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

Mark A Phuong¹, Gusti N Mahardika²

¹Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA
²Animal Biomedical and Molecular Biology Laboratory, Faculty of Veterinary Medicine, Udayana University Bali, Jl Sesetan-Markisa 6, Denpasar, Bali 80225, Indonesia

Corresponding author:
M. A. Phuong
E-mail: markphuong@gmail.com
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 gene 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 1959a; Daltry et al. 1996; Li et al. 2005; 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. 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. 2014). 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 1). We sequenced all samples on a single Illumina HiSeq2000 lane, producing an average of 12.8 million reads per sample (Table 1). 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 1). 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 1). 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 1). Average coverage across the phylogenetic markers was 95.9X, while the average coverage for the conotoxin exons was 149.6X (Table 1).

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. S1). 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 S1). 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 S1). 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). We recovered a higher percentage of previously sequenced conotoxins if the species was included in the bait design (52.35%, Table S2) compared to species not included in the bait design (39.5%, Table S2).

**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, Table S3). 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 S3). 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. 3). In contrast, the prepro region was more frequently distributed across more than one exon (Fig. 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 S4, Fig. 4, S2). 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 S4, Fig. 4, S2). 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 S4, Fig. 4, S2). For comparison, exons from non-conotoxin genes were more conserved than their adjacent noncoding regions (average relative ratio < 1, Fig. 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 S5, Fig. S3) 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 S5, Fig. S3), and in other cases, all copies were expressed in the transcriptome (e.g., *C. californicus*, O3 gene superfamily, 3/3 copies expressed, Table S5, Fig. S3). The average proportion of gene copies expressed per gene superfamily per species was 45% (range: 24% – 63%, Table S5).

**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. 5). Total conotoxin gene diversity ranged from as low as 120 in *C. papilliferus* to as high as 859 in *C. coronatus* (Fig. 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 S6). 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. 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. 5, S4).

**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. 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 1) 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. 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 S3). 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 S5, Fig. S3). 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 1) is on par with studies that employed similar techniques in Anuran frogs (mean = 60.2% (Portik et al. 2016)) and Salamanders (mean = 18.21% (McCarty-Melstad et al. 2016)). The proportion of reads marked as duplicates were higher than previous studies (mean = 24.5% (McCarty-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 1), 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 S1). 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 divMKFPLLFISSL gene superfamily and we were unable to recover full sequences from any of the other species (Table S1). Second, the genetic organization of gene superfamilies may 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 S1, S3, Fig. S1). 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 S4, Fig. 4, S2).

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 S7). 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). 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 S3) and the number of exons per gene (range = 1 – 6 exons, Table S3) 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. 2015a). For example, (Barghi et al. 2015a) 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. 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. 2015a).

**Conotoxin molecular evolution**

A previous analysis of patterns of conotoxin divergence suggested that introns within conotoxin 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 S4, Fig. 4, S2). This is similar to the pattern found in non-conotoxin exons (Fig. 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 S5, Fig. S3), concurring with similar reports from smaller sets of gene superfamilies (Chang & Duda 2012, 2014; Barghi et al. 2015a). 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. 2015b; 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 1959a, 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 pseudogenized. 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 1959b). 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 S6, Fig. 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. 5, S4). 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. 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. 2015b; 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 S1). 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 1). 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 (Altschul et al. 1990) with an evalue threshold of $1 \times 10^{-10}$ 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 S5.

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 S8). 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 S8. 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 redund 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. 2014).

**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).

**Figure 1.** A superfamily conotoxins from *Conus lividus* described in the transcriptome from Phuong et al. 2016. Protein alignment generated using Geneious. Amino acids are colored based on disagreement to a consensus sequence generated from the alignment, not shown. Cysteines are colored and bolded. Signal region and mature toxin coding region are annotated based on the presence of these functional regions at a particular position in the alignment in any of the sequences.

**Figure 2.** Exon length distribution. Exon length distribution across all conotoxin gene superfamilies sequenced in this study. Analysis only includes sequences from species that had transcriptome data available.
Figure 3. Histograms showing the frequency of the largest proportion of each conotoxin precursor peptide region found on a single exon in Conidae genomes. Analysis only includes sequences from species that had transcriptome data available.

Figure 4. Scatterplot of uncorrected pairwise distances for select gene superfamilies and non-conotoxin loci between exons and adjacent noncoding regions. 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 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 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.

Figure S1. Conotoxin diversity per gene family per conotoxin functional region in a phylogenetic context.

Figure S2. 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 S3. 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 S4. 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 S5. Bioinformatic workflow for conotoxin assembly.** We outline the overall bioinformatic workflow and provide complementary figures to illustrate the bioinformatic processing of the data.
Table 1. Sample information and exon capture statistics.

<table>
<thead>
<tr>
<th>Species</th>
<th># reads</th>
<th>% reads on target</th>
<th>% duplicates</th>
<th># bases recovered</th>
<th># exons recovered</th>
<th>Exon coverage</th>
<th>conotoxin fragments recovered</th>
<th>Collection Source</th>
<th>Year collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>magus</td>
<td>10,348,142</td>
<td>48.17%</td>
<td>56.18%</td>
<td>652113</td>
<td>5258</td>
<td>79.0</td>
<td>332</td>
<td>Australian Museum (Catalog No. C487620)</td>
<td>2009</td>
</tr>
<tr>
<td>textile</td>
<td>8,130,024</td>
<td>47.86%</td>
<td>51.79%</td>
<td>649193</td>
<td>5236</td>
<td>64.8</td>
<td>764</td>
<td>Australian Museum (Catalog No. C487632)</td>
<td>2009</td>
</tr>
<tr>
<td>lischkeanus</td>
<td>8,942,888</td>
<td>45.30%</td>
<td>52.64%</td>
<td>627806</td>
<td>5024</td>
<td>64.2</td>
<td>834</td>
<td>Australian Museum (Catalog No. C480270)</td>
<td>2011</td>
</tr>
<tr>
<td>aristophanes</td>
<td>15,276,200</td>
<td>52.96%</td>
<td>53.68%</td>
<td>655444</td>
<td>5310</td>
<td>92.9</td>
<td>2278</td>
<td>Australian Museum (Catalog No. C480269)</td>
<td>2011</td>
</tr>
<tr>
<td>papilliferus</td>
<td>9,733,028</td>
<td>46.97%</td>
<td>46.97%</td>
<td>659001</td>
<td>5318</td>
<td>82.5</td>
<td>281</td>
<td>Australian Museum (Catalog No. C480190)</td>
<td>2011</td>
</tr>
<tr>
<td>californicus</td>
<td>11,257,992</td>
<td>40.29%</td>
<td>46.23%</td>
<td>581181</td>
<td>4659</td>
<td>89.2</td>
<td>649</td>
<td>Monterey Bay, California (provided by WF Gilly)</td>
<td>2014</td>
</tr>
<tr>
<td>virgo</td>
<td>8,319,348</td>
<td>51.90%</td>
<td>44.74%</td>
<td>659419</td>
<td>5312</td>
<td>84.9</td>
<td>406</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>quercinus</td>
<td>12,260,196</td>
<td>47.95%</td>
<td>53.40%</td>
<td>667897</td>
<td>5409</td>
<td>92.6</td>
<td>500</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>ebraeus</td>
<td>16,311,454</td>
<td>49.57%</td>
<td>55.37%</td>
<td>669852</td>
<td>5427</td>
<td>120.7</td>
<td>849</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>flavidaus</td>
<td>12,175,022</td>
<td>46.25%</td>
<td>41.37%</td>
<td>671106</td>
<td>5446</td>
<td>110.0</td>
<td>614</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>lividaus</td>
<td>13,561,122</td>
<td>51.65%</td>
<td>54.67%</td>
<td>663079</td>
<td>5365</td>
<td>93.5</td>
<td>987</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>miles</td>
<td>18,577,352</td>
<td>50.13%</td>
<td>49.06%</td>
<td>680191</td>
<td>5541</td>
<td>175.9</td>
<td>594</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>ratti</td>
<td>16,948,836</td>
<td>52.08%</td>
<td>55.04%</td>
<td>677626</td>
<td>5497</td>
<td>139.7</td>
<td>727</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>distans</td>
<td>12,706,016</td>
<td>50.85%</td>
<td>57.47%</td>
<td>658791</td>
<td>5411</td>
<td>100.0</td>
<td>384</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>sponsalis</td>
<td>11,168,242</td>
<td>46.70%</td>
<td>44.51%</td>
<td>657141</td>
<td>5313</td>
<td>77.8</td>
<td>2033</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>imperialis</td>
<td>8,838,484</td>
<td>43.70%</td>
<td>50.19%</td>
<td>652250</td>
<td>5242</td>
<td>69.0</td>
<td>397</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>marmoreus</td>
<td>11,494,484</td>
<td>52.20%</td>
<td>50.27%</td>
<td>669298</td>
<td>5430</td>
<td>103.8</td>
<td>524</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>varius</td>
<td>15,035,404</td>
<td>50.78%</td>
<td>52.57%</td>
<td>666670</td>
<td>5387</td>
<td>113.6</td>
<td>743</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>musicus</td>
<td>12,761,042</td>
<td>46.59%</td>
<td>56.15%</td>
<td>649666</td>
<td>5245</td>
<td>71.9</td>
<td>1970</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>planoris</td>
<td>13,623,080</td>
<td>44.80%</td>
<td>58.94%</td>
<td>658461</td>
<td>5330</td>
<td>76.0</td>
<td>899</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>miliaris</td>
<td>13,039,246</td>
<td>52.88%</td>
<td>55.59%</td>
<td>640651</td>
<td>5152</td>
<td>67.2</td>
<td>2106</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>vexillum</td>
<td>12,197,882</td>
<td>49.41%</td>
<td>52.89%</td>
<td>668942</td>
<td>5417</td>
<td>97.6</td>
<td>603</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>arenatus</td>
<td>14,607,148</td>
<td>47.23%</td>
<td>48.61%</td>
<td>665490</td>
<td>5397</td>
<td>98.3</td>
<td>1466</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>muriculatus</td>
<td>17,328,206</td>
<td>47.25%</td>
<td>59.85%</td>
<td>662812</td>
<td>5359</td>
<td>102.3</td>
<td>1072</td>
<td>Bali, Indonesia (Field collection)</td>
<td>2013</td>
</tr>
<tr>
<td>Species</td>
<td>Total</td>
<td>MA</td>
<td>FM</td>
<td>XII</td>
<td>LMA</td>
<td>LMA%</td>
<td>LMA#/1000</td>
<td>MA#/1000</td>
<td>Life Span</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
<td>-----------</td>
</tr>
<tr>
<td>chaldaeus</td>
<td>11,778,640</td>
<td>50.84%</td>
<td>55.12%</td>
<td>661691</td>
<td>5371</td>
<td>91.3</td>
<td>704</td>
<td>131.8</td>
<td>Bali, Indonesia (Field collection)</td>
</tr>
<tr>
<td>coronatus</td>
<td>14,873,018</td>
<td>49.92%</td>
<td>45.09%</td>
<td>670729</td>
<td>5442</td>
<td>100.8</td>
<td>2158</td>
<td>179.3</td>
<td>Bali, Indonesia (Field collection)</td>
</tr>
<tr>
<td>moreleti</td>
<td>10,843,692</td>
<td>49.41%</td>
<td>57.47%</td>
<td>653436</td>
<td>5282</td>
<td>78.8</td>
<td>485</td>
<td>142.8</td>
<td>Bali, Indonesia (Field collection)</td>
</tr>
<tr>
<td>anemone</td>
<td>10,662,156</td>
<td>42.78%</td>
<td>41.61%</td>
<td>670558</td>
<td>5440</td>
<td>94.2</td>
<td>321</td>
<td>130.6</td>
<td>Museum Victoria (Registration No. F237402)</td>
</tr>
<tr>
<td>mustelinus</td>
<td>15,317,812</td>
<td>50.86%</td>
<td>58.11%</td>
<td>668295</td>
<td>5416</td>
<td>115.5</td>
<td>679</td>
<td>153.8</td>
<td>Australian Museum (Catalog No. C487551)</td>
</tr>
<tr>
<td>litteratus</td>
<td>13,002,630</td>
<td>51.11%</td>
<td>56.64%</td>
<td>673310</td>
<td>5451</td>
<td>99.8</td>
<td>538</td>
<td>148.4</td>
<td>Australian Museum (Catalog No. C487581)</td>
</tr>
<tr>
<td>capitaneus</td>
<td>16,366,304</td>
<td>51.30%</td>
<td>57.35%</td>
<td>672762</td>
<td>5461</td>
<td>120.4</td>
<td>909</td>
<td>151.9</td>
<td>Australian Museum (Catalog No. C487552)</td>
</tr>
<tr>
<td>emaciatus</td>
<td>13,391,964</td>
<td>51.38%</td>
<td>54.88%</td>
<td>665699</td>
<td>5387</td>
<td>101.7</td>
<td>554</td>
<td>168.3</td>
<td>Australian Museum (Catalog No. C487593)</td>
</tr>
</tbody>
</table>
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.

Literature cited


<table>
<thead>
<tr>
<th>Species</th>
<th>Total conodont diversity</th>
<th>Select gene superfamilies</th>
<th>Daily breadth</th>
<th>Neoprotrocoela</th>
<th>Heteropleidae</th>
<th>Monoplacophora</th>
<th>Mollusca</th>
<th>Annelida</th>
</tr>
</thead>
<tbody>
<tr>
<td>arenatus</td>
<td>352</td>
<td>A Q T</td>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>miliaris</td>
<td>60</td>
<td>A Q T</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aristophanes</td>
<td>400</td>
<td>A Q T</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arenatus</td>
<td>326</td>
<td>A Q T</td>
<td>1.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>moriellus</td>
<td>72</td>
<td>A Q T</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tamhinius</td>
<td>192</td>
<td>A Q T</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>turfus</td>
<td>349</td>
<td>A Q T</td>
<td>1.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>squinicus</td>
<td>161</td>
<td>A Q T</td>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>muticus</td>
<td>28</td>
<td>A Q T</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>capitanus</td>
<td>430</td>
<td>A Q T</td>
<td>1.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>miles</td>
<td>278</td>
<td>A Q T</td>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rattsus</td>
<td>162</td>
<td>A Q T</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>planorbis</td>
<td>436</td>
<td>A Q T</td>
<td>1.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>virkius</td>
<td>299</td>
<td>A Q T</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>imperialis</td>
<td>144</td>
<td>A Q T</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>papilliferus</td>
<td>120</td>
<td>A Q T</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aeolus</td>
<td>137</td>
<td>A Q T</td>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>distans</td>
<td>197</td>
<td>A Q T</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>californicus</td>
<td>207</td>
<td>A Q T</td>
<td>2.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>