Expression Differentiation is Constrained to Low-Expression Proteins Over Ecological Timescales

Mark J. Margres,* Kenneth P. Wray,* Margaret Seavy,* James J. McGivern,* Nathanael D. Herrera,* Darin R. Rokyta*

November 5, 2015

*Department of Biological Science, Florida State University, Tallahassee, FL, 32306
Running Head: Constraints on Expression Differentiation

Key Words: Protein expression, Selective constraints, Evolutionary rates, Adaptation

Corresponding Author:

Darin R. Rokyta

Department of Biological Science
Florida State University
319 Stadium Drive
4058 King Life Sciences
Tallahassee, FL 32306
(850) 645-8812 (ph.)
(850) 645-8447 (fax)
drokyta@bio.fsu.edu
Abstract

Protein expression level is one of the strongest predictors of protein sequence evolutionary rate, with high-expression protein sequences evolving at slower rates than low-expression protein sequences, largely because of constraints on protein folding and function. Expression evolutionary rates have also been shown to be negatively correlated with expression level across human and mouse orthologs over relatively long divergence times (i.e., approximately 100 million years). Long-term evolutionary patterns, however, often cannot be extrapolated to microevolutionary processes (and vice versa), and whether this relationship holds for traits evolving under directional selection within a single species over ecological timescales (i.e., <5,000 years) is unknown and not necessarily expected. Expression is a metabolically costly process, and the expression level of a particular protein is predicted to be a trade-off between the benefit of its function and the costs of its expression. Selection should drive the expression level of all proteins close to values that maximize fitness, particularly for high-expression proteins because of the increased energetic cost of production. Therefore, stabilizing selection may reduce the amount of standing expression variation for high-expression proteins and, in combination with physiological constraints that may place an upper-bound on the range of beneficial expression variation, these constraints could severely limit the availability of beneficial expression variants. To determine whether rapid expression evolution was restricted to low-expression proteins due to these constraints on highly expressed proteins over ecological timescales, we compared venom protein expression levels across mainland and island populations for three species of pit vipers. We detected significant differentiation in protein expression levels in two of the three species and found that rapid expression differentiation was restricted to low-expression proteins. Our results suggest that various constraints on high-expression proteins reduce the availability of beneficial expression variants relative to
low-expression proteins, enabling low-expression proteins to evolve, and potentially lead to adaptation, more rapidly.
The expression level of a protein is one of the strongest predictors of protein sequence evolutionary rate; sequences of highly expressed proteins evolve more slowly than low-expression proteins (Duret and Mouchiroud 1999; Pal et al. 2001; Gout et al. 2010; Nabholz et al. 2013; Park et al. 2013; Yang et al. 2012). This relationship may be a function of specific selective constraints on sequences to avoid protein misfolding (Drummond et al. 2005; Geiler-Samerotte et al. 2011), protein misinteractions (Yang et al. 2012), a decrease in protein function (Gout et al. 2010; Cherry 2010), and/or mRNA misfolding (Park et al. 2013). Analyses of microarray data have shown that expression evolutionary rate is also negatively correlated with expression level across human and mouse orthologs (Liao and Zhang 2006). Although the selective constraints imposed on the sequences of highly expressed proteins are well-documented (Zhang and Yang 2015), the mechanistic basis of the negative correlation between expression evolutionary rate and expression level remains unclear (Liao and Zhang 2006). This relationship between expression level and expression evolutionary rate has only been documented when comparing orthologous genes across species with relatively long divergence times (e.g., human and mouse diverged approximately 100 million years ago; (Liao and Zhang 2006)). Long-term evolutionary patterns, however, are often unpredictable/intractable because of stochastic environmental fluctuations and irregular ecological changes (Grant and Grant 2002). Microevolutionary processes, on the other hand, can often be predicted simply based on selection and heritability (Grant and Grant 2002) and, therefore, present an opportunity to identify the mechanisms of divergence that often become lost in macroevolutionary patterns (e.g., species differences versus speciation-generating changes (Coyne and Orr 2004)). Although this relationship between expression level and expression evolutionary rate has been documented as a long-term evolutionary pattern (Liao and Zhang 2006), whether this relationship holds for traits evolving under directional selection within a single species over ecological timescales (i.e., microevolution; <5,000 years) is unknown.

Expression is a metabolically costly process, requiring energy for transcription, transla-
tion, and the mobilization of the translational machinery (DEKEL and ALON 2005; GOUT et al. 2010). The expression level of a protein is predicted to be a trade-off between the beneficial effects of its function and the energetic costs of its expression (GOUT et al. 2010; CHERRY 2010), and selection should drive the expression level of all proteins close to values that maximize fitness (NABHOLZ et al. 2013; DEKEL and ALON 2005). Although most proteins will be expressed near their optimal levels, abundantly expressed proteins should be highly optimized due to the increased energetic cost of production (GOUT et al. 2010; VISHNOI et al. 2010). Stabilizing selection on expression level should, therefore, be stronger for high-expression proteins, and this constraint should reduce the amount of standing expression variation for high-expression proteins relative to low-expression proteins. High-expression proteins could also eventually reach an upper bound on expression due to physiological and biophysical constraints (i.e., only so much of a particular protein can be made per cell or tissue). Because high-expression proteins are closer to the upper bound than low-expression proteins, the range of beneficial expression variation available to these abundant proteins is reduced. We may, therefore, expect that over short, ecological timescales, adaptive divergence in expression level would be limited to low-expression proteins because of reductions in the standing expression variation for high-expression proteins due to stabilizing selection and the range of beneficial expression variation available to high-expression proteins because of the upper-bound constraint.

Rapid adaptation is often associated with strong directional selection in novel environments following colonization events or dietary changes (REZNICK and GHALAMBOR 2001). Therefore, most studies of adaptation on ecological timescales involve translocations, trait manipulations, or other perturbations to assess the speed of adaptation (REZNICK and GHALAMBOR 2001; FRASER et al. 2011) and may not accurately reflect natural conditions. Focusing on populations inhabiting young barrier islands, however, could alleviate these concerns because of the colonization of a novel environment, changes in resource availability, and potentially limited gene flow (KOLBE et al. 2012; DOLEY et al. 2008; VINCENT et al. 2009;
Comparative studies of sympatric taxa can identify which evolutionary processes produced the observed patterns of differentiation (Gomulkiewicz et al. 2007). Several genera of North American pit vipers inhabit barrier islands of the southeastern United States, making them ideal for studying rapid adaptation.

Snake venoms are comprised of approximately 20 to 100 toxic peptides and proteins (Calvete et al. 2010; Margres et al. 2014; Margres et al. 2015) that collectively function in predation and defense. Although most quantitative traits are the products of developmental pathways where changes in expression level may have effects mediated through complex interaction networks, toxin expression variation directly changes the phenotype because relative amounts of venom components determine venom efficacy. Expression is typically measured at the mRNA level (Rokyta et al. 2012; Margres et al. 2013; Rokyta et al. 2013). The proteome, however, is more representative of the actual phenotype (Diz et al. 2012), particularly for venoms (Casewell et al. 2014), and the specialization of the venom-gland makes venom genetically tractable (Margres et al. 2014; Margres et al. 2015). Because venom is a secretion (Gibbs et al. 2009), protein expression can be directly measured by reversed-phase high-performance liquid chromatography (RP-HPLC).

To determine whether rapid expression evolution was restricted to low-expression proteins because of constraints on highly expressed proteins, we compared protein expression across mainland and island populations for three species of pit vipers native to the southeastern United States: the eastern diamondback rattlesnake (Crotalus adamanteus), the pygmy rattlesnake (Sistrurus miliarius), and the cottonmouth (Agkistrodon piscivorus). Because of the young age of the islands (e.g., <5,000 years, see Materials and Methods; Lopez and Rink 2006) and the selective constraints potentially limiting the evolvability of highly expressed proteins, we predicted that abundant proteins would exhibit less differentiation in expression than low-expression proteins, with low-expression proteins exhibiting patterns of rapid, adaptive, differentiation.
MATERIALS AND METHODS

Sampling: We collected venom and blood samples from eight *C. adamanteus*, 10 *S. miliarius*, and 12 *A. piscivorus* from the Florida mainland and 11 *C. adamanteus*, 18 *S. miliarius*, and 10 *A. piscivorus* from St. Vincent, Little St. George, and St. George islands (Fig. 1). These islands are Holocene formations (<5,000 years old) located in the Gulf of Mexico seven km from the mouth of the Apalachicola River Delta (Lopez and Rink 2006). All *C. adamanteus* were used in the analyses of Margres et al. (2015). We recorded sex, snout-vent length, and total length for each individual. We limited our analyses to adults to avoid the potentially confounding effects of ontogenetic protein expression variation, which has been previously documented (Mackessy 1988; Calvete et al. 2010; Durban et al. 2013; Margres et al. 2015). Samples were collected under the following permits: Florida Fish and Wildlife Conservation Commission (FWC) LSSC-13-00004 and LSSC-09-0399 and St. Vincent National Wildlife Refuge Permit 41650-2012-08. The above procedures were approved by the Florida State University Institutional Animal Care and Use Committee (IACUC) under protocols #0924 and #1333.

Transcriptomic Analysis: We followed the exact approach of Rokyta et al. (2013) for transcriptomic assembly and analysis. Briefly, we performed two *de novo* assemblies using the Extender program with 1,000 merged reads and three additional *de novo* assemblies using NGen. We identified and annotated toxin sequences following blastx searches against the NCBI non-redundant protein database. Only complete protein-coding sequences were retained. All raw reads were deposited in the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA), and the toxin transcripts were deposited in the NCBI Transcriptome Shotgun Assembly (TSA) databases.

Mass Spectrometry Analysis: Chromatographic separation and tandem mass spectrometry of the 25 RP-HPLC *C. adamanteus* peaks have been analyzed previously (Margres
et al. 2014; Margres et al. 2015). Thirty-five RP-HPLC peaks for *S. miliarius* and 42 RP-HPLC peaks for *A. piscivorus* were collected as previously described (Margres et al. 2014; Margres et al. 2015). Briefly, samples were run in triplicate on an externally calibrated Thermo LTQ Orbitrap Velos nLC-ESI-LIT-Orbitrap. Tandem mass spectra were extracted by Proteome Discoverer version 1.4.0.288 and Sequest (Thermo Fisher Scientific, San Jose, CA, USA; version 1.4.0.288) was used to search the species-specific transcriptome databases with signal peptides removed and assuming the digestion enzyme trypsin, allowing one missed digestion site. Scaffold version 4.3.2 (Proteome Software Inc., Portland, OR) was used to validate MS/MS-based peptide and protein identifications for both species. Peptide identifications were accepted if they could be established at greater than 95% probability by the Scaffold local FDR algorithm, contained at least two identified peptides, and possessed unique peptide evidence. To identify the major toxins within each peak, only proteins with >10% of the total spectral matches within each peak were reported. All data for *S. miliarius* and *A. piscivorus* are shown in Table S5. The raw proteomic data for *C. adamanteus* has been previously published (Margres et al. 2014; Margres et al. 2015).

**Protein Quantification and Statistical Analyses:** Reversed-phase high-performance liquid chromatography was performed on a Beckman System Gold HPLC (Beckman Coulter, Fullerton, CA) equipped with Beckman 32 Karat Software Version 8.0 for peak quantification as described by Margres et al. (2014, 2015) for 30 µg of total protein for *S. miliarius* (Fig. S1A), 100 µg of total protein for *C. adamanteus* (Fig. S1B), and 50 µg of total protein for *A. piscivorus* (Fig. S1C). The raw data is contained in Table S6. This approach produces compositional data subject to constant-sum constraints and inherently biased toward negative correlation among components (Aitchison 1986). Therefore, we followed the approach of Margres et al (2015) and Wray et. al (2015) and used centered logratio (clr) and isometric log ratio (ilr) transformations (Egozcue et al. 2003), when appropriate, to transform the data using the robCompositions package (Templ et al. 2011) in R prior to sta-
stistical analyses (Filzmoser et al. 2009). We used the multiplicative replacement strategy (Martin-Fernandez et al. 2003) implemented in the R package zCompositions assuming a detection threshold of 0.01% (the smallest measured value) to resolve the issue of zeros. We used the adonis function from the vegan package (Oksanen et al. 2007) in R and euclidean distances to perform a permutational or non-parametric MANOVA (McArdle and Anderson 2001) on the ilr-transformed data to test for significant protein expression variation as previously described (Margres et al. 2015). To determine if the variation detected was restricted to low-abundance proteins and highly expressed proteins were conserved, we divided the RP-HPLC peaks for each species into low-abundance and high-abundance data sets prior to conducting the non-parametric MANOVA as described above. Here, we first calculated the percent mean for each peak and then clr transformed these values. If the mean for an individual peak was less than the geometric mean, it was classified as low-expression. If the mean for an individual peak was greater than the geometric mean, it was classified as highly expressed. All statistical analyses of high-expression and low-expression were performed following this approach. For Fig. 2, however, we first clr transformed the raw percentage data for all samples and then calculated the mean directly from the clr values because this allowed us to estimate the standard error shown in the figure. We performed a linear discriminant function analysis using the lda function in R on the ilr-transformed data for each species to assess group membership placement probabilities across populations as previously described (Margres et al. 2015).

DNA Sequencing: Sistrurus miliarius and A. piscivorus DNA was extracted from whole blood samples drawn from the caudal vein using the Omega bio-tek E.Z.N.A Tissue DNA Kit according to the manufacturer’s protocol. A fragment of cytochrome b, 841 bp and 1,003 bp respectively, was amplified in 25 µL PCR runs using the H16064 and L14910 primers and thermal cycling protocol described by Burbrink et al. (2000). A 1,018 bp fragment for C. adamanteus (accessions KJ730289, KJ730300, KJ730314, KJ730321, KJ730327, KJ730342,
KJ730344, KJ730345, KJ730351, KJ730357, KJ730358, KJ730362, KJ730364, KJ730370, KJ730377, KJ730383, KJ730389, KJ730394, KJ730396) and C. horridus (outgroup; accession KJ730366) were taken from Margres et al. (2015) following the same approach. PCR products were purified using the QIagen QIAquick PCR Purification Kit and sequencing was on the Applied Biosystems 3730 Genetic Analyzer.

All individuals from both populations across all species were used in species-specific phylogenetic analyses. For each species, sequences were aligned using the MegAlign module of the DNAStar Lasergene version 11 software suite. Model selection was performed using jModeltest v. 0.1.1 under default settings (Guindon and Gascuel 2003; Darriba et al. 2012), with the Akaike information criterion used to determine the most appropriate model for each species (Akaike 1974). A maximum likelihood analysis was run in PAUP* 4.0b10 (Swofford 1998) using a heuristic search with 100 stepwise random-addition sequence replicates using the tree bisection-reconnection method. To assess support for the ML tree, we also performed a non-parametric bootstrap analysis using 1,000 pseudoreplicates with 10 stepwise random-addition sequence replicates. Base frequencies, rate matrix, proportion of invariable sites, and shape were estimated from the data.

RESULTS AND DISCUSSION

Rapid Differentiation in Protein Expression Following the Colonization of an Island: We performed a non-parametric MANOVA comparing protein expression levels across island and mainland populations separately for each species (Fig. 1) and detected significant expression differentiation in S. miliarius ($P<0.01$) and C. adamanteus ($P = 0.05$), but not A. piscivorus ($P = 0.93$). This significant differentiation in expression represents rapid (i.e., <5,000 years) phenotypic divergence in S. miliarius and C. adamanteus. As previously mentioned, comparative studies of sympatric taxa can help identify which evolutionary processes produced the observed patterns of differentiation (Gomulkiewicz et al. 2007). The lack of expression differentiation in A. piscivorus highlights the significance of the differentiation
in *S. miliarius* and *C. adamanteus* and suggests that different evolutionary processes are responsible for these different patterns. The geographic variation in protein expression for *C. adamanteus* and *S. miliarius* is consistent with local adaptation as a result of variable selective pressures due to genotype-by-genotype-by-environment interactions, and the lack of expression variation in *A. piscivorus* may be a result of diffuse selection due to its generalist diet (VINCENT et al. 2004) or high levels of gene flow (but see below). We next examined whether this differentiation in expression level for *C. adamanteus* and *S. miliarius* was biased towards high- or low-expression proteins.

**Expression Differentiation and Variation is Constrained to Low-Expression Proteins:** To determine whether the variation detected in *S. miliarius* and *C. adamanteus* was restricted to low-abundance proteins and highly expressed proteins were conserved, we divided the RP-HPLC peaks for each species into low-abundance and high-abundance data sets based on the centered logratio (clr) transformed mean for each peak and conducted a non-parametric MANOVA as described above. If the mean for an individual peak was less than the geometric mean, it was classified as low-expression. If the mean for an individual peak was greater than the geometric mean, it was classified as highly expressed. Of the 25 peaks in *C. adamanteus* and 28 peaks in *S. miliarius*, 13 and 14 were classified as low-expression proteins, respectively. We detected significant expression variation only in the low expression data sets for both *S. miliarius* (*P*$_{\text{Low}} < 0.01$, *P*$_{\text{High}} = 0.25$) and *C. adamanteus* (*P*$_{\text{Low}} = 0.05$, *P*$_{\text{High}} = 0.29$). As expected, neither class exhibited significant variation in *A. piscivorus* (*P*$_{\text{Low}} = 0.92$, *P*$_{\text{High}} = 0.83$).

We next looked at the covariance matrix of the clr-transformed data sets to identify the most variable peaks relative to the classification of each protein as high- or low-expression as described above. Low-expression proteins accounted for 95.4% of the variance in *C. adamanteus*, 86.2% of the variance in *A. piscivorus*, and 57.7% of the variance in *S. miliarius* (Supplementary Table S1), indicating that differentiation in protein expression and/or
standing expression variation, especially in *C. adamanteus*, was restricted to low-expression loci. The lower proportion of variance accounted for by low-expression proteins in *S. miliarius* may have been a reflection of the high amount of variation detected in peak 26 (16.1%; Supplementary Table S1). This peak was classified as highly expressed, but had the lowest expression level of any highly expressed protein.

We then compared the clr mean for each RP-HPLC peak across mainland (x-axis) and island (y-axis) populations for all three species (Fig. 2). We calculated the coefficient of determination \( (R^2) \) and found a good fit for all three species in high-expression proteins \( (R^2_{\text{Cadam}} = 0.785, R^2_{\text{Smili}} = 0.669, R^2_{\text{Apisc}} = 0.785) \), indicating that high-expression proteins were conserved across populations in all three species. Consistent with our previous analyses, low-expression proteins exhibited significant differentiation in *C. adamanteus* \( (R^2 = 0.019) \) and *S. miliarius* \( (R^2 = 0.328) \), but not *A. piscivorus* \( (R^2 = 0.746) \).

To determine if the increased variance in low-expression proteins was biological or a result of technical biases, we conducted six RP-HPLC analyses on a single venom sample from a mainland *C. adamanteus* and plotted the clr mean (x-axis) and variance (y-axis) for all RP-HPLC peaks (Fig. S2). If the increased variance in low-expression proteins was because of a limitation of our approach to accurately quantify low-abundance peaks, we would expect to see a significant negative correlation (i.e., a substantial reduction in variance as expression increased). We found a lack of correlation between the clr mean and variance among all peaks \( (R^2 = 0.0426, R = -0.2063, P = 0.3224; \text{Fig. S2a}) \), and even less so following the removal of a single, outlying low-abundance peak \( (R^2 = 0.0052, R = -0.0720, P = 0.7380; \text{Fig. S2b}) \); \( R^2 \) is the coefficient of determination and \( R \) is Pearson’s correlation coefficient), indicating that the increased variance in low-expression proteins was not an artifact of our method but rather biological and consistent with our expectations for stabilizing selection.

Overall, abundant proteins exhibited significantly less differentiation in expression than low-expression proteins, with the latter exhibiting patterns of rapid differentiation. These results are consistent with previous work that found a negative correlation between expres-
sion evolutionary rate and expression level between human and mouse orthologs (Liao and Zhang 2006). Our results, however, demonstrate that this pattern holds over ecological timescales as well as 100 million years of divergence. We predicted that stabilizing selection on expression level should be stronger for high-expression proteins and reduce the amount of standing expression variation relative to low-expression proteins. We found strong evidence supporting this expectation (Fig. 2), especially in *A. piscivor us* (Fig. 2C), despite a lack of expression differentiation. Our results suggest that the expression level of highly expressed proteins evolve under considerable constraints, potentially because these proteins are already expressed at or near their physiological maxima, and the expression level of a protein is a strong predictor of both protein sequence and protein expression evolutionary rates. These constraints may limit the range of beneficial expression variation available to high-expression proteins, indicating that rapid, adaptive divergence would be restricted to low-expression proteins over ecological timescales.

**The Process Driving Expression Differentiation in Low-Expression Proteins:** The differentiation in protein expression detected may be a result of genetic drift following a relaxation of evolutionary constraints (Khan et al. 2013), selection following the colonization of a novel environment (Fraser et al. 2011), founder effects (Kolbe et al. 2012), and/or phenotypic plasticity (Hunt et al. 2011).

We did not specifically test for plasticity in this study because venom expression differences have been repeatedly shown to be under genetic control and not environmentally induced (Daltry et al. 1996; Margres et al. 2015; Gibbs et al. 2011; Holding et al. 2015), and the feeding ecology of venomous snakes makes adaptive plasticity unlikely. Because venom is stored for long periods of time (e.g., over winter and between infrequent feeding events) and previous meals may not be robust predictors of future meals, plasticity would be unlikely to provide any adaptive advantage for this trait. Gibbs et al. (2011) fed different groups of *S. miliarius* different prey items over extended time periods and did not
find any significant changes in venom expression, Margres et al. (2015) documented the ontogenetic shift in venom expression in *C. adamanteus* in lab-raised individuals and found that geographic differences in venom expression held over long periods of time despite the animals being raised under identical conditions in captivity, and Holding et al. (2015) recently showed that prey preference is also under genetic control and not affected by previously fed upon items in *S. miliarius*. Because these studies failed to identify any plastic changes in venoms in the two species in our study that exhibited significant expression differentiation, we rejected the hypothesis that plasticity played any role in generating the observed variation in venom expression.

Although relaxed purifying selection on protein expression levels is believed to be rare, a recent study by Khan et al. (2013) proposed a theoretical framework for determining whether expression differentiation was consistent with directional selection or the relaxation of evolutionary constraints. The authors stated that a shift in mean expression level associated with high within-lineage variation is indicative of drift following a relaxation of constraints, and a shift in mean expression level associated with low within-lineage variation is indicative of directional selection within that particular population. The non-parametric MANOVA identified mean differences in the expression levels of low-expression proteins between island and mainland *C. adamanteus* and *S. miliarius* populations, and we again used the covariance matrix of the clr-transformed data sets to examine within-population variances across island and mainland populations for both species. According to Khan et al. (2013), if the variance for a particular protein or class of proteins is greater within the mainland population, this would suggest directional selection in the island population. To compare the magnitude of the differences in variances across island and mainland populations, we calculated the total variances across all peaks in both populations for all three species. For the two species with significant mean expression differences (*C. adamanteus* and *S. miliarius*), the total variances across all peaks were much larger in the mainland populations relative to the island populations (313 mainland vs. 228 island in *C. adamanteus*, 341 mainland vs. 15
239 island in *S. miliarius*). We did not find a significant mean expression difference for *A. piscivorus*, and the total variances across all peaks were nearly identical (204 mainland vs. 192 island). In *C. adamanteus*, the total variance for low-expression proteins was also greater within the mainland population than within the island population (301 mainland vs. 214 island), and nine of the 13 low-expression proteins had greater variances within the mainland population than within the island population (although this frequency was not significantly different than what we would expect by chance; *P*=0.17; Fig. 2). Similarly in *S. miliarius*, the total variance for low-expression proteins was greater within the mainland population than within the island population (199 mainland vs. 128 island), and 11 of the 14 low-expression proteins had greater variances within the mainland population than within the island population (*P*=0.03; Fig. 2). Again, in *A. piscivorus* the total variances for low-expression proteins across mainland and island populations were nearly identical (172 mainland vs. 171 island). We next performed the multivariate analog to a Levene’s test to test for homogeneity of group variances dispersions across island and mainland populations for all peaks, high-expression peaks only, and low-expression peaks only. To determine if the variances of the island and mainland populations were significantly different, we ran an ANOVA to compare the distances of group members (e.g., an island individual) to the group centroid (e.g., island population centroid) across populations. We failed to identify any significant differences in these comparisons, although the low-expression analysis did approach significance in *S. miliarius* (*P*= 0.07) and we typically saw a reduction in the *P*-value in the low-expression comparisons relative to the high-expression only analyses (e.g., in *C. adamanteus* *P*<sub>High</sub> = 0.69, *P*<sub>Low</sub> = 0.32). Power analyses suggested that these non-significant results, however, may have been the result of a small sample size; for a one-way ANOVA comparing two groups with a moderate effect size (*f* = 0.25) and a significance level of 0.05, we would need a shared sample size of approximately 63 individuals to obtain a power (i.e., confidence) of 0.80. Our largest shared sample size was *n* = 10 (i.e., 10 island individuals and 10 mainland individuals) in *S. miliarius* and *A. piscivorus*. The shared sample size in
*C. adamanteus* was *n* = 8. Even when assuming a large effect size (*f* = 0.40) and reducing the power to 0.75, we still have less than half the adequate sample size required (*n* = 23) to detect a significant result.

The significant differentiation in mean protein expression across island and mainland populations along with the reduction in intra-island expression variation could also be explained by an alternative selection model where selection would be maintaining diversity on the mainland rather than driving expression differentiation on the island. If we assume that dietary variation on the mainland is much larger than on the island, a reasonable assumption because islands typically exhibit a significant reduction in species diversity (*MacArthur* and *Wilson* 1967), selection could maintain the higher variation in venom expression on the mainland because of the higher variation in diet. Therefore, under this hypothesis, the significant mean differences in expression between island and mainland populations would be the result of founder effects, and the larger variance on the mainland would be the result of diversifying selection. This alternative hypothesis, however, would require some sort of biogeographic structure in the mainland population for selection to maintain this expression variation. Our sampling is from a very small region of contiguous and uniform habitat. The only potential biogeographic barriers on the mainland are the Ochlockonee and Apalachicola Rivers. Although the latter has been repeatedly documented as a biogeographic barrier to a number of organisms (*Burbrink* *et al.* 2000; *Baer* 1998), the former has never been known, to the best of our knowledge, to impede gene flow in squamates, and the majority of our sampling was west of the Apalachicola River (i.e., only 4 of the 10 mainland *S. miliarius* were collected west of the Apalachicola River). Additionally, the island system sits at the mouth of the Apalachicola River, and immigration is equally likely to have occurred from either side of the Apalachicola River. Margres *et al.* (2015) recently sequenced a 986-bp fragment of ND5 and a 1018-bp fragment of cytochrome *b* for *C. adamanteus* and found two haplotypes in panhandle Florida. Both haplotypes were present in island and mainland populations and were found on either side of both rivers with no obvious frequency differences across the
Ochlockonee River. Considering that we failed to detect any population structure within the mainland population in *C. adamanteus* when sequencing two loci with dense sampling (i.e., 70 individuals in panhandle Florida), we find it reasonable to assume a lack of population structure within the mainland populations for all species and, therefore, that neither river is a barrier that could restrict gene flow to the degree necessary for this alternative model.

Overall, our results are most consistent with the predictions of Khan et al. (2013) for directional selection in the island populations of *C. adamanteus* and *S. miliarius*. Lower within-population variance in the island populations, however, is not surprising given the likelihood of a smaller effective population size than on the mainland, and this may reflect founder effects rather than directional selection.

Founder effects, genetic or phenotypic changes in a population as a result of being initially colonized by relatively few individuals, can cause divergence among populations, particularly island populations (Kolbe *et al.* 2012; Spurgin *et al.* 2014). Although the variance analysis suggested directional selection produced the expression differentiation identified in low-expression proteins across populations (Khan *et al.* 2013), this divergence in phenotype along with the lack of within-population variance could be a result of the island being initially colonized by a small number of individuals. To determine if the identified differentiation in protein expression between island and mainland populations was a result of demographic histories, we sequenced cytochrome *b* for all sampled individuals. A reduction in genetic diversity within the island populations relative to the mainland populations would indicate founder effects, while similar genetic diversity across island and mainland populations would indicate a lack of founder effects due to multiple colonization events and/or continuous gene flow. We found a complete lack of genetic diversity across *C. adamanteus* island and mainland populations (0.0% sequence divergence), a single variable site across *S. miliarius* island and mainland populations (0.1% sequence divergence; Fig. S3), and five variable sites across *A. piscivorus* island and mainland populations (0.3% sequence divergence; Fig. S4). We did identify a monophyletic clade of island *A. piscivorus*, although not all island specimens
were in this clade. Determining the demographic histories of these populations was difficult because of the lack of genetic diversity within *C. adamanteus* and *S. miliarius*. This absence of genetic variation across island and mainland individuals could indicate ongoing gene flow (i.e., a lack of founder effects), or could be a result of the young age of these island populations (i.e., 5,000 years was not enough time for neutral differentiation to occur). Identifying a genetic bottleneck following a founding event has been known to be sensitive to the number of loci examined, and only using a single locus can result in a type II error (Spurgin et al. 2014). Therefore, we currently cannot rule out founder effects, although strong selection can overwhelm founder effects over ecological timescales (Kolbe et al. 2012), and the only species that exhibited any degree of neutral population differentiation (i.e., *A. piscivorus*) did not significantly differ in mean expression.

Demonstrating that this expression differentiation was a result of selection, and not founder effects or another neutral process, would require fitness comparisons across populations. In the absence of fitness data, comparing the differentiation of traits under putative selection to that of neutral markers may allow the identification of adaptive variation (Savolainen et al. 2013). The differentiation in protein expression among *S. miliarius* and *C. adamanteus* populations that can be accounted for by divergence at neutral markers may reflect neutral processes, but variation that exceeds this neutral divergence may be indicative of directional selection (Whitehead and Crawford 2006; Richter-Boix et al. 2010; Margres et al. 2015). To determine if the identified differentiation in protein expression between island and mainland populations was a result of adaptive or neutral processes, we used the sequence data discussed above to compare neutral divergence and protein expression differentiation. The lack of neutral differentiation and the significant phenotypic divergence across *S. miliarius* and *C. adamanteus* populations suggests that the observed protein expression differentiation was a result of directional selection (Margres et al. 2015), potentially in spite of gene flow, although founder effects cannot currently be ruled out.

Expression variation is typically attributed to *cis*-regulatory mutations (Carroll 2008).
Cis-regulatory mutations, however, are not the only mechanism for altering the amounts of protein produced (Hastings et al. 2009). Polymorphisms at a much larger genomic scale, such as gene duplications and deletions (Stranger et al. 2007), can also alter the expression level of a particular protein (Nguyen et al. 2006). The correlation between gene copy number differences and changes in gene expression has been previously documented (Cheng et al. 2005; Freeman et al. 2006; Nair et al. 2008), including in venoms (Margres et al. 2015), and venom protein families are believed to be the result of gene duplication and positive selection (Casewell et al. 2011) via the birth-and-death model of protein evolution (Fry et al. 2008). The significant expression variation we detected, therefore, could be the result of variation in copy number, assuming that variation in copy number would affect low-expression (and presumably low-copy) genes more than high-expression, high-copy genes (e.g., the difference between 10 copies and 12 copies for a particular protein may not be significant, but the difference between 2 copies and 4 copies may be). Genomic drift (i.e., the random duplication and deletion of genes) has been shown to play an important role in generating copy-number variation (Nozawa et al. 2007; McCarroll et al. 2008; Nei et al. 2008), and it has even been proposed that neutral processes are responsible for maintaining the vast majority of all identified variation in copy number (Nozawa et al. 2007). Genomic drift, however, is stochastic and would only affect our estimates of variance, not the mean. Therefore, regardless of the mechanism (e.g., cis-regulatory mutation, copy-number variation, miRNA regulation, translational efficiency), our results suggest that the identified expression differentiation and variation was the result of selection rather than neutral processes, although, again, founder effects cannot be ruled out.

The Rate of Fixation of Expression Levels: The significant variation detected in C. adamanteus and S. miliarius demonstrated that, on average, expression levels for low-expression proteins differed between island and mainland populations. To determine if this protein expression variation was fixed within each population, we used a linear discrimi-
nant function analysis to assess group membership placement probabilities (Margres et al. 2015). If the expression phenotypes have been fixed in the island populations, we would expect placement probabilities near 100% for island S. miliarius and C. adamanteus. This analysis, however, is problematic if the sample size (i.e., number of individuals per species) does not exceed the number of variables (i.e., RP-HPLC peaks). For S. miliarius, the number of variables equaled the sample size \( n = 28 \) while the number of variables \( \text{var} = 25 \) exceeded the sample size \( n = 19 \) in C. adamanteus. Therefore, we performed the analysis on the low-abundance and high-abundance data sets independently. Based on our previous analyses, we would expect relatively low placement probabilities for the high-abundance data sets and higher placement probabilities for the low-abundance data sets given that low-expression proteins explained the majority of the variation in our data.

Analysis of the high-abundance data sets accurately assigned 45.5% of island and 50.0% of mainland C. adamanteus and 77.8% of island and 70.0% of mainland S. miliarius. Analysis of the low-abundance data sets accurately assigned 81.2% of island and 75.0% of mainland C. adamanteus and 77.8% of island and only 60.0% of mainland S. miliarius. The slightly lower placement probability for mainland S. miliarius in the low-expression data set may have been a reflection of the high amount of variation detected in high abundance peak 26 as discussed above. We did see, however, a significant improvement in placement probabilities for C. adamanteus in the low-abundance data set as expected. These placement probability percentages, although indicative of significant population differentiation in the expression of low-expression proteins, demonstrated that the island expression patterns were not yet fixed in either species. Because the island phenotypes were not fixed, founder effects (discussed above) were unlikely to cause the observed differentiation in expression levels because following a severe bottleneck with a relatively short recovery period (<5,000 years old), we would expect near fixation of the expression phenotype. Therefore, our results suggest that these expression differences were the result of directional selection. Local adaptation is predicted to act as a barrier to migration due to reduced immigrant fitness, and subsequent genetic
drift will eventually result in neutral genetic structure across populations (Spurgin et al. 2014). The age of the islands (<5,000 years old), however, may be insufficient to allow the completion of this process. Therefore, this isolation by adaptation (Spurgin et al. 2014) may be incipient, and the lack of expression level fixation and neutral differentiation despite, on average, significant variation in low-abundance protein expression, may simply reflect the young age of the island populations. This could also indicate ongoing gene flow in these two species, which is predicted to increase the probability of successful establishment and persistence in novel environments (Forsman 2014) as well as potentially promote local adaptation in coevolutionary contexts (North et al. 2010).

**Protein Identification:** To identify the individual proteins present in each RP-HPLC peak, we used the approach of Margres et al. (2014, 2015) to correlate specific toxin transcripts with specific venom proteins. We identified 122 and 157 unique putative toxin transcripts in the venom-gland transcriptomes of *S. miliarius* (GenBank Transcriptome Shotgun Assembly accession number GDBJ02000000) and *A. piscivorus* (GenBank Transcriptome Shotgun Assembly accession number GDAZ02000000), respectively, and these toxin transcripts were grouped into 63 and 76 clusters on the basis of <1% nucleotide divergence in their coding sequences as previously described (Rokyta et al. 2012; Rokyta et al. 2013; Margres et al. 2013; Margres et al. 2015). The venom-gland transcriptome for *C. adamanteus* was previously assembled and annotated (NCBI Sequence Read Archive accession number SRA050594, GenBank Transcriptome Shotgun Assembly accession number GBEX01000000; (Rokyta et al. 2012; Margres et al. 2014; Margres et al. 2015)); 76 unique putative toxin transcripts that grouped into 44 clusters were identified (Margres et al. 2015). Following transcriptome assembly and analysis, we identified unique proteomic evidence for 24 of the 63 *S. miliarius* toxin clusters (Table S2; Fig. S1A) and 30 of the 76 *A. piscivorus* toxin clusters (Table S3; Fig. S1C). Proteomic analysis of *C. adamanteus* venom was also previously described (Margres et al. 2014; Margres et al. 2015). We re-analyzed
this data using different parameters (see Materials and Methods) and identified 18 of the 44
*C. adamanteus* toxin clusters (Table S4; Fig. S1B). Table S5 contains peptide reports for *S.
miliarius* and *A. piscivorus*.

Our previous analyses demonstrated that high-expression proteins were conserved across
populations and that low-expression proteins exhibited significant expression variation. To
determine whether particular protein families were over- or underrepresented in high- (i.e.,
less variable) and low-expression (i.e., more variable) RP-HPLC peaks across all three
species (Table S1), we compared protein family presence/absence across expression classes.
Cysteine-rich secretory proteins (CRISPs) were only identified in low-expression peaks 16 and
17 in *A. piscivorus* (Table S3). These peaks were the fourth most and most variable peaks,
respectively, suggesting that CRISP expression was not only biased towards low-expression
in *A. piscivorus* but also exhibited the most variation within populations. Myotoxin (peaks
1b and 2) and a single phospholipase A2 protein (peak 10) were only identified in high-
expression peaks in *C. adamanteus*. These proteins were the third and second least variable,
respectively, indicating that these proteins were highly expressed with little variation. The
C-type lectin protein family was represented by a single protein in peaks 28, 29, and 32 in
*S. miliarius*. All of these peaks were highly expressed but possessed very different variances.
Peak 29 was the fifth least variable peak, peak 32 was the thirteenth least variable peak, but
peak 28 was the second most variable peak. Detecting the same toxic protein in multiple
peaks suggests that this protein undergoes post-transcriptional modifications (*Casewell
et al. 2014; Margres et al. 2015*), and the expression variation (or lack thereof) detected
in this protein appears to be post-transcriptional variant-specific. Overall, the lack of bias
towards a particular expression level for the most diverse protein families (e.g., snake venom
metalloproteinases, snake venom serine proteinases) indicated that expression variation was
locus-specific, or sometimes post-transcriptional variant-specific, rather than gene-family
specific.
CONCLUSION

We compared the rates of expression evolution for high- and low-expression proteins and found that, over ecological timescales, expression levels of abundant proteins were significantly conserved and rapid expression evolution was restricted to low-expression proteins. Our results are consistent with microarray studies examining human and mouse orthologs (Liao and Zhang 2006) and suggest that stabilizing selection on high-expression proteins reduced the amount of standing expression variation in these abundant proteins. This reduction in standing variation, in combination with the upper-bound constraint, limited the rate at which adaptive expression variation was generated in high-expression proteins. Therefore, the expression level of a protein is a strong predictor of both protein expression and protein sequence evolutionary rate. Gibbs et al. (2009) found that highly expressed proteins exhibited less variation in presence-absence variation than low-expression proteins in Sistrurus rattlesnakes. The authors speculated that highly expressed proteins performed generic killing functions and low-expression proteins were not only prey-specific, but also more evolvable, consistent with our findings. The greater evolvability of low-expression proteins relative to high-expression proteins may allow them to more rapidly respond to novel selective pressures, and although the optimal expression of these proteins is relatively low, the fitness effects of a regulatory-mutation affecting expression may be high (Gout et al. 2010).

Although our results were consistent with theoretical predictions (Gout et al. 2010) and previous work in mammalian systems (Liao and Zhang 2006; Zhang and Yang 2015), these findings contradicted previous work showing that venom loci expressed at all levels contributed to protein expression divergence among adult C. adamanteus (Margres et al. 2015). This study, however, used range-wide sampling with much older divergence times than the current study (i.e., 1.27 million years versus <5,000 years; (Margres et al. 2015)). Together, these results suggest that although both high and low-expression venom proteins exhibited significant expression variation over large spatial and temporal scales (Margres
et al. 2015), rapid expression evolution was confined to low-expression venom proteins. We may see different patterns of expression differentiation in venoms over different timescales because of the selective and physiological constraints acting on high-expression proteins. These constraints may reduce the potential of highly expressed proteins to generate beneficial sequence and expression variation for a given venom protein, suggesting that rapid, adaptive divergence would be restricted to low-expression venom proteins over ecological timescales. Given enough time, however, beneficial expression variation can arise in high-expression venom proteins (Margres et al. 2015). Our results suggest that, in the context of proteins evolving under directional selection, the initial steps in the adaptive process may be restricted to mutations affecting low-expression proteins due to constraints on highly expressed proteins, with expression differentiation in proteins expressed at higher levels occurring over larger temporal scales. Therefore, this microevolutionary bias in expression evolutionary rate may result in the long-term evolutionary pattern previously documented (Liao and Zhang 2006; Margres et al. 2015), suggesting that short-term processes can, at least occasionally, be extrapolated to a macroevolutionary level.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (DEB 1145987 to D.R.R.), Florida State University (to M.J.M.), and the Gopher Tortoise Council (to M.J.M.). Sanger sequences were submitted to the National Center for Biotechnology Information (NCBI) Trace Archive under accession numbers KP881369–KP881418. Annotated transcriptome sequences were submitted to the GenBank Transcriptome Shotgun Assembly database under accession number GDBJ02000000 for *S. miliarius* and GDAZ02000000 for *A. piscivorus*. We thank Pierson Hill, Jacob Loyacano, Joe Pfaller, Mark S. Margres, and Flavio Morrissiey for help in acquiring venom samples. We thank Megan Lamb, Danielle Jones, Jennifer Wanat, and Rebecca Bernard with the Florida DEP and Apalachicola River NERR and Bradley Smith and Shelley Stiaes with US Fish and Wildlife Service and St. Vincent NWR.
for access to field sites. Samples were collected under the following permits: Florida Fish and Wildlife Conservation Commission (FWC) LSSC-13-00004 and LSSC-09-0399, Florida Department of Environmental Protection - Permit #04101310, and St. Vincent National Wildlife Refuge - Permit #41650-2012-08. Sample collection was approved by the Florida State University Institutional Animal Care and Use Committee (IACUC) under protocols #0924 and #1333.

LITERATURE CITED


Casewell, N., S. Wagstaff, W. Wuster, D. Cook, F. Bolton, S. King, D. Pia,


Nair, S., B. Miller, M. Barends, A. Jaidee, J. Patel, M. Mayxay, P. Newton, F. Nosten, M. Ferdig, and T. Anderson, 2008 Adaptive copy number evolution in


ROKYTA, D. R., K. P. WRAY, and M. J. MARGRES, 2013 The genesis of an exceptionally


Vishnoi, A., S. Kryazhimskiy, G. Bazykin, S. Hannenhalli, and J. Plotkin, 2010
Young proteins experience for variable selection pressures than old proteins. Genome
Res. 20: 1574–1581.

Whitehead, A. and D. Crawford, 2006 Neutral and adaptive variation in gene expres-

Wray, K. P., M. J. Margres, M. Seavy, and D. R. Rokyta, 2015 Early significant

Yang, J., B. Liao, S. Zhuang, and J. Zhang, 2012 Protein misinteraction avoidance
causes highly expressed proteins to evolve slowly. P. Natl. Acad. Sci. USA 109(14): E831–
E840.

Figure 1: Sampling of island and mainland populations across three genera of pit vipers. We collected venom and blood samples from 19 *Crotalus adamanteus*, 28 *Sistrurus miliarius*, and 22 *Agkistrodon piscivorus* across island and mainland populations. Locations of the study sites within the state of Florida are indicated in the inset map.
Figure 2: Expression differentiation and variation was constrained to low-expression proteins. We plotted the centered logratio (clr) mean for each reversed-phase high-performance liquid chromatography peak across mainland (x-axis) and island (y-axis) populations for (A) *Crotalus adamanteus*, (B) *Sistrurus miliarius*, and (C) *Agkistrodon piscivorus*. High-expression proteins were highly-correlated across populations across all species, and low-expression proteins exhibited a much larger degree of differentiation and variance within populations, particularly for the two species that exhibited significant differentiation (A and B). The larger variance for low-expression proteins relative to high-expression proteins in *A. piscivorus* (C), despite a lack of population differentiation, was strong evidence supporting our expectation that strong stabilizing selection would reduce the amount of standing expression variation for high-expression proteins. Bars indicate standard error, the solid line indicates a perfect agreement, dashed lines indicate the origin (i.e., the geometric mean), and proteins less than these values were considered low-expression proteins.