Effects of geographic heterogeneity in species interactions on the evolution of venom genes

Dan Chang[1,2,$] , Amy M. Olenzek[1], Thomas F. Duda, Jr[1,3]

1. Department of Ecology and Evolutionary Biology and Museum of Zoology, University of Michigan, Ann Arbor, Michigan, 48109, USA

2. Department of Statistics, University of Michigan, Ann Arbor, Michigan, 48109, USA

3. Smithsonian Tropical Research Institute, Balboa, Ancón, Republic of Panama

$. Current address: 1156 High Street – Mail Stop EEBiology, University of California Santa Cruz, Santa Cruz, CA 95064

Supplementary Information

Figure S1-S5

Table S1-S10

Supplementary references

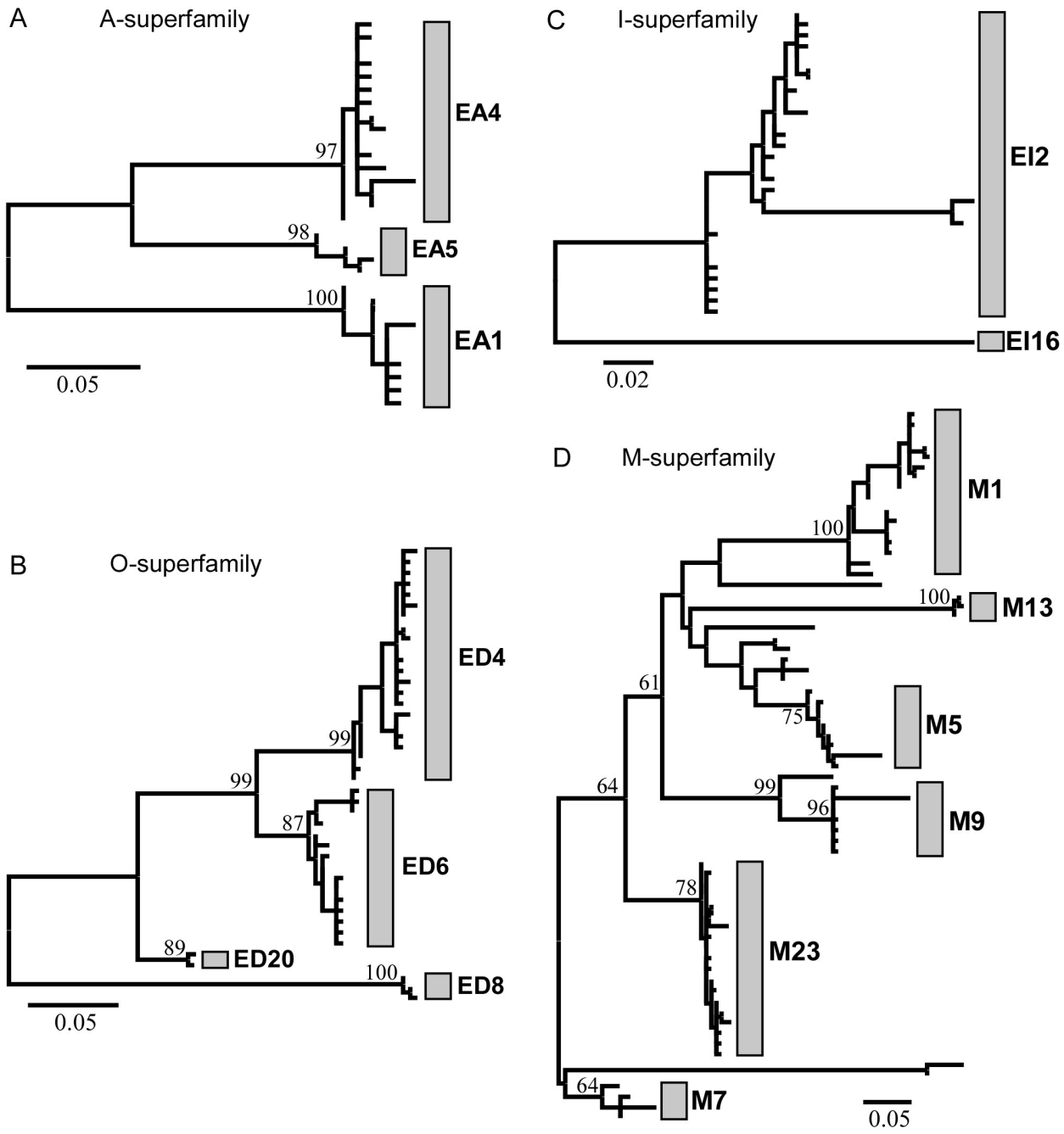


Figure S1. Gene trees of unique conotoxin gene mRNA sequences recovered from venom ducts of *C. ebraeus* individuals at three locations constructed using the Maximum-Likelihood estimation and mid-point rooting. Major clades labeled with grey bars are of putative single loci, and numbers on internal branches are bootstrap values of major clades (except for I-superfamily).

(A) Gene tree of 30 A-superfamily sequences recovered from two individuals at American Samoa, three at Guam and two at Hawaii, constructed with the Tajima 3-parameter [1] +G model. These sequences occur in three major clades ('EA1', 'EA4' and 'EA5') that we interpreted to represent three distinct loci. Sequences within clades differed at between one and five nucleotides (nt) (out of a total of 169-185 nt); sequences among clades differed at between 23 and 39 nt (out of a total of 160 nt).

(B) Gene tree of 45 unique O-superfamily sequences obtained from two individuals at American Samoa, five at Guam and two at Hawaii, constructed with the Tamura-Nei [2] +I model. These sequences fell into four major clusters ('ED4', 'ED6', 'ED8' and 'ED20') that we interpreted to represent four distinct loci. Sequences within clades differed at between one and ten nt (out of a total of 266-278 nt); sequences among clades differed at between 21 and 61 nt (out of a total of 266 nt).

(C) Gene tree of 22 unique I-superfamily sequences from two individuals at American Samoa, five at Guam and one at Hawaii, constructed with the HKY [3] model. These sequences are grouped into two major clades ('EI2' and 'EI16'). Sequences within clade EI2 differed at between one and 23 nt (out of a total of 229 nt); sequences between the two clades differed at between 43 and 64 nt (out of a total of 226 nt).

(D) Gene tree of 67 M-superfamily sequences from three individuals at American Samoa, six at Guam and two at Hawaii (amplified with the primer set MPr2 (Table S1)) constructed with the Tamura-Nei+G model. These sequences fell into more than six major clades. Sequences within clades (except clade 'M1') differed at between one and nine nt out of 217-233 nt while sequences among the six clades differed at between 29 and 59 nt out of 214 nt. Sequences of clade 'M1' differ at maximum of 20 nt, indicating the possibility that these sequences represent two loci. Out of the 37 colonies sequenced from two individuals at American Samoa, one at Guam and two at Hawaii, we only recovered seven sequences with the primer set MPr1 (Table S1; GenBank accession numbers JX177162 - JX177168). These sequences differed at a maximum of two nt (out of a total of 233 nt) and represent one putative locus.

Locus	Allele	Predicted Amino Acid Sequences
ED4	ED4a	K G **C** V Q T G G S **C** P S T T G **C** **C** N G L **C** N V D K **C** T
	ED4b	K R **C** V Q T G G S **C** P S T T G **C** **C** N G L **C** N V D K **C** T
	ED4c	K R **C** V Q T G S S **C** P S T T G **C** **C** S G L **C** N V N K **C** T
	ED4d	K G **C** V Q T G G S **C** P S T N G **C** **C** N G L **C** N V N R **C** A
	ED4e	- - - - - T G G S **C** P S T T G **C** **C** N G L **C** N V D K **C** T
	ED4f	- - - - - T G G S **C** P S T T G **C** **C** S G L **C** N V N K **C** T
	ED4g	- - - - - T G S S **C** P S T T G **C** **C** S G L **C** N V N K **C** A
	ED4h	- - - - - T G S S **C** P S T T G **C** **C** N G L **C** N V N K **C** T
	ED4i	- - - - - T G G S **C** P S T T G **C** **C** N G L **C** N V N K **C** T
ED6	ED6a	A **C** V N R G D P **C** Q R T V R **C** **C** S R R **C** G I N G **C**
	ED6b	R **C** V N S G D P **C** Q R T V R **C** **C** S R R **C** G I N G **C**
	ED6c	A **C** V N R G D P **C** Q R T V R **C** **C** S R R **C** S I N G **C**
	ED6d	A **C** V N R G D P **C** Q R T V R **C** **C** S R R **C** G V N G **C**
	ED6e	A **C** I N S G D P **C** Q R T V R **C** **C** S R R **C** G V N S **C**
	ED6f	A **C** V N R G D P **C** Q R T V R **C** **C** S R R **C** G V N S **C**
	ED6g	- - - - R G D P **C** Q R T V R **C** **C** S R R **C** S I N S **C**
E1	E1a	T H S G G A **C** N S H D Q **C** **C** N A F **C** D T A T R T **C** V
	E1b	T D S G G A **C** N S H D Q **C** **C** N E F **C** S T A T R T **C** I
	E1bii	T D S G G A **C** N S H D Q **C** **C** N E F **C** S T A T R T **C** I
	E1c	T R S G G A **C** Y S H N Q **C** **C** D D F **C** S T A T S T **C** V
	E1d	T R S G G A **C** N S H T Q **C** **C** D D F **C** S T A T S T **C** I
	E1e	T R S G G A **C** Y S H N Q **C** **C** D D F **C** S T A T S T **C** I
	E1f	T H S G G A **C** N S H D Q **C** **C** A N F **C** R K A T S T **C** M
	E1g	T R S G G A **C** N S H T Q **C** **C** D H F **C** S T A T S T **C** I
	E1h	T R S G G A **C** N S H D Q **C** **C** A N F **C** R K A T S T **C** M
ED20	ED20a	K G E P **C** N S S V P **C** **C** S G I **C** G Y F N **C** A
	ED20b	K G E P **C** N W S V P **C** **C** S G I **C** G Y F N **C** A
	ED20c	K G E P **C** N S S V P **C** **C** S G I * G Y F N **C** A
EA4	EA4a	R I A L I A T R E **C** **C** A N P Q **C** W A K N **C** R
	EA4b	R I A L I A T R E **C** **C** A N P Q **C** W G K N **C** R

Figure S2. **Alignment of predicted amino acid sequences of alleles of five conotoxin loci of *C. ebraeus*.** The cysteine backbone of each predicted peptide is highlighted in bold; amino acid replacements among alleles are highlighted in grey. *: stop codon. Because one of the locus-specific primers of locus ED4 could only be designed in the region where this site occurs, we do not have sequence data for all individuals at the first polymorphic site. The nucleotide composition of the first three segregating sites are not known for allele ED6g because the allele-specific primer for locus ED6 (Table S3) occurred in this region and allele ED6g is inferred from the sequence chromatogram obtained with this primer set.

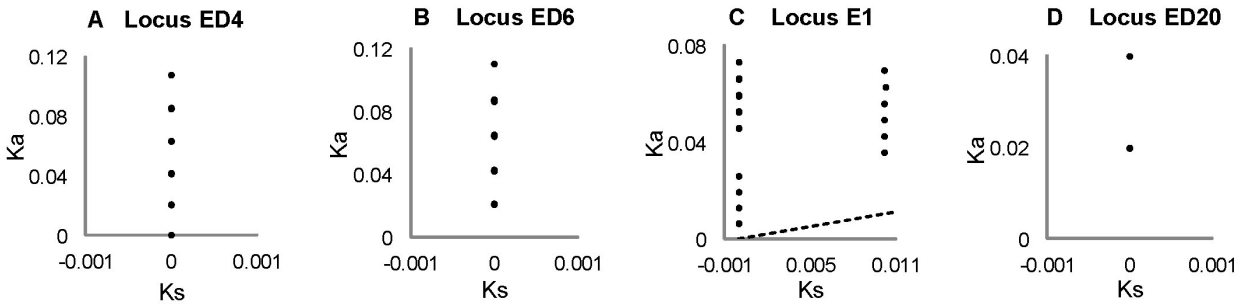
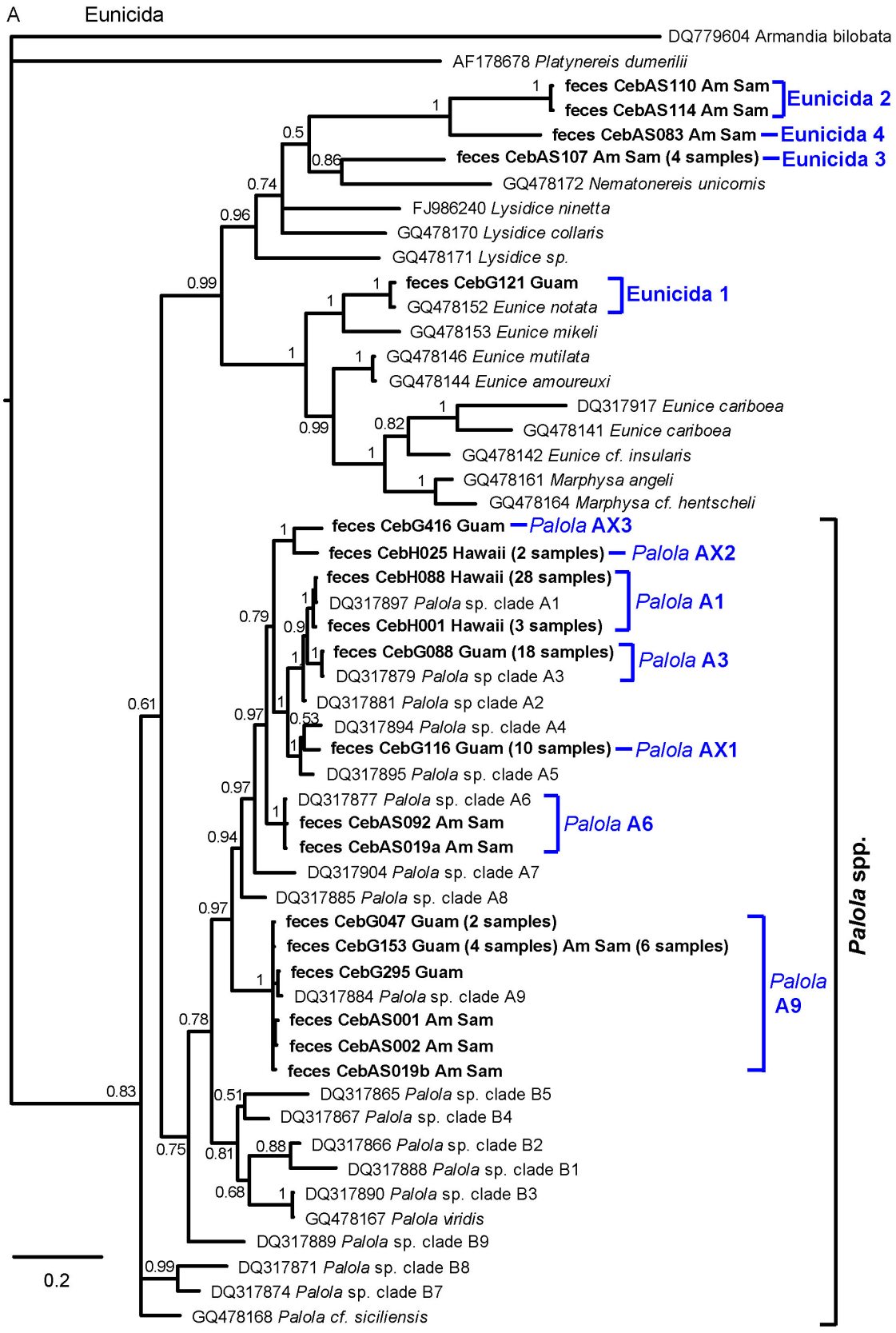


Figure S3. Pairwise Ka vs Ks values estimated based on allelic compositions of four conotoxin loci among populations of *C. ebraeus* at Hawaii, Guam and American Samoa. The dashed line in (C) has a slope of 1 (i.e. Ka=Ks), and all data points in this panel are above the dashed line. Ks of dots in the other panels are all 0.


A
Eunicida
DQ779604 Armandia bilobata
AF178678 Platynereis dumerilii
feces CebAS110 Am Sam
feces CebAS114 Am Sam
Eunicida 2
feces CebAS083 Am Sam
Eunicida 4
feces CebAS107 Am Sam (4 samples)
Eunicida 3
GQ478172 Nematonereis unicornis
FJ986240 Lysidice ninetta
GQ478170 Lysidice collaris
GQ478171 Lysidice sp.
feces CebG121 Guam
GQ478152 Eunice notata
Eunicida 1
GQ478153 Eunice mikeli
GQ478146 Eunice mutilata
GQ478144 Eunice amoureuxi
DQ317917 Eunice cariboea
GQ478141 Eunice cariboea
GQ478142 Eunice cf. insularis
GQ478161 Marphysa angeli
GQ478164 Marphysa cf. hentscheli
feces CebG416 Guam
Palola AX3
feces CebH025 Hawaii (2 samples)
Palola AX2
feces CebH088 Hawaii (28 samples)
DQ317897 Palola sp. clade A1
feces CebH001 Hawaii (3 samples)
Palola A1
feces CebG088 Guam (18 samples)
DQ317879 Palola sp clade A3
Palola A3
DQ317881 Palola sp clade A2
DQ317894 Palola sp. clade A4
feces CebG116 Guam (10 samples)
Palola AX1
DQ317895 Palola sp. clade A5
DQ317877 Palola sp. clade A6
feces CebAS092 Am Sam
feces CebAS019a Am Sam
Palola A6
DQ317904 Palola sp. clade A7
DQ317885 Palola sp. clade A8
feces CebG047 Guam (2 samples)
feces CebG153 Guam (4 samples) Am Sam (6 samples)
feces CebG295 Guam
DQ317884 Palola sp. clade A9
feces CebAS001 Am Sam
feces CebAS002 Am Sam
feces CebAS019b Am Sam
Palola A9
DQ317865 Palola sp. clade B5
DQ317867 Palola sp. clade B4
DQ317866 Palola sp. clade B2
DQ317888 Palola sp. clade B1
DQ317890 Palola sp. clade B3
GQ478167 Palola viridis
DQ317889 Palola sp. clade B9
DQ317871 Palola sp. clade B8
DQ317874 Palola sp. clade B7
GQ478168 Palola cf. siciliensis
Palola spp.
0.2

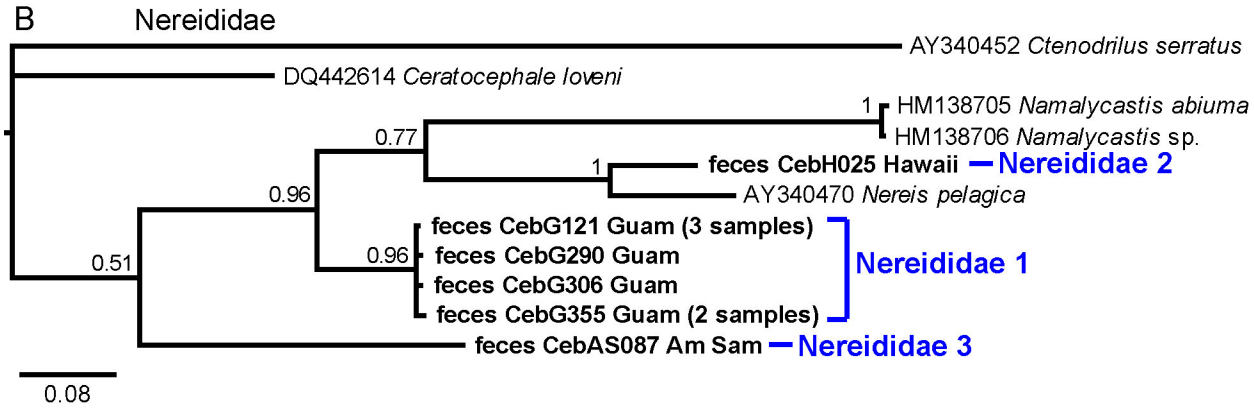


Figure S4. Bayesian consensus phylogenies constructed from sequences of a region of the mitochondrial 16S gene recovered from fecal samples of *C. ebraeus* individuals at American Samoa, Guam and Hawaii (GenBank accession numbers JX177300-JX177352, FJ804537-FJ804572 and FJ907334-FJ907342) and downloaded from GenBank (GenBank accession numbers included in the names of sequences).

Posterior probabilities are labeled at nodes of major clades. Sequences obtained from *C. ebraeus* fecal samples are highlighted in bold. Names of fecal sequences include the location and the number of identical samples from each location if identical sequences were obtained from more than one individual. Putative *Palola* species (order Eunicida) were determined based on the individual clades in the species tree and classifications proposed by Schulze [4]. Classification of putative prey species are labeled in blue next to the clades. Am Sam: American Samoa.

(A) Phylogeny of sequences of Eunicida species constructed with the HKY+I+G model, rooted with the outgroup *Armandia bilobata*.

(B) Phylogeny of sequences of Nereididae species constructed with the GTR+G model, rooted with the outgroup *Ctenodrilus serratus*.

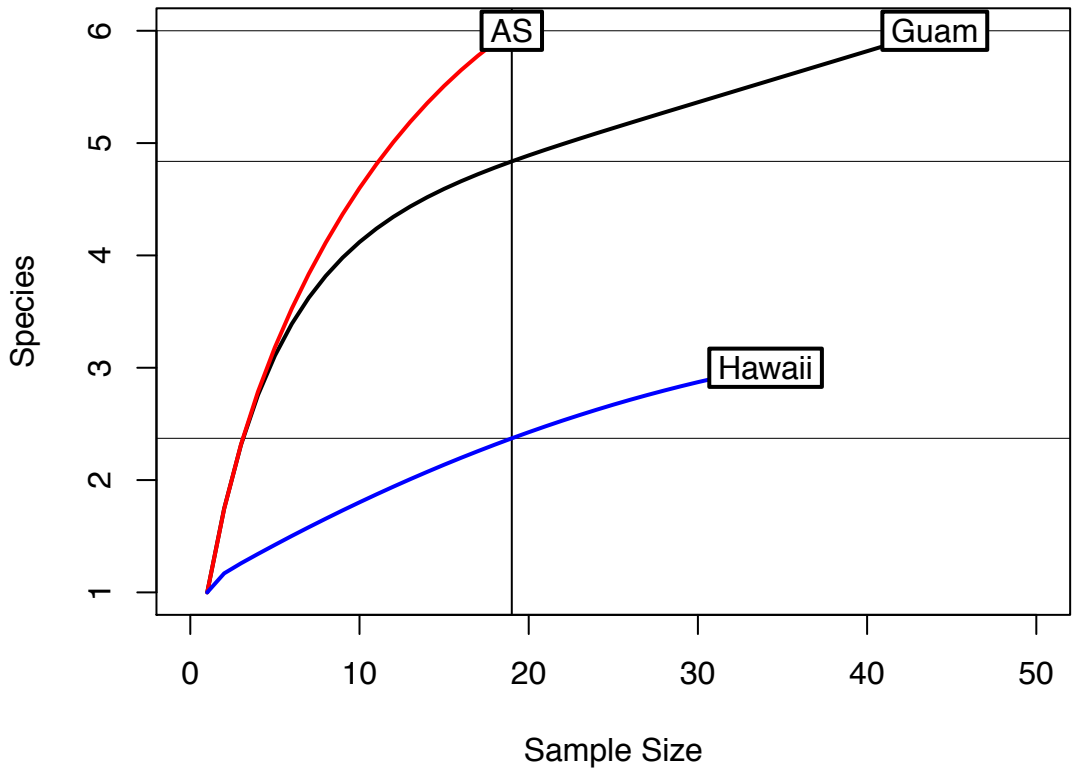


Figure S5. Rarefaction curves of prey species richness versus sample size for Hawaii, Guam and American Samoa populations. The vertical line is given to illustrate prey species richness for the populations given the lowest sample size observed (N=19); the horizontal lines project to expected species richness at size 19 for the three populations. AS: American Samoa.

Table S1. General primers for each conotoxin superfamily. 3'UTR: 3' untranslated region.

Conotoxin superfamily	Toxin type	Primer location	Primer sequences
A	α-conotoxin	Prepro	5'ATGGGCATGCGGATGATGTTCAC 3'
		3'UTR	5' GTCGTGGTTCAGAGGGTCCTGG 3'
O	δ-conotoxin	Prepro	5'CATCACCAAGATGAAACTGACGTG 3'
		3'UTR	5' GCGCCAATCAAAGATCAAGCC 3'
M (primer set MPr1)	μ-conotoxin	Prepro	5' CATGATGTCTAAACTGGGAGT 3'
		3'UTR	5' GCAAATCTGAAGGAGACTGCAATC 3'
M (primer set MPr2)		Prepro	5' GTTGAAAATGGGAGTGGTGCT 3'
		3'UTR	5' ATGATATCAACAAACGCTGTCGTTG 3'
I	-	Prepro	5' ATGATGTTTCGATTGACGTCAGTCAG 3'
		3'UTR	5' ACGTCAGGCTTGAGTTATCTGCC 3'

Table S2. Locus-specific primers used to genotype each locus.

Locus	Location	Primer Sequences
ED4	Forward	5'GTAAAACGACGGCCAGTATTGCACCAGAAAAGATGCGTACAG3'
	Reverse	5'CAGGAAACAGCTATGACCGCGCCAATCAAAGATCAAGCC3'
ED6	Forward	5'GTAAAACGACGGCCAGTAATTGCACCAGAAAAGATGCRTAAAC3'
	Reverse	5'CAGGAAACAGCTATGACCGCGCCAATCAAAGATCAAGCC3'
ED20	Forward	5'GTAAAACGACGGCCAGTTAAATTGCACGAGAAATCATGCATTAG3'
	Reverse	5'CAGGAAACAGCTATGACCGCGCCAATCAAAGATCAAGCC3'
EA4	Forward	5'GTAAAACGACGGCCAGTGATCGCTCTGATCGCCACACGC3'
	Reverse	5'CAGGAAACAGCTATGACCTGGAGTAGCAGCGTCTTCAACG3'

Table S3. Allelic-specific primers to verify and differentiate alleles.

Locus	Location	Primer Sequences	Purpose
ED4	Forward	5'GTAAAACGACGGCCAGTCATCAG CAAGATGAAACTGAC3'	differentiate allele 40 from allele 5, verified by sequencing
	Reverse	5'CAGGAAACAGCTATGACCCGATG GACACGAACCACCCGTC3'	
ED6	Forward	5'GTAAAACGACGGCCAGTGCACCA GAAAGCATGCGTAAACAG3'	Differentiate 7+39 allele pairs and 6+38 pairs, verified by sequencing
	Reverse	5'CAGGAAACAGCTATGACCGCGCC AATCAAAGATCAAGCC3'	

Table S4. Results of hierarchical Analysis of Molecular Variance (AMOVA) for the highly polymorphic loci (ED4, ED6 and E1) with the Tamura-Nei [2] model. Three types of grouping were tested for each locus: H, (G, A) represents grouping of Guam with American Samoa; G, (H, A) represents grouping of Hawaii and American Samoa; A, (H, G) represents grouping of Hawaii and Guam. Significance of F_{SC}, F_{ST} and F_{CT} values was evaluated by 10,100 random permutations. The negative percentage of covariance among groups may result from the linear restriction of the model and large variations within groups. Results showed that levels of variance among groups for the H, (G, A) grouping is much larger than levels of variance among populations within groups, F_{CT} is large, and the *P*-value is the smallest among the three groupings.

Locus	Grouping	Variation among groups (%)	Variation among pops within groups (%)	Variation within pops (%)	F_{SC}	F_{ST}	F_{CT}
ED4	H, (G,A)	20.20	0.34	79.46	$0.004^{P=0}$	$0.205^{P=0}$	$0.202^{P=0.33}$
	G, (H,A)	-1.75	18.58	83.17	$0.183^{P=0}$	$0.168^{P=0}$	$-0.017^{P=0.66}$
	A, (H,G)	-13.45	24.34	89.11	$0.215^{P=0}$	$0.109^{P=0}$	$-0.134^{P=1}$
ED6	H, (G,A)	30.98	1.83	67.19	$0.03^{P=0}$	$0.33^{P=0}$	$0.31^{P=0.34}$
	G, (H,A)	-20.90	44.68	76.22	$0.37^{P=0}$	$0.24^{P=0}$	$-0.21^{P=1}$
	A, (H,G)	-7.29	32.54	74.75	$0.30^{P=0}$	$0.25^{P=0}$	$-0.07^{P=0.67}$
E1	H, (G,A)	17.53	-0.65	83.12	$-0.01^{P=0.01}$	$0.17^{P=0}$	$0.18^{P=0.33}$
	G, (H,A)	-7.01	18.81	88.20	$0.18^{P=0}$	$0.12^{P=0}$	$-0.07^{P=1}$
	A, (H,G)	-5.82	17.32	88.50	$0.16^{P=0}$	$0.12^{P=0}$	$-0.06^{P=0.67}$

Table S5. Sample sizes, numbers of total (and unique) alleles, gene diversity, nucleotide diversity and their standard errors (SE) of the five conotoxin loci at three locations. AS: American Samoa.

Locus	Location	Sample size	Alleles (unique)	Gene Diversity (SE)	Nucleotide Diversity (SE)
ED4	**AS**	10	6 (0)	0.632 (0.113)	0.017 (0.010)
	Guam	24	8 (3)	0.714 (0.041)	0.028 (0.017)
	Hawaii	28	4 (0)	0.201 (0.070)	0.005 (0.004)
ED6	**AS**	13	5 (0)	0.785 (0.041)	0.053 (0.030)
	Guam	24	6 (1)	0.638 (0.064)	0.041 (0.024)
	Hawaii	30	3 (1)	0.242 (0.070)	0.011 (0.009)
E1	**AS**	21	7 (1)	0.678 (0.064)	0.025 (0.014)
	Guam	29	6 (0)	0.682 (0.041)	0.023 (0.013)
	Hawaii	48	4 (2)	0.620 (0.031)	0.022 (0.012)
ED20	**AS**	11	2 (1)	0.091 (0.081)	0.001 (0.002)
	Guam	25	2 (0)	0.040 (0.038)	0.001 (0.002)
	Hawaii	20	2 (0)	0.050 (0.047)	0.001 (0.002)
EA4	**AS**	14	2 (0)	0.198 (0.092)	0.004 (0.004)
	Guam	36	2 (0)	0.178 (0.056)	0.003 (0.004)
	Hawaii	15	2 (0)	0.186 (0.088)	0.004 (0.004)

Table S6. Posterior probabilities of the number of populations (K), assuming a uniform prior of K=2, 3, and 4. Log probabilities of the data given K (i.e. log (P(data|K))) were estimated with Structure 2.3.3.

K	log(P(data|K))	Posterior Probability
K=2	-332.3	1.0
K=3	-352.4	0.0
K=4	-367.7	0.0

Table S7. Tests of neutrality of conotoxin loci with Tajima's *D*. Tajima's *D* values were estimated for each locus in each location and *P*-values were estimated by percentages of values in 10,000 simulations that were larger than or equal to observed values. *: $P<0.05$, **: $P<0.001$. *D* values highlighted in bold are significantly different from zero after the strict Bonferroni correction for multiple tests.

Locus	Population	Tajima's *D*
ED4	AS	0.466
	Guam	**2.205***
	Hawaii	-1.649*
ED6	AS	1.805*
	Guam	1.108
	Hawaii	-0.187
E1	AS	0.837
	Guam	0.842
	Hawaii	**3.216****
ED20	AS	-1.162
	Guam	-1.103*
	Hawaii	-1.124*
EA4	AS	-0.477
	Guam	-0.225
	Hawaii	-0.537

Table S8. Coefficients of the slope of the fitted line in simple regression analyses of the haplotype and nucleotide diversities of five conotoxin genes and the COI gene with the diversities of prey (*H'* and genetic distance). Haplotype diversities of the mitochondrial COI gene for populations at Hawaii, Guam and American Samoa are nearly equivalent (0.963 at Hawaii, 0.978 at Guam, 0.947 at American Samoa; retrieved from Duda and Lessios [5]). The Guam population exhibits slightly higher nucleotide diversity (0.009 at Guam, 0.006 at American Samoa and Hawaii; estimated with Tamura-Nei model from the COI gene sequences reported in Duda and Lessios [5]).

Locus	Haplotype diversity vs *H*'	Haplotype diversity vs genetic distance	Nucleotide diversity vs *H*'	Nucleotide diversity vs genetic distance
ED4	0.423	2.597	0.016	0.079
ED6	0.422	3.069	0.032	0.237
E1	0.054	0.344	0.001	0.016
ED20	0.014	0.199	0	0
EA4	0.002	0.054	0	-0.001
COI	-0.001	-0.069	0.001	0.002

Table S9. Pearson [6], Spearman [7] and Kendall [8] correlation coefficients of the pairwise Φ_{ST} matrices of each of the three highly polymorphic conotoxin genes with the pairwise divergence indices of prey (PS_I and D_{ST}). Coefficients of Φ_{ST} with Pianka's overlap index are identical to those with PS_I.

Locus	PS_I			D_{ST}		
	Pearson	**Spearman**	**Kendall**	**Pearson**	**Spearman**	**Kendall**
ED4	-0.999	-0.866	-0.817	0.727	1.000	1.000
ED6	-0.985	-0.866	-0.817	0.746	1.000	1.000
E1	-0.999	-0.866	-0.817	0.655	1.000	1.000

Table S10. Gene and nucleotide diversities of two O-superfamily conotoxin genes *MIL2* and *MIL3* and the mitochondrial COI gene of *C. miliaris* populations at Easter Island (abbreviated as EI), Guam and American Samoa (abbreviated 'AS'). Standard deviations of indices are presented in parentheses. Distances among haplotypes are calculated with respective models used in Duda and Lee [9]: K80 [10] model for locus *MIL2*, Jukes-Cantor [11] model for locus *MIL3*, and Tamura-Nei model for the COI gene.

Locus	Gene Diversity (Standard Deviation)			Nucleotide Diversity (Standard Deviation)		
	EI	**Guam**	**AS**	**EI**	**Guam**	**AS**
MIL2	0.635 (0.043)	0.271 (0.084)	0.381 (0.094)	0.015 (0.009)	0.010 (0.007)	0.014 (0.009)
MIL3	0.747 (0.036)	0.594 (0.070)	0.631 (0.064)	0.021 (0.001)	0.017 (0.002)	0.017 (0.001)
COI	0.961 (0.014)	1.000 (0.017)	0.979 (0.016)	0.008 (0.004)	0.010 (0.005)	0.010 (0.005)

References

1. Tamura K. 1992 Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. *Mol Biol Evol* **9**, 678-687.
2. Tamura K., Nei M. 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**, 512-526.
3. Hasegawa M., Kishino H., Yano T. 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160-174.
4. Schulze A. 2006 Phylogeny and Genetic Diversity of Palolo Worms (Palola, Eunicidae) from the Tropical North Pacific and the Caribbean. *Biol Bull* **210**(1), 25-37.
5. Duda T.F., Lessios H.A. 2009 Connectivity of populations within and between major biogeographic regions of the tropical Pacific in *Conus ebraeus*, a widespread marine gastropod. *Coral Reefs* **28**(3), 651-659.
6. Rodgers J.L., Nicewander W.A. 1988 Thirteen Ways to Look at the Correlation Coefficient. *Am Stat* **42**(1), 59-66.
7. Spearman C. 1910 Correlation calculated from faulty data. *Brit J Psychol, 1904-1920* **3**(3), 271-295.
8. Kendall M.G. 1948 *Rank correlation methods.* Oxford, England, Griffin.
9. Duda T.F., Lee T. 2009 Ecological release and venom evolution of a predatory marine snail at Easter Island. *PLoS One* **4**(5), e5558.
10. Kimura M. 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**(2), 111-120.
11. Jukes T.H., Cantor C.R. 1969 Evolution of protein molecules. In *Mammalian protein metabolism* (ed. Munro H.N.), pp. 21-123. New York, Academic Press.