased by the sequencing method.

Phylogenetic relationships

Below, we discuss the results of our phylogenetic analyses, how the phylogenetic relationships compare to past work, and their implications for Conidae taxonomy. Unless otherwise noted, the results we highlight below have at least 90% bootstrap support in the RAxML analyses and 90% posterior probabilities from the ASTRAL-II analyses (Figure S3-7, S3-8). When present results on subgeneric relationships starting from the top of the tree shown in Figure S3-6.

We recovered all six major deep lineages representing genera in Conidae that were previously described in recent molecular phylogenetic studies using mtDNA (Puillandre *et al.* 2014a; Uribe *et al.* 2017), Fig. S3-1, S3-6, S3-7, S3-8). Specifically, we find strong support for *Profundiconus*, *Californiconus*, *Lilliconus*, *Pygmaeconus*, *Conasprella*, and *Conus*, as separate and distinct lineages. We also confirm the branching order of these six genera that were recently

described using mtDNA genomes (Uribe *et al.* 2017), with *Profundiconus* being sister to all other genera, *Pygmaeconus* + *Lilliconus* sister to *Californiconus*, *Californiconus* + *Lilliconus* + *Pygmaeconus* sister to *Conasprella*, and these four genera sister to *Conus*. Given that *Pygmaeconus* and *Lilliconus* both originated after the other four major Conidae lineages, we propose to synonymize both these genera as *Lilliconus*.

Based on the molecular phylogeny from three mtDNA genes, monophyletic groupings of species from *Conasprella* were classified into several subgenera (Puillandre *et al.* 2014a; b). We note several differences between past results and our study in the relationships among these genera and their monophyly:

(1) *Ximeniconus* is sister to all other *Conasprella* in some trees, or we reconstructed a polytomy at the base of *Conasprella*, which contrasts with *Conasprella* (*Kohniconus*) *arcuata* recovered at the base of *Conasprella* in previous work.

(2) *Kohniconus* is polyphyletic. In (Puillandre *et al.* 2014a), only a single species from *Kohniconus* was included in the study and we find evidence for the non-monophyly of *Kohniconus* when we included the additional species, *C. centurio*. Given these results, we propose that *C. emarginatus*, *C. delssertii*, and *C. centurio* be placed in the subgenus *Kohniconus* and *C. arcuata* placed in _____.

(3) *Endemoconus* is paraphyletic. When including an additional species (*C. somalica*) not sequenced in (Puillandre *et al.* 2014a), we find that *Endemoconus* is not monophyletic. Based on these results, *C. somalica* should be transferred to *Conasprella*.

Within *Conus*, our results largely confirm previous findings that *C. distans* is sister to all other *Conus* species and the relationships among subgenera remain tenuous and difficult to resolve

(Puillandre *et al.* 2014a). We note the following differences in subgenera relationships and classification between our results and past work:

(1) We found support the sister relationship between *Turriconus* and *Stephanoconus*, which has not been recovered in a previous study (Puillandre *et al.* 2014a).

(2) We found support for the monophyly of *Pyruconus* across our RAxML analyses, but not our ASTRAL-II analyses. The monophyly of *Pyruconus* was not supported in (Puillandre *et al.* 2014a).

(3) *C. trigonus* and *C. lozeti* were classified into the subgenus (*Plicaustraconus*) based on morphological characters (Puillandre *et al.* 2014b). We found this subgenus to be polyphyletic when sequence data was obtained. We propose that additional data is required to classify *C. trigonus* into the appropriate subgenus while we transfer *C. lozeti* to ____.

(4) Similar to (Puillandre *et al.* 2014a), we found that *Textila* + *Afonsoconus* is sister to *Pionoconus*. However, instead of the unsupported relationship of *Asprella* as sister to these three subgenera, we found support for *Gastridium* as the sister group.

(5) We found support for the sister relationship between *Asprella* and *Phasmoconus*, which conflicts with the unsupported relationship shown (Puillandre *et al.* 2014a), where these subgenera branch in different parts of the phylogeny.

(6) We find support for the following successional branch order: *Tesselliconus*, *Plicaustraconus*, *Eugeniconus*, and *Conus*. We found that *Conus* is sister to *Leptoconus*, *Darioconus*, and *Cylinder*, but the relationships among these three subgenera remained unresolved. This conflicts with (Puillandre *et al.* 2014a), as *Cylinder* was paraphyletic

with the inclusion *C. nobilis*, whereas in our results with increased sampling of *Eugeniconus*, *Cylinder* became monophyletic.

(7) We did not find strong support for the subgenus *Calibanus*, contrasting with previous work (Puillandre *et al.* 2014a). In our results, we found that *C. thalassiarchus* and *C. furvus* were not sister to each other, or their relationship resulted in an unresolved polytomy. Additional investigation into the subgeneric status of these two species is necessary before removing them from *Calibanus*.

(8) *C. sanderi* was classified into its own subgenus (*Sandericonus*) based on morphological characters (Puillandre *et al.* 2014b). Here, when sequence data were obtained, we found it nested within *Dauciconus*. No other species within this subgenus have been sequenced up until this point. Therefore, we synonymize *Sandericonus* with *Dauciconus* because *C. sanderi* is the type species for *Sandericonus*.

(9) *C. granulatus* was classified into its own subgenus (*Atlanticonus*) based on morphological characters (Puillandre *et al.* 2014b). Here, we found that it was nested within *Dauciconus*. No other species within this subgenus have been sequenced up until this point. Therefore, we synonymize *Atlanticonus* with *Dauciconus* because *C. granulatus* is the type species for *Atlanticonus*.

(10) *C. hyaena* was classified into *Rhizoconus* based on morphological characters. Based on sequence results from this study, we found that *C. hyaena* is actually nested within *Asprella*. Therefore, we transfer *C. hyaena* to the subgenus *Asprella*.

(11) Three species (*C. pergrandis*, *C. moncuri*, and *C. darkini*) sequenced in this study were placed into the subgenus *Embrikena* (Puillandre *et al.* 2014b). Our results do not support the monophyly of *Embrikena*, as *C. darkini* was found to be nested within

Turriconus and the sister relationship between *C. moncuri* and *C. pergrandis* was not supported in 5/6 trees. Based on these results, we transfer *C. darkini* to the subgenus, *Turriconus* while additional data is required to classify *C. moncuri* and *C. pergrandis* into the appropriate subgenus.

(12) *C. cocceus* was placed into *Floraconus* based on morphological characters in (Puillandre *et al.* 2014b). With sequence data, we found that it was actually nested within *Phasmoconus*. Therefore, we transfer *C. cocceus* to the subgenus, *Phasmoconus*.

(13) *C. bajanensis* was placed into the *Conasprella* subgenus *Dalliconus* based on morphological characters. Here, we found it nested within *Dauciconus*. Therefore, we transfer the species *bajaensis* to the genus *Conus* and the subgenus *Dauciconus*.

Classification within Conidae is known to be highly unstable (Jiménez-Tenorio & Tucker 2013; Puillandre *et al.* 2014a; b; Puillandre & Tenorio 2018). Although the phylogeny presented here improved understanding of subgeneric relationships and monophyly of subgenera, resolving relationships within Conidae still remains a significant challenge. Given the underperformance of our capture experiment (Table S3-1), it is unclear if the reason for the moderate power in resolving relationships is due to insufficient data/incomplete data or due to short internal branches during the origination of Conidae subgenera that are extremely difficult to resolve. Overall, our results suggest that both additional data and increased sampling of Conidae species are reasonable pursuits to continue attempting to resolve the phylogeny and classification of this family of marine snails.

Timing of diversification

The timing of splits between major are largely congruent with past estimates from a study using mtDNA genomes (Uribe *et al.* 2017), Fig. S3-1, S3-9). However, our age estimates for the branching events between *Californiconus*, *Lilliconus*, and *Pygmaeconus* are much younger (occurring across the Oligocene into the Miocene) than previous estimates (occurring across the Eocene into the Oligocene, (Uribe *et al.* 2017), Fig. S3-1, S3-9). This discrepancy may have been caused by differences in fossil calibration, as we included many more fossils in this study compared to previous studies. The Conidae fossil record and analyses of several molecular phylogenetic studies suggest a major radiation of *Conus* during the Miocene (Kohn 1990; Duda Jr. *et al.* 2001; Uribe *et al.* 2017). While we noted that many branching events within *Conus* ocurred during the Miocene into the present, we did not detect a shift in diversification on the branch leading to the origin of *Conus* (Fig. S3-1, S3-10). This is congruent with diversification rates estimated from the fossil record (Kohn 1990), suggesting that the accumulation of species during the Miocene may have been a function of an increased number of lineages present rather than an increase in diversification rates. While BAMM did not detect a shift in diversification rates at the origin of *Conus*, BAMM detected a shift in diversification rates on the branch leading to the subgenus *Lautoconus*, a known and documented radiation of cone snails (Duda & Rolán 2005; Cunha *et al.* 2005) , Fig. S3-1, S3-10). Several factors are thought to have contributed to the radiation of *Lautoconus*, including geographic isolation and reduced dispersal due to short-lived and non-dispersing larval stages (Cunha *et al.* 2005)

Speciation rates and conotoxin gene diversity

Contrary to macroevolutionary expectations, we were unable to detect any relationship between speciation rates and conotoxin gene diversity (Fig. S3-1, S3-11, Table S3-7). BAMM

detected moderate levels of diversification rate heterogeneity within the genus *Conus*, which likely led to the non-significant result in the STRAPP analysis (Rabosky & Huang 2016), Fig. S3-1, S3-10). STRAPP is known to have little power in identifying trait dependent diversification in phylogenies with only a single shift in diversification rates and BAMM analyses on our data only detected a single shift within *Conus* (Rabosky & Huang 2016), Fig. S3-1, S3-10). Even when performing the analyses with BiSSE, a method in recent years that has become the subject of criticism due to high false positive rates (Abosky 2017; Rabosky & Goldberg 2017), our analyses did not detect an impact of conotoxin gene diversity on diversification rates (Table S3-7). These analyses suggest that conotoxin gene diversity may not be a rate-limiting control on speciation rates in *Conus*.

Several factors may explain this decoupling between conotoxin gene diversity and speciation rates. A critical assumption in *Conus* biology is that ecological diversification driven by diet specialization is a major factor governing diversification dynamics in cone snails (Duda & Palumbi 1999; Duda Jr. *et al.* 2001). Past studies have shown that cone snail venom repertoires track their dietary breadth, providing a link between diet and venom evolution (Phuong *et al.* 2016; Phuong & Mahardika 2017). However, it is unclear whether or not the relationship between diet and venom evolution leads to ecological speciation due to divergence in prey preference. Ecological speciation is often difficult to detect in marine ecosystems and long-term diversification patterns may be better explained by traits that limit dispersal and promote isolation (Bowen *et al.*). Indeed, BAMM identified a shift in diversification rates on the branch leading to *Lautoconus*, a radiation of almost exclusively endemic cone snail species that is thought to be due to presence of nonplanktonic larvae. Another possibility is that conotoxin phenotypic divergence may not be the rate-limiting factor in prey specialization and divergence

(Duda Jr. *et al.* 2001). Conotoxin genes are under continuous positive selection and gene duplication that allow venom components to change rapidly in response to the environment (Duda & Palumbi 1999; Duda Jr. *et al.* 2001; Chang & Duda 2012; Phuong & Mahardika 2017). This persistent evolutionary change in the venom cocktail suggests that perhaps venom evolution is not necessarily the factor limiting dietary shifts among species and ultimately, speciation among taxa. Ecological opportunity is hypothesized as a necessary component for diversification (Losos 2010) and may be a more critical factor shaping Conidae diversification. Indeed, evidence from the fossil record and past Conidae molecular phylogenetic studies indicate a concentration of lineage formation during the Miocene (Kohn 1990; Duda Jr. *et al.* 2001; Uribe *et al.* 2017), a period that is coincident with the formation of coral reefs in the Indo-Australian Archipelago (Cowman & Bellwood 2011). Our results also show a concentration of branching events during this period as well, though we do not detect a shift in diversification rates (Fig. S3-1). Regardless, these results together suggests that further research should be devoted to the impact of environmental factors governing diversification dynamics in cone snails.

A potential limitation with testing the relationship between venom gene diversity and speciation rates is the surprising lack of heterogeneity in diversification rates across *Conus* (Fig. S3-1). Although we incorporated the proportion of taxa sampled into our diversification analyses, we still only obtained phylogenetic markers for ~33% of the species described in *Conus*. This incomplete sampling may have hindered our ability to detect greater variation in diversification rates across *Conus*. Therefore, future studies should continue to expand the phylogenetic sampling of *Conus* and re-evaluate the results reported here.

Venom evolution is assumed to be a key innovation that led to the evolutionary success of venomous animal lineages (Pyron & Burbrink 2011; Sunagar *et al.* 2016) and a large body of

work is devoted towards understanding how venom evolves and responds to the environment over time (Kordis & Gubensek 2000; Wong & Belov 2012; Casewell *et al.* 2013). However, the impact of venom evolution on higher-level diversification patterns is rarely tested. Here, we examined the effect of variation in the adaptive capacity of venom across Conidae species and found it had no influence on speciation patterns. Although these results suggest that conotoxin gene diversity does not limit speciation rates in Conidae, it does not refute the importance of venom evolution in adaptation and prey specialization, as diet is well documented to have an impact on venom composition at multiple biological scales (Duda *et al.* 2009; Safavi-Hemami *et al.* 2015; Chang & Duda 2016; Phuong *et al.* 2016; Phuong & Mahardika 2018). Rather, our results suggest that venom may be necessary, but not sufficient, to promote speciation. Future work in other venomous animal systems may shed light on whether or not the ability to adapt to different prey through venom evolution translates to the long-term evolutionary success of taxa.

Data availability

Raw read data will be made available at the National Center for Biotechnology Information Sequence Read Archive. Bait sequences, conotoxin sequences, scripts, and final datasets used for analyses will be uploaded onto Dryad following publication.

Acknowledgements

A majority of the sampling material in this paper originates from numerous shore-based expeditions and deep sea cruises, conducted respectively by MNHN and Pro-Natura International (PNI) as part of the Our Planet Reviewed programme (SANTO 2006, ATIMO VATAE, MAINBAZA, INHACA 2011, GUYANE 2014, PAPUA NIUGINI, KAVIENG 2014), by

MNHN and AAMP (Pakaihi i Te Moana), and/or by MNHN and Institut de Recherche pour le Développement (IRD) as part of the Tropical Deep-Sea Benthos programme (AURORA 2007, BIOPAPUA, EBISCO, EXBODI, MADEEP, MIRIKY, TAIWAN 2013, NANHAI 2014, BIOPAPUA, SALOMONBOA 3, CONCALIS, EXBODI, SALOMON BOA3, KARUBENTHOS 2015, NORFOLK 2, TERRASSES). Scientific partners included the University of Papua New Guinea (UPNG); National Fisheries College, Kavieng; Institut d'Halieutique et Sciences Marines (IH.SM), Université de Tuléar, Madagascar; Universidade Eduardo Mondlane, Maputo; the Madagascar bureau of the Wildlife Conservation Society (WCS); and Instituto Español de Oceanografia (IOE). Funders and sponsors included the Total Foundation, Prince Albert II of Monaco Foundation, Stavros Niarchos Foundation, Richard Lounsbery Foundation, Vinci Entrepose Contracting, Fondation EDF, European Regional Development Fund (ERDF), the Philippines Bureau of Fisheries and Aquatic Research (BFAR), the French Ministry of Foreign Affairs, Fonds Pacifique and the Government of New Caledonia. Additional field work included PANGLAO 2004 and PANGLAO 2005 (joint projects of MNHN and University of San Carlos, Cebu City, and the Philippines Bureau of Fisheries and Aquatic Research); KARUBENTHOS 2012 (a joint project of MNHN with Parc National de la Guadeloupe and Université des Antilles); sampling in Western Australia arranged by Hugh Morrison, with support of the Western Australian Museum. The Taiwan and South China Sea cruises were supported by bilateral cooperation research funding from the Taiwan Ministry of Science and Technology (MOST 102-2923-B-002-001-MY3, PI Wei-Jen Chen) and the French National Research Agency (ANR 12-ISV7-0005-01, PI Sarah Samadi). All expeditions operated under the regulations then in force in the countries in question and satisfy the conditions set by the Nagoya Protocol for access to genetic resources. We thank the Indonesian Ministry of State

for Research and Technology (RISTEK) for providing permission to MAP and GNM to conduct fieldwork Bali in 2014 (permit number 277/SIP/FRP/SM/VIII/ 2013) and providing permission to PWHV TvR RMM and REP to conduct fieldwork across Indonesia in 2016 (permit number 414/SIP/FRP/E5/Dit.KI/X/2015). We thank DST Hariyanto, MBAP Putra, MKAA Putra, and the staff at the Indonesian Biodiversity Research Center in Denpasar, Bali for assistance during the 2014 field season; the staff at the Museum Zoologicum Bogoriense for assistance during the 2016 field season; F Criscione, F Köehler, A Moussalli, A Hogget, and L Vail for logistical assistance for fieldwork at the Lizard Island Research Station in Australia; M Reid, A Hallan, and J Waterhouse for access to specimens at the Australian Museum in Sydney, Australia; J Finn, M Mackenzie, and M Winterhoff for access to specimens at the Museum Victoria in Melbourne, Australia; G Pauley and AM Bemis for access to specimens at the Florida Museum of Natural History at the University of Florida; TF Duda Jr. and T Lee for access to specimens at the University of Michigan Museum of Zoology; L Kirkendale and C Whisson for access to specimens at the Western Australian Museum in Perth, Australia; WF Gilly for access to the *C. californicus* specimen; H Safavi-Hemami and Q Li for access to 10 additional Conidae transcriptomes; K Bi for advice on bait design; A Devault and MYcroarray for great service and technical support for bait synthesis; L Smith and the Evolutionary Genetics Lab at UC Berkeley for laboratory support; J Chang, MCW Lim and EM McCartney-Melstad for thoughtful advice and discussions throughout the entire process. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575. This work was supported by two Grants-in-Aid of research from Sigma Xi, a Grants-in-Aid of Research from the Society for Integrative and Comparative Biology, a research grant from the Society of Systematic Biologists, a Student

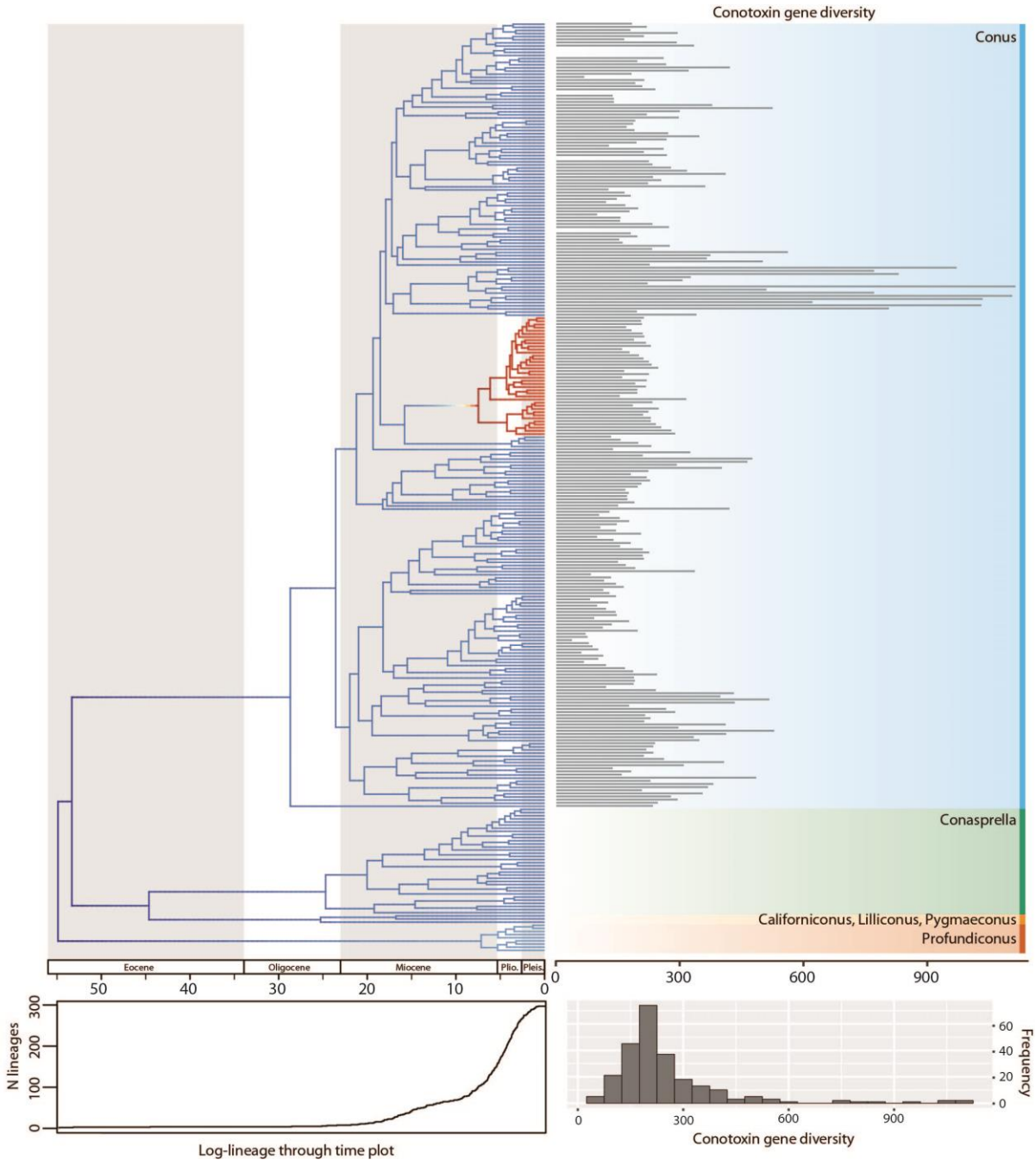
Research Award from the American Society of Naturalists, a National Science Foundation Graduate Research Opportunities Worldwide to Australia, the Lerner Gray Fund for Marine Research from the American Museum of Natural History, research grants from the Department of Ecology and Evolutionary Biology at UCLA, the Melbourne R. Carriker Student Research Award from the American Malacological Society, an Academic Grant from the Conchologists of America, a Student Research Award from Unitas Malacologicas, a Lemelson Fellowship from the UCLA Indonesian Studies program, the Lewis and Clark Fund from the American Philosophical Society, a Young Explorer's Grant from the National Geographic Society, a Travel Award from the UCLA Graduate Division, a Research Grant from the American Institute for Indonesian Studies and the Council of American Overseas Research Centers, a small award from the B Shaffer Lab, a National Science Foundation Graduate Research Fellowship, an Edwin W. Pauley fellowship, a Fulbright Fellowship to Indonesia, and a Chateaubriand fellowship awarded to MAP. This work was supported by the Service de Systématique Moléculaire (UMS 2700 CNRS-MNHN) and the CONOTAX project funded by the French National Research Agency (grant number ANR-13-JSV7-0013-01). This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. The *C. californicus* specimen was collected under a California Department of Fish and Wildlife collecting permit granted to WF Gilly (SC-6426).

List of Figures

in the alignment. Colors across the phylogeny show instantaneous diversification rates and are averaged across all rate models sampled from a BAMM analysis. Warmer colors indicate higher speciation rates. Log-lineage through time plot is shown below the phylogeny. Bars are shown at tips depicting variation conotoxin gene diversity across the phylogeny. If bar is not shown, data is not available or were excluded from downstream diversification analyses. Histogram on the bottom right shows variation in conotoxin gene diversity. Abbreviations: Plio. = Pliocene; Pleis. = Pleistocene.

Figures

Figure 3-1



List of supplementary material

Figure S3-1. Node placement of fossils. Numbers correspond to node placement justification in the supplementary information on node assignment. Tree was generated from a RAxML analysis of a concatenated alignment where loci were kept if at least 20% of species was present in the locus. Best tree is shown and erroneous and intraspecific tips were pruned.

Figure S3-2. Boxplots showing impact of phylogeny (categorized by Conidae genus) on capture efficiency metrics. Graph title shows resultant P value from ANOVA analyses.

Figure S3-3. Boxplots showing impact of tissue quality (estimated by maximum DNA fragment lengths assessed via gel electrophoresis) on capture efficiency metrics. Categories are either "g" for genomic band, or 1500, 1000, or 500 for bands beginning at 1500bp, 1000bp, or 500bp. Graph title shows resultant P value from ANOVA analyses.

Figure S3-4. Scatterplot showing relationship between the number of phylogenetic markers recovered and the change in total conotoxin gene diversity between this study and (Phuong & Mahardika 2017). Results showed a positive relationship between the two parameters, suggesting that if a sample performed poorly in the capture experiment, it performed poorly in recovering data across all loci (phylogenetic loci or conotoxin loci).

Figure S3-5. Histograms showing absolute change in conotoxin sequence diversity per gene superfamily between this study and (Phuong & Mahardika 2017). Graphs are partitioned by conotoxin functional region, where sequences were categorized based on whether they contained

the entire coding region or mostly the signal, prepro, mature, or post regions. On average, estimates of conotoxin diversity per gene superfamily varied slightly.

Figure S3-6. Maximum likelihood phylogeny inferred using RAxML, where 20% of the taxa needed to be present within a locus to be included in the final concatenated alignment. The six major genera are colored and subgenera re noted for *Conasprella* and *Conus*.

Figure S3-7. Phylogenies inferred through the coalescent-based method, ASTRAL-II. Individual loci were inferred under default paramenters in RAxML. Nodes are collapsed when posterior probailities are <90%. Trees are colored and labeled by genus. We varied the level of missing data for each ASTRAL run, where we only retained loci if (a) 80% of taxa had sequences, (b) 50% had sequences, and (c) 20% of taxa had sequences.

Figure S3-8. Maximum likelihood phylogenies generated using a concatenated alignment. Nodes are collapsed when boostrap support values are <90%. Trees are colored and labeled by genus. We varied the level of missing data for each RAxML run, where we only retained loci for the nal concatenated alignment if (a) 80% of taxa had sequences, (b) 50% had sequences, and (c) 20% of taxa had sequences.

Figure S3-9. Maximum likelihood phylogeny dated with 13 fossil node calibrations in MCMCtree. 95% confidence intervals shown at nodes. The final concatenated alignment consisted of loci where 20% of the taxa needed to be present within the locus to be included.

Figure S3-10. 95% credible set of distinct shift configurations from BAMM. Each graph is labeled by the posterior probability of each shift configuration. Warmer, red colors represent faster speciation rates than cooler, blue colors. We note that in all shift configurations, there is a shift in diversification rates in the clade leading to *Lautoconus*.

Figure S3-11. Scatterplot showing relationship between conotoxin gene diversity and BAMM speciation rates. STRAPP results produced a Spearman's correlation coefficient (r) of 0.01 and a p value of 0.93.

Figure S3-12. Diversity estimates for the A gene superfamily signal region and the T gene superfamily post region. Estimates are plotted next to the RAxML phylogeny where 20% of taxa had sequences in each locus.

Table S3-1. Sample information and capture efficiency metrics. We first list the species name listed in all phylogenetic analyses ("species") and the accepted taxonomic classification in the WoRMS database at the genus, subgenus, and species level ("WoRMS genus", "WoRMS subgenus", "WoRMS species"). We then list the specimen ID ("ID"), the collection source ("Collection"), the year the sample was collected ("Year Collected"), how the sample was preserved ("Preservation type"), and the country the sample originated from ("Country"). We then list "Gelsize", or the largest fragment size visualized via gel electrophoresis as a way to measure tissue quality. Values can be either "g" for genomic band, or 1500, 1000, or 500 for bands beginning at 1500bp, 1000bp, or 500bp. We list the data collection method ("Data collection method"), the number of reads sequenced ("# of reads sequenced") and several capture

efficiency metrics. Finally, we list the total estimated number of conotoxin genes per species ("Conotoxin gene diversity").

Table S3-2. Information on fossils used for calibration. We list the fossil taxon ("Fossil Species"), the genus ("Clade assignment"), extant species related to the fossil ("Compared with"), the formation ("Formation"), the age of the fossil ("Age"), and its citation ("Reference").

Table S3-3. Table showing the number of conotoxin sequences recovered per species for each gene superfamily. Within each gene superfamily, conotoxin sequences were categorized based on whether they contained the entire coding region or mostly the signal, prepro, mature, or post regions.

Table S3-4. Comparison of conotoxin gene diversity estimates between this study and (Phuong & Mahardika 2017). These represent comparisons between technical replicates (capture experiment was performed on the same libraries in both studies).

Table S3-5. Comparison of conotoxin gene diversity estimates between this and (Phuong & Mahardika 2017), broken down by gene superfamily. Within each gene superfamily, conotoxin sequences were categorized based on whether they contained the entire coding region or mostly the signal, prepro, mature, or post regions. These represent comparisons between technical replicates (capture experiment was performed on the same libraries in both studies).

Table S3-6. Number of nodes resolved depending on the amount of missing data and the tree inference method. Phylogenetic trees were inferred using either RAxML or ASTRAL-II. The "% taxa per locus was" the percent of samples needed per locus in order to retain the locus for phylogenetic inference.

Table S3-7. BiSSE AIC results. "Threshold" represents the conotoxin gene diversity value used to decide between "high" and "low" conotoxin diversity. Values above the threshold value were categorized as "high" and values below were categorized as "low". "AIC – variable rates" shows AIC values for a model where speciation and extinction rates were allowed to vary depending on a trait. "AIC – equal rates" represents AIC values for the null model, where rates were not allowed to vary by trait.

References

Abosky DALR (2017) Model Inadequacy and Mistaken Inferences of Trait-Dependent Speciation. , **64**, 340–355.

Adams DC (2014) A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. *Evolution*, **68**, 2675–2688.

Altschup SF, Gish W, Pennsylvania T, Park U (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**, 403–410.

Aman JW, Imperial JS, Ueberheide B *et al.* (2015) Insights into the origins of fish hunting in venomous cone snails from studies of *Conus tessulatus*. *Proceedings of the National Academy of Sciences*, **112**, 5087–5092.

Ashkenazy H, Penn O, Doron-faigenboim A *et al.* (2012) FastML : a web server for probabilistic reconstruction of ancestral sequences. , **40**, 580–584.

Bankevich A, Nurk S, Antipov D *et al.* (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, **19**, 455–477.

Bansal V, Bafna V (2008) HapCUT: An efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, **24**, 153–159.

Barghi N, Concepcion GP, Olivera BM, Lluisma AO (2014) High conopeptide diversity in Conus tribblei revealed through analysis of venom duct transcriptome using two high-throughput sequencing platforms. *Marine biotechnology*, **17**, 81–98.

Barghi N, Concepcion GP, Olivera BM, Lluisma AO (2015a) Comparison of the venom peptides

and their expression in closely related Conus species: insights into adaptive post-speciation evolution of Conus exogenomes. *Genome Biology and Evolution*, **7**, 1797–1814.

Barghi N, Concepcion GP, Olivera BM, Lluisma AO (2015b) Structural features of conopeptide genes inferred from partial sequences of the Conus tribblei genome. *Molecular Genetics and Genomics*, **291**, 411–422.

Barlow A, Pook CE, Harrison RA, Wüster W (2009) Coevolution of diet and prey-specific venom activity supports the role of selection in snake venom evolution. *Proceedings of the Royal Society B*, **276**, 2443–2449.

Benson G (1999) Tandem Repeats Finder: a program to analyse DNA sequences. *Nucleic Acids Res.*, **27**, 573–578.

Bi K, Vanderpool D, Singhal S *et al.* (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**, 403.

Binford GJ (2001) Differences in venom composition between orb-weaving and wandering Hawaiian Tetragnatha (Araneae). *Biological Journal of the Linnean Society*, **74**, 581–595.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bolnick DI, Svanbäck R, Araújo MS, Persson L (2007) Comparative support for the niche variation hypothesis that more generalized populations also are more heterogeneous. *Proceedings of the National Academy of Sciences*, **104**, 10075–10079.

Bowen BW, Rocha LA, Toonen RJ, Karl SA The origins of tropical marine biodiversity.

Brahma RK, McCleary RJR, Kini RM, Doley R (2015) Venom gland transcriptomics for identifying, cataloging, and characterizing venom proteins in snakes. *Toxicon : official journal of the International Society on Toxinology*, **93**, 1–10.

Brawand D, Wagner CE, Li YI *et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, **51.**

Bresler M, Sheehan S, Chan AH, Song YS (2012) Telescoper: De novo assembly of highly repetitive regions. *Bioinformatics*, **28**, 311–317.

Casewell NR, Harrison RA, Wüster W, Wagstaff SC (2009) Comparative venom gland transcriptome surveys of the saw-scaled vipers (Viperidae: Echis) reveal substantial intra-family gene diversity and novel venom transcripts. *BMC Genomics*, **10**, 564.

Casewell NR, Wagstaff SC, Harrison RA, Renjifo C, Wu W (2011) Domain Loss Facilitates Accelerated Evolution and Neofunctionalization of Duplicate Snake Venom Metalloproteinase Toxin Genes. , **28**, 2637–2649.

Casewell NR, Wüster W, Vonk FJ, Harrison RA, Fry BG (2013) Complex cocktails: the evolutionary novelty of venoms. *Trends in ecology & evolution*, **28**, 219–29.

Chang D, Duda TF (2012) Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Molecular biology and evolution*, **29**, 2019–29.

Chang D, Duda TF (2014) Application of community phylogenetic approaches to understand gene expression: differential exploration of venom gene space in predatory marine gastropods. *BMC evolutionary biology*, **14**, 123.

Chang D, Duda TF (2016) Age-related association of venom gene expression and diet of

predatory gastropods. *BMC Evolutionary Biology*, **16**, 27.

Chang D, Olenzek AM, Duda Jr. TF (2015) Effects of geographical heterogeneity in species interactions on the evolution of venom genes. *Proceedings of the Royal Society B*, **282**, 20141984.

Conticello SG, Gilad Y, Avidan N *et al.* (2001) Mechanisms for evolving hypervariability: the case of conopeptides. *Molecular biology and evolution*, **18**, 120–31.

Cornetti L, Valente LM, Dunning LT *et al.* (2015) The genome of the "great speciator" provides insights into bird diversification. *Genome Biology and Evolution*, **7**, 2680–2691.

Cowman PF, Bellwood DR (2011) Coral reefs as drivers of cladogenesis: expanding coral reefs, cryptic extinction events, and the development of biodiversity hotspots. *Journal of evolutionary biology*, **24**, 2543–2562.

Creer S, Malhotra A, Thorpe RS *et al.* (2003) Genetic and ecological correlates of intraspecific variation in pitviper venom composition detected using matrix-assisted laser desorption time-of-flight mass spectrometry (MALDI-TOF-MS) and isoelectric focusing. *Journal of Molecular Evolution*, **56**, 317–329.

Crow KD, Wagner GP (2006) What is the role of genome duplication in the evolution of complexity and diversity? *Molecular biology and evolution*, **23**, 887–92.

Cunha RL, Castilho R, Rüber L, Zardoya R (2005) Patterns of cladogenesis in the venomous marine gastropod genus Conus from the Cape Verde islands. *Systematic biology*, **54**, 634–50.

Daltry JC, Wüster W, Thorpe RS (1996) Diet and snake venom evolution. *Nature*, **379**, 537–540.

Deutsch M, Long M (1999) Intron – exon structures of eukaryotic model organisms. *Nucleic Acids Research*, **27**, 3219–3228.

Dowell NL, Giorgianni MW, Kassner VA *et al.* (2016) The deep origin and recent loss of venom toxin genes in rattlesnake. *Current Biology*, **26**, 2434–2445.

Duda TF (2008) Differentiation of venoms of predatory marine gastropods: divergence of orthologous toxin genes of closely related Conus species with different dietary specializations. *Journal of molecular evolution*, **67**, 315–21.

Duda TF, Bolin MB, Meyer CP, Kohn AJ (2008) Hidden diversity in a hyperdiverse gastropod genus: discovery of previously unidentified members of a Conus species complex. *Molecular phylogenetics and evolution*, **49**, 867–76.

Duda TF, Chang D, Lewis BD, Lee T (2009) Geographic variation in venom allelic composition and diets of the widespread predatory marine gastropod Conus ebraeus. *PloS one*, **4**, e6245.

Duda Jr. TF, Kohn AJ, Palumbi SR (2001) Origins of diverse feeding ecologies within Conus, a genus of venomous marine gastropods. *Biological Journal of the Linnean Society*, **73**, 391–409.

Duda TF, Lee T (2009) Ecological release and venom evolution of a predatory marine snail at Easter Island. *PloS one*, **4**.

Duda TF, Palumbi SR (1999) Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod Conus. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6820–3.

Duda TF, Palumbi SR (2000) Evolutionary diversification of multigene families: allelic selection

of toxins in predatory cone snails. *Molecular biology and evolution*, **17**, 1286–93.

Duda TF, Remigio EA (2008) Variation and evolution of toxin gene expression patterns of six closely related venomous marine snails. *Molecular ecology*, **17**, 3018–32.

Duda TF, Rolán E (2005) Explosive radiation of Cape Verde Conus, a marine species flock. *Molecular ecology*, **14**, 267–72.

Dutertre S, Jin A, Kaas Q *et al.* (2013) Deep venomics reveals the mechanism for expanded peptide diversity in cone snail venom. *Molecular & cellular proteomics*, **12**, 312–29.

Dutertre S, Jin A-H, Vetter I *et al.* (2014) Evolution of separate predation- and defence-evoked venoms in carnivorous cone snails. *Nature communications*, **5**, 3521.

Elliger CA, Richmond TA, Lebaric ZN *et al.* (2011) Diversity of conotoxin types from Conus californicus reflects a diversity of prey types and a novel evolutionary history. *Toxicon*, **57**, 311–22.

Endean R, Rudkin C (1963) Studies of the venoms of some Conidae. *Toxicon*, **1**, 49–64.

Endean R, Rudkin C (1965) Further studies of the venoms of Conidae. *Toxicon*, **69**, 225–249.

Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic biology*, **61**, 717–726.

FitzJohn RG (2012) Diversitree : comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, **3**, 1084–1092.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates.

Molecular marine biology and biotechnology, **3**, 294–9.

Fry BG, Roelants K, Champagne DE *et al.* (2009) The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. *Annual review of genomics and human genetics*, **10**, 483–511.

Fry BG, Wüster W, Kini RM *et al.* (2003a) Molecular evolution and phylogeny of elapid snake venom three-finger toxins. *Journal of Molecular Evolution*, **57**, 110–129.

Fry BG, Wüster W, Ryan Ramjan SF *et al.* (2003b) Analysis of Colubroidea snake venoms by liquid chromatography with mass spectrometry: evolutionary and toxinological implications. *Rapid communications in mass spectrometry : RCM*, **17**, 2047–2062.

Gibbs HL, Rossiter W (2008) Rapid evolution by positive selection and gene gain and loss: PLA(2) venom genes in closely related Sistrurus rattlesnakes with divergent diets. *Journal of molecular evolution*, **66**, 151–66.

Gibbs HL, Sanz L, Sovic MG, Calvete JJ (2013) Phylogeny-based comparative analysis of venom proteome variation in a clade of rattlesnakes (Sistrurus sp.). *PloS one*, **8**.

Gilles A, Meglécz E, Pech N *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC genomics*, **12**, 245.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing. *Nature Biotechnology*, **27**, 182–189.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, **29**, 644–52.

Guillén Y, Rius N, Delprat A *et al.* (2014) Genomics of ecological adaptation in cactophilic Drosophila. *Genome Biology and Evolution*, **7**, 349–366.

Gutiérrez JM, Avila C, Camacho Z, Lomonte B (1990) Ontogenetic changes in the venom of the snake Lachesis muta stenophrys (bushmaster) from Costa Rica. *Toxicon*, **28**, 419–426.

Han M V., Thomas GWC, Lugo-Martinez J, Hahn MW (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution*, **30**, 1987–1997.

Haney RA, Ayoub NA, Clarke TH, Hayashi CY, Garb JE (2014) Dramatic expansion of the black widow toxin arsenal uncovered by multi-tissue transcriptomics and venom proteomics. *BMC genomics*, **15**, 366.

Hendricks JR (2009) *The Genus Conus (Mollusca:Neogastropoda) in the Plio-Pleistocene of the Southeastern United States*. Paleontological Research Institution, Ithaca, NY.

Hendricks JR (2015) Glowing Seashells : Diversity of Fossilized Coloration Patterns on Coral Reef-Associated Cone Snail (Gastropoda : Conidae) Shells from the Neogene of the Dominican Republic. *Plos One*, **10**, 1–59.

Hendricks JR (2018) Diversity and preserved shell coloration patterns of Miocene Conidae (Neogastropoda) from an exposure of the Gatun Formation, Colón Province, Panama. *Journal of Paleontology*, **In press**.

Henrissat B, Martínez AT, Otillar R *et al.* (2012) The Paleozoic Origin of Enzymatic from 31 Fungal Genomes. *Science*, **336**, 1715–1719.

Hoekstra HE, Coyne JA (2007) The Locus of Evolution: Evo Devo and the Genetics of

Adaptation. *Evolution*, **61**, 995–1016.

Hu H, Bandyopadhyay PK, Olivera BM, Yandell M (2011) Characterization of the Conus bullatus genome and its venom-duct transcriptome. *BMC genomics*, **12**.

Hu H, Bandyopadhyay PK, Olivera BM, Yandell M (2012) Elucidation of the molecular envenomation strategy of the cone snail Conus geographus through transcriptome sequencing of its venom duct. *BMC genomics*, **13**.

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Research*, **9**, 868–877.

Hughes AL, Yeager M (1997) Comparative evolutionary rates of introns and exons in murine rodents. *J Mol Evol*, **45**, 125–130.

Jiménez-Tenorio M, Tucker JK (2013) *Illustrated catalog of the Living Cone snails*. MdM Publishing, Wellington, FL.

Jin A, Dutertre S, Kaas Q *et al.* (2013) Transcriptomic messiness in the venom duct of Conus miles contributes to conotoxin diversity. *Molecular & cellular proteomics*, **12**, 3824–33.

Jin A, Israel MR, Inserra MC *et al.* (2015) delta-Conotoxin SuVIA suggests an evolutionary link between ancestral predator defence and the origin of fish-hunting behaviour in carnivorous cone snails. *Proceedings of the Royal Society B*, **282**.

Jones AG, Arnold SJ, Bürger R (2007) The mutation matrix and the evolution of evolvability. *Evolution*, **61**, 727–45.

Jones M, Good J (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.

Kaas Q, Westermann J-C, Craik DJ (2010) Conopeptide characterization and classifications: an analysis using ConoServer. *Toxicon*, **55**, 1491–509.

Kaas Q, Yu R, Jin A-H, Dutertre S, Craik DJ (2012) ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic acids research*, **40**, D325-30.

Katoh K, Kuma K-I, Toh H, Miyata T (2005) MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**, 511–518.

Kembel SW, Cowan PD, Helmus MR *et al.* (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**, 1463–1464.

Kirschner M, Gerhart J (1998) Evolvability. *Proceedings of the National Academy of Sciences*, **95**, 8420–8427.

Kohn AJ (1959a) The ecology of Conus in Hawaii. *Ecological Monographs*, **29**, 47–90.

Kohn AJ (1959b) Ecological notes on Conus (Mollusca: Gastropoda) in the Trincomalee region of Ceylon. *Annals and Magazine of Natural History*, **2**, 308–320.

Kohn AJ (1966) Food specialization in Conus in Hawaii and California. *Ecology*, **47**, 1041–1043.

Kohn AJ (1968) Microhabitats, abundance and food of Conus on atoll reef's in the Maldive and Chagos islands. *Ecology*, **49**, 1046–1062.

Kohn AJ (1978) Ecological Shift and Release in an Isolated Population: Conus miliaris at Easter Island. *Ecological Monographs*, **48**, 323–336.

Kohn AJ (1981) Abundance, diversity, and resource use in an assemblage of Conus species in Enewetak Lagoon. *Pacific Science*, **34**, 359–369.

Kohn AJ (1990) Tempo and mode of evolution in conidae. *Malacologia*, **32**, 55–67.

Kohn AJ (2001) Maximal species richness in Conus: diversity, diet and habitat on reefs of northeast Papua New Guinea. *Coral Reefs*, **20**, 25–38.

Kohn AJ (2003) Biology of Conus on shores of the Dampier Archipelago, Northwestern Australia.

Kohn AJ (2015) Ecology of Conus on Seychelles reefs at mid-twentieth century: comparative habitat use and trophic roles of co-occurring congeners. *Marine Biology*, **162**, 2391–2407.

Kohn A, Almasi K (1993) Comparative ecology of a biogeographically heterogeneous Conus assemblage. In: *Proceedings of the Fifth International Marine Biological Workshop: The Marine Flora and Fauna of Rottnest Island.*, pp. 509–521.

Kohn AJ, Curran KM, Mathis BJ (2005) Diets of the predatory gastropods Cominella and Conus at Esperance, Western Australia. In: *The Marine Flora and fauna of Esperance, Western Australia*, pp. 235–244.

Kohn AJ, Nybakken JW (1975) Ecology of Conus on eastern Indian Ocean fringing reefs: diversity of species and resource utilization. *Marine Biology*, **29**, 211–234.

Kordis D, Gubensek F (2000) Adaptive evolution of animal toxin multigene families. *Gene*, **261**, 43–52.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**, 357–9.

Lavergne V, Dutertre S, Jin A *et al.* (2013) Systematic interrogation of the Conus marmoreus venom duct transcriptome with ConoSorter reveals 158 novel conotoxins and 13 new gene

superfamilies. *BMC genomics*, **14**.

Lavergne V, Harliwong I, Jones A *et al.* (2015) Optimized deep-targeted proteotranscriptomic profiling reveals unexplored Conus toxin diversity and novel cysteine frameworks. *Proceedings of the National Academy of Sciences*, **112**.

Lewis RJ (2009) Conotoxins: Molecular and therapeutic targets. In: *Marine Toxins as Research Tools*, pp. 45–65.

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, **12**.

Li M, Fry BG, Kini RM (2005a) Eggs-only diet: Its implications for the toxin profile changes and ecology of the marbled sea snake (Aipysurus eydouxii). *Journal of Molecular Evolution*, **60**, 81–89.

Li M, Fry BG, Kini RM (2005b) Putting the brakes on snake venom evolution: The unique molecular evolutionary patterns of Aipysurus eydouxii (marbled sea snake) phospholipase A2 toxins. *Molecular Biology and Evolution*, **22**, 934–941.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–9.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.

Liu Z, Li H, Liu N *et al.* (2012) Diversity and evolution of conotoxins in Conus virgo, Conus eburneus, Conus imperialis and Conus marmoreus from the South China Sea. *Toxicon*, **60**, 982–989.

Losos JB (2010) Adaptive radiation, ecological opportunity, and evolutionary determinism. *The American naturalist*, **175**, 623–639.

Lu A, Yang L, Xu S, Wang C (2014) Various conotoxin diversifications revealed by a venomic study of Conus flavidus. *Molecular & cellular proteomics*, **13**, 105–18.

Maddison WP, Midford PE, Otto SP (2007) Estimating a binary character's effect on speciation and extinction. *Systematic biology*, **56**, 701–10.

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–63.

Malmstrøm M, Matschiner M, Tørresen OK *et al.* (2016) Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, **48**.

Marsh H (1971) Observations on the food and feeding of some vermivorous Conus on the Great Barrier Reef. *The veliger*, **14**, 45–55.

Mayrose I, Zhan SH, Rothfels CJ *et al.* (2011) Recently Formed Polyploid Plants Diversify at Lower Rates. , **333**.

McCartney-Melstad E, Mount GG, Shaffer HB (2016) Exon capture optimization in amphibians with large genomes. *Molecular Ecology Resources*, **16**, 1084–1094.

McIntosh J, Santos A, Olivera B (1999) Conus peptides targeted to specific nicotinic acetylcholine receptor subtypes. *Annual review of biochemistry*, **68**, 59–88.

Metz EC, Robles-Sikisaka R, Vacquier VD (1998) Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial DNA. *Proceedings of the National Academy of Sciences*, **95**, 10676–10681.

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols*, **2010**, doi:10.1101/pdb.prot5448.

Millard EL, Daly NL, Craik DJ (2004) Structure-activity relationships of alpha-conotoxins targeting neuronal nicotinic acetylcholine receptors. *European journal of biochemistry / FEBS*, **271**, 2320–6.

Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, **12**, R112.

Mirarab S, Warnow T (2015) ASTRAL-II : coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, **45**, 44–52.

Nakashima K, Nobuhisa I, Deshimaru M *et al.* (1995) Accelerated evolution in the protein-coding regions is universal in crotalinae snake venom gland phospholipase A2 isozyme genes. *Proceedings of the National Academy of Sciences*, **92**, 5605–5609.

Nakashima K, Ogawa T, Oda N *et al.* (1993) Accelerated evolution of Trimeresurus flavoviridis venom gland phospholipase A2 isozymes. *Proceedings of the National Academy of Sciences*, **90**, 5964–5968.

Nybakken J (2009) Ontogenetic change in the conus radla, its form, distribution among the radula tpes, and significance in systematics and ecology. *Malacologia*, **32**, 35–54.

Nybakken J, Perron F (1988) Ontogenetic change in the radula of Conus magus (Gastropoda). *Marine Biology*, **98**, 239–242.

Oksanen J, Blanchet FG, Friendly M *et al.* (2016) vegan: Community Ecology Package.

Olivera BM (1997) Conus venom peptides, receptor and ion channel targets, and drug design: 50 million years of neuropharmacology. *Molecular biology of the cell*, **8**, 2101–9.

Olivera BM (2002) Conus venom peptides: reflections from the biology of clades and species. *Annual Review of Ecology and Systematics*, **33**, 25–47.

Olivera BM, Rivier J, Clark C *et al.* (1990) Diversity of Conus neuropeptides. *science*.

Olivera BM, Teichert RW (2007) Diversity of the neurotoxic Conus peptides. *Molecular interventions*, **7**, 251–260.

Olivera BM, Walker C, Cartier GE *et al.* (1999) Speciation of cone snails and interspecific hyperdivergence of their venom peptides: potential evolutionary significance of introns. *Annals of the New York Academy of Sciences*, **870**, 223–237.

Orme D (2013) The caper package : comparative analysis of phylogenetics and evolution in R. , 1–36.

Pagel M (1997) Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, **26**, 331–348.

Pahari S, Bickford D, Fry BG, Kini RM (2007) Expression pattern of three-finger toxin and phospholipase A2 genes in the venom glands of two sea snakes, Lapemis curtus and Acalyptophis peronii: comparison of evolution of these toxins in land snakes, sea kraits and sea snakes. *BMC evolutionary biology*, **7**, 175.

Paradis E, Claude J, Strimmer K (2018) APE : Analyses of Phylogenetics and Evolution in R language. , **20**, 289–290.

Parry L, Tanner A, Vinther J (2014) The origin of annelids (A Smith, Ed,). *Palaeontology*, **57**, 1091–1103.

Pease JB, Haak DC, Hahn MW, Moyle LC (2016) Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLoS Biology*, **14**, 1–24.

Phuong M, Mahardika G (2017) Targeted sequencing of venom genes from cone snail genomes reveals coupling between dietary breadth and venom diversity. *biorxiv*.

Phuong MA, Mahardika GN (2018) Targeted Sequencing of Venom Genes from Cone Snail Genomes Improves Understanding of Conotoxin Molecular Evolution. , 1–15.

Phuong MA, Mahardika GN, Alfaro ME (2016) Dietary breadth is positively correlated with venom complexity in cone snails. *BMC Genomics*, **17**, 401.

Pigliucci M (2008) Is evolvability evolvable? *Nature reviews. Genetics*, **9**, 75–82.

Portik DM, Smith LL, Bi K (2016) An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class : Amphibia , Order : Anura). *Molecular Ecology Resources*, **16**, 1069–1083.

Prjibelski AD, Vasilinetc I, Bankevich A *et al.* (2014) ExSPAnder: A universal repeat resolver for DNA fragment assembly. *Bioinformatics*, **30**, 293–301.

Puillandre N, Bouchet P, Duda TF *et al.* (2014a) Molecular phylogeny and evolution of the cone snails (Gastropoda, Conoidea). *Molecular phylogenetics and evolution*, **78**, 290–303.

Puillandre N, Duda TF, Meyer C, Olivera BM, Bouchet P (2014b) One, four or 100 genera? A new classification of the cone snails. *Journal of Molluscan Studies*, 1–23.

Puillandre N, Kantor YI, Sysoev a. *et al.* (2011) The dragon tamed? A molecular phylogeny of

the Conoidea (Gastropoda). *Journal of Molluscan Studies*, **77**, 259–272.

Puillandre N, Koua D, Favreau P, Olivera BM, Stöcklin R (2012) Molecular phylogeny, classification and evolution of conopeptides. *Journal of molecular evolution*, **74**, 297–309.

Puillandre N, Tenorio MJ (2018) Molluscan Studies. , 200–210.

Puillandre N, Watkins M, Olivera BM (2010) Evolution of Conus peptide genes: duplication and positive selection in the A-superfamily. *Journal of molecular evolution*, **70**, 190–202.

Pyron RA, Burbrink FT (2011) EXTINCTION , ECOLOGICAL OPPORTUNITY , AND THE ORIGINS OF GLOBAL SNAKE DIVERSITY. , 163–178.

Rabosky DL (2014) Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PloS one*, **9**, e89543.

Rabosky DL, Goldberg EE (2017) FiSSE : A simple nonparametric test for the effects of a binary character on lineage diversification rates. , 1432–1442.

Rabosky DL, Grundler M, Anderson C *et al.* (2014) BAMMtools : an R package for the analysis of evolutionary dynamics on phylogenetic trees. , 701–707.

Rabosky DL, Huang H (2016) A Robust Semi-Parametric Test for Detecting Trait-Dependent Diversification. *Systematic biology*, **65**, 181–193.

Rabosky DL, Santini F, Eastman J *et al.* (2013) Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature communications*, **4**, 1–8.

Rash LD, Hodgson WC (2002) Pharmacology and biochemistry of spider venoms. *Toxicon*, **40**, 225–254.

Reichelt R, Kohn A (1995) Feeding and distribution of predatory gastropods on some great barrier reef platforms. *Proceedings of the Fifth International Coral Reef Congress*, **1985**, 191–196.

Remigio EA, Duda TF (2008) Evolution of ecological specialization and venom of a predatory marine gastropod. *Molecular ecology*, **17**, 1156–62.

Revell LJ (2012) phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223.

Revell L, Harmon L, Collar D (2008) Phylogenetic Signal, Evolutionary Process, and Rate. *Systematic Biology*, **57**, 591–601.

Richards DP, Barlow A, Wüster W (2012) Venom lethality and diet: Differential responses of natural prey and model organisms to the venom of the saw-scaled vipers (Echis). *Toxicon*, **59**, 110–116.

Robertson FM, Gundappa MK, Grammes F *et al.* (2017) Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. , 1–14.

Robinson S, Norton R (2014) Conotoxin Gene Superfamilies. *Marine Drugs*, **12**, 6058–6101.

Robinson SD, Safavi-Hemami H, McIntosh LD *et al.* (2014) Diversity of conotoxin gene superfamilies in the venomous snail, Conus victoriae. *PloS one*, **9**, e87648.

Rokyta DR, Lemmon AR, Margres MJ, Aronow K (2012) The venom-gland transcriptome of the eastern diamondback rattlesnake (Crotalus adamanteus). *BMC Genomics*, **13**, 213.

Rokyta DR, Wray KP, Margres MJ (2013) The genesis of an exceptionally lethal venom in the

timber rattlesnake (Crotalus horridus) revealed through comparative venom-gland transcriptomics. *BMC genomics*, **14**, 394.

Roy SW (2016) Is Mutation Random or Targeted ?: No Evidence for Hypermutability in Snail Toxin Genes. *Molecular Biology and Evolution*, **33**, 2642–2647.

Ruby JG, Bellare P, Derisi JL (2013) PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3*, **3**, 865–80.

Safavi-Hemami H, Gajewiak J, Karanth S *et al.* (2015) Specialized insulin is used for chemical warfare by fish-hunting cone snails. *Proceedings of the National Academy of Sciences*, **112**, 1743–1748.

Safavi-Hemami H, Lu A, Li Q *et al.* (2016) Venom Insulins of Cone Snails Diversify Rapidly and Track Prey Taxa. *Molecular Biology and Evolution*, **33**, 1–27.

Sakharkar MK, Chow VTK, Kangueane P (2004) Distributions of exons and introns in the human genome. *In Silico Biology*, **4**, 387–393.

Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, **19**, 301–302.

Santini F, Harmon LJ, Carnevale G, Alfaro ME (2009) Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC evolutionary biology*, **9**, 194.

Saunders P, Wolfson F (1961) Food and feeding behavior in Conus californicus Hinds, 1844. *Veliger*, **3**, 73–76.

Sayyari E, Mirarab S (2017) Fast Coalescent-Based Computation of Local Branch Support from

Quartet Frequencies Article Fast Track. , **33**, 1654–1668.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Schoener TW (1968) The Anolis lizards of Bimini: resource partitioning in a complex fauna. *Ecology*, **49**, 704–726.

da Silva Jr. NJ, Aird SD (2001) Prey specificity, comparative lethality and compositional differences of coral snake venoms. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*.

Simakov O, Marletaz F, Cho S-J *et al.* (2013) Insights into bilaterian evolution from three spiralian genomes. *Nature*, **493**, 526–31.

Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, **6**, 31.

Smit AFA, Hubley R, Green P (2015) RepeatMasker Open-4.0.

Soltis DE, Albert V a, Leebens-Mack J *et al.* (2009) Polyploidy and angiosperm diversification. *American journal of botany*, **96**, 336–48.

Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Sunagar K, Casewell NR, Varma S *et al.* (2016) Deadly Innovations: Unraveling the Molecular Evolution of Animal Venoms. , 1–27.

Tank DC, Eastman JM, Pennell MW *et al.* (2015) Nested radiations and the pulse of angiosperm diversification : increased diversification rates often follow whole genome duplications.

Taylor JD (1978) Habitats and diet of predatory gastropods at addu atoll, maldives. *Journal of Experimental Marine Biology and Ecology*, **31**, 83–103.

Taylor JD (1984) A partial food web involving predatory gastropods on a pacific fringing reef. *Journal of Expermental Marine Biology and Ecology*, **14**, 273–290.

Taylor J (1986) Diets of sand-living predatory gastropods at Piti Bay, Guam. *Asian Marine Biology*, **3**, 47–58.

Taylor JD, Reid DG (1984) The abundance and trophic classification of molluscs upon coral reefs in the Sudanese Red Sea. *Journal of Natural History*, **18**, 175–209.

Teasdale LC, Köhler F, Murray KD, O'Hara T, Moussalli A (2016) Identification and qualification of 500 nuclear, single-copy, orthologous genes for the Eupulmonata (Gastropoda) using transcriptome sequencing and exon capture. *Molecular Ecology Resources*, **16**, 1107–1123.

Uribe JE, Puillandre N, Zardoya R (2017) Phylogenetic relationships of Conidae based on complete mitochondrial genomes. *Molecular Phylogenetics and Evolution*, **107**, 142–151.

Van Valen L (1965) Morphological variation and width of ecological niche. *The American Naturalist*, **99**, 377–390.

Vonk FJ, Jackson K, Doley R *et al.* (2011) Snake venom: from fieldwork to the clinic. *Bioessays*, **33**, 269–279.

Wagner GP, Altenberg L (1996) Perspective: Complex Adaptations and the Evolution of Evolvability. *Evolution*, **50**, 967–976.

Wagner GP, Kin K, Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data:

RPKM measure is inconsistent among samples. *Theory in biosciences*, **131**, 281–5.

Williams V, White J, Schwaner TD, Sparrow A (1988) Variation in venom proteins from isolated populations of tiger snakes (Notechis ater niger, N. scutatus) in South Australia. *Toxicon*, **26**, 1067–1075.

Wong ESW, Belov K (2012) Venom evolution through gene duplications. *Gene*, **496**, 1–7.

Worms Editorial Board (2017) World Register of Marine Species.

Wu Y, Wang L, Zhou M *et al.* (2013) Molecular evolution and diversity of Conus peptide toxins, as revealed by gene structure and intron sequence analyses. *PloS one*, **8**, e82495.

Yang AS (2001) Modularity, evolvability, and adaptive radiations: a comparison of the hemi- and holometabolous insects. *Evolution & development*, **3**, 59–72.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, **24**, 1586–91.

Zhan SH, Glick L, Tsigenopoulos CS, Otto SP, Mayrose I (2014) Comparative analysis reveals that polyploidy does not decelerate diversification in fish. , **27**, 391–403.

Zhang Y, Chen J, Tang X *et al.* (2010) Transcriptome analysis of the venom glands of the Chinese wolf spider Lycosa singoriensis. *Zoology*, **113**, 10–8.