

1 **Phylogenomics of scorpions reveal a co-diversification of scorpion mammalian**
2 **predators and mammal-specific sodium channel toxins**

3
4 CARLOS E. SANTIBÁÑEZ-LÓPEZ^{1,+,*}, SHLOMI AHARON², JESÚS A. BALLESTEROS¹, GUILHERME
5 GAINETT¹, CAITLIN M. BAKER^{1,3}, EDMUNDO GONZÁLEZ-SANTILLÁN⁴, MARK S. HARVEY⁵,
6 MOHAMED K. HASSAN⁶, ALI HUSSIN ABU-ALMAATY⁶, SHOROUK MOHAMED ALDEYARBI⁶, LIONEL
7 MONOD⁷, ANDRÉS OJANGUREN-AFFILASTRO⁸, ROBERT J. RAVEN⁹, RICARDO PINTO-DA-ROCHA¹⁰,
8 YORAM ZVIK¹¹, EFRAT GAVISH-REGEV^{2,#}, PRASHANT P. SHARMA^{1,#}
9

10 ¹Department of Integrative Biology, University of Wisconsin–Madison, Madison, WI 53706; ²National Natural
11 History Collections, The Hebrew University of Jerusalem, Jerusalem, Israel; ³Department of Organismic and
12 Evolutionary Biology, Harvard University, Cambridge, MA 02138; ⁴Laboratorio de Aracnología, Facultad de
13 Ciencias, Departamento de Biología Comparada, Universidad Nacional Autónoma de México, C.P. 04510 Mexico
14 City, Mexico; ⁵Department of Terrestrial Zoology, Western Australian Museum, Locked Bag 49, Welshpool DC
15 6986, Western Australia, Australia; ⁶Zoology Department, Faculty of Science, Port Said University, Egypt; ⁷
16 Département des arthropodes et d'entomologie I, Muséum d'histoire naturelle, Route de Malagnou 1, 1208 Geneve,
17 Switzerland; ⁸División Aracnología, Museo Argentino de Ciencias Naturales, Av. Ángel Gallardo 470, Buenos
18 Aires, Argentina; ⁹Department of Terrestrial Biodiversity, Queensland Museum, Grey Street, South Bank, Brisbane,
19 Queensland, Australia; ¹⁰Departamento de Zoología, Instituto de Biociências, Universidade de São Paulo, Rua do
20 Matão, travessa 14, 321, São Paulo, Cep: 05508-900, Brasil; ¹¹Hoopoe Ornithology & Ecology Center, Yeroham,
21 Israel & Ben-Gurion University of the Negev, Be'er-Sheva, Israel; ⁺Present address: Department of Biology, Eastern
22 Connecticut State University, 83 Windham Street, Willimantic, CT 06226, USA
23 * To whom correspondence may be addressed. Email: santibanezlopezc@easternc.edu.
24 [#] Co-senior authors.

25

26 Abstract

27 Scorpions constitute a charismatic lineage of arthropods and comprise more than 2,500 described
28 species. Found throughout various tropical and temperate habitats, these predatory arachnids
29 have a long evolutionary history, with a fossil record that began in the Silurian. While all
30 scorpions are venomous, the asymmetrically diverse family Buthidae harbors nearly half the
31 diversity of extant scorpions, and all but one of the 58 species that are medically significant to
32 humans. Many aspects of scorpion evolutionary history are unclear, such as the relationships of
33 the most toxic genera and their constituent venom peptides. Furthermore, the diversification age
34 of toxins that act specifically on mammalian ion channels have never been inferred. To redress
35 these gaps, we assembled a large-scale phylogenomic dataset of 100 scorpion venom
36 transcriptomes and/or genomes, emphasizing the sampling of highly toxic buthid genera. To
37 infer divergence times of venom gene families, we applied a phylogenomic node dating approach
38 for the species tree in tandem with phylostratigraphic bracketing to estimate minimum ages of
39 mammal-specific toxins. Our analyses establish a robustly supported phylogeny of scorpions,
40 particularly with regard to relationships between medically significant taxa. Analysis of venom
41 gene families shows that mammal-specific sodium channel toxins have independently evolved in
42 five lineages within Buthidae. The temporal windows of mammal-specific toxin origins are
43 contiguous with the basal diversification of major scorpion mammal predators such as
44 carnivores, shrews, bats and rodents. These results suggest an evolutionary arms race model
45 comprised of co-diversification of mammalian predators and NaTx homologs in buthid venom.

46

47 Arachnida | venom | phylostratigraphy | dating | co-diversification

48

PHYLOGENOMIC ANALYSIS OF SCORPIONS

49 To scientists and laypersons, scorpions are fascinating for the potency of their venom, a complex
50 mixture of bioactive compounds secreted in specialized organs in the ultimate body segment, and
51 used to disrupt biochemical and physiological processes in target organisms (King and Hardy
52 2013; Casewell et al. 2013). This lineage of arachnids diversified in the Permian and harbors
53 considerable modern diversity; nearly 2500 species have been described in the presently 22
54 recognized families, with this diversity distributed across all continents except Antarctica
55 (Sissom 1990; Santibáñez-López et al. 2019a, b; Tropea and Onmis 2020). Scorpions are
56 particularly well-adapted to survival in extreme habitats, and their ability to produce and deliver
57 venoms is inferred to be a key factor to their success. A major dimension of their biodiversity is
58 found in their venoms, which are rich in toxins that have a broad array of biological targets
59 (including those affecting Na^+ , K^+ , Cl^- and Ca^{2+} ion channels), mucopolysaccharides, and
60 enzymes (Cao et al. 2013; He et al. 2013). Scorpion envenomation causes nearly an order of
61 magnitude greater fatalities worldwide than snakebites, and particularly so in developing or rural
62 subtropical regions. Intensive functional study of specific peptides has uncovered significant
63 biomedical applications in scorpion venom, such as the identification of antimicrobial and
64 antitumor agents, fluorescent “tumor paint”, and transport molecules for molecular cargo (Veiseh
65 et al. 2007; Rapôso 2017; Díaz-Perlas et al. 2018).
66 While all scorpions are venomous, both the extant diversity of scorpions, as well as the toxicity
67 of their venom to mammals, is asymmetrical distributed; ca. 1200 species (47% of described
68 diversity) of scorpions are members of the family Buthidae (“thick-tailed scorpions”; Fig. 1),
69 which includes nearly all significantly venomous scorpion species, as well most of the known
70 molecular diversity of scorpion venom (Santibáñez-López et al. 2019b). All scorpions are
71 thought to possess insect-specific ion channel toxins, which facilitate prey capture. Within

72 buthids, salient components of the venom cocktail are ion channel toxins that are specific to
73 mammalian targets and are inferred to function as anti-predator deterrents (Niermann et al.
74 2020). Such neurotoxins operate by blocking action potentials at nerve synapses, precipitating
75 symptoms of neurotoxicosis such as intense pain, hypersalivation, muscle spasms, asphyxia, and
76 paralysis. The genera *Androctonus*, *Buthus*, *Centruroides*, *Hottentotta*, *Leiurus*, *Parabuthus* and
77 *Tityus* all contain multiple highly toxic and medically significant species known for the potency
78 of their neurotoxins (Santos et al. 2016; Niermann et al. 2020).

79 Surprisingly, little is understood as to about the evolutionary relationships of medically relevant
80 scorpions. The higher-level molecular phylogeny of buthids was first inferred using a 296-bp
81 fragment of 16S rRNA and sampling 17 genera, with weak resolution of basal relationships (Fet
82 et al. 2003). Subsequently, most molecular phylogenetic studies of Buthidae have used a handful
83 of Sanger-sequenced loci to address relationships of derived groups, such as individual genera or
84 subfamilies, or of buthids restricted to specific geographic terranes (Ojanguren-Affilastro et al.
85 2017; Suranse et al. 2017). Mitochondrial DNA surveys of medically significant buthids in
86 particular have not yielded supported basal relationships (Borges et al. 2014). Furthermore, in the
87 first phylogenomic study of scorpions, sampling of Buthidae was limited to just four species
88 (Sharma et al. 2015). While subsequent analyses of venom evolution added buthid terminals to
89 the scorpion tree, these additional datasets were derived from Sanger-sequenced EST libraries or
90 454-pyrosequenced datasets, resulting in high proportions of missing data and attendant
91 instability in buthid relationships (Santibáñez-López et al. 2018). Other recent phylogenomic
92 investigations have prioritized the systematically complex Iurida, without adding to the sampling
93 of the original four Buthidae transcriptomes (Sharma et al. 2018; Santibáñez-López et al. 2019a,
94 b; Santibáñez-López et al. 2020).

PHYLOGENOMIC ANALYSIS OF SCORPIONS

95 The lack of a robust and densely sampled molecular phylogeny of scorpions has hindered various
96 evolutionary analyses pertaining to the evolution of scorpion morphology, biogeography, and
97 venomics. In particular, undersampling of buthids risks omitting key nodes for understanding the
98 mode and tempo of scorpion and scorpion venom evolution. Targeting this group in
99 phylogenomic analyses is key to testing hypotheses about single versus multiple origins of
100 mammal-specific toxins, as well as their potential co-diversification with mammalian predators.
101 To close these knowledge gaps, we assembled a phylogenomic dataset of 100 scorpion terminals.
102 Within Buthidae, we emphasized sampling of highly toxic species from fieldwork theaters in the
103 southwestern US, the Neotropics, and the Middle East. Here, we show that mammal-specific
104 toxins have originated independently in five major clades within Buthidae and the timing of their
105 origins coincides with the diversification of major mammalian predators of scorpions.

106

107 MATERIALS AND METHODS

108 Extended methods are provided in the Supplementary Text.

109

110 *Fieldwork, RNA-Seq, and Phylogenomic Analyses.*

111 Scorpions were collected by hand in field theaters across Brazil, Egypt, Israel, and the US,
112 commonly with the aid of ultraviolet lighting. A subset of well-studied species was obtained
113 through captive breeding programs. Milking and dissection of venom glands, RNA extraction,
114 and paired-end transcriptome sequencing was performed on the Illumina HiSeq 2500 platform
115 for 42 species, following our previous approaches (Sharma et al. 2015; Santibáñez-López et al.
116 2018). New datasets were combined with 45 venom gland RNA-Seq datasets and one genome
117 we previously generated (Santibáñez-López et al. 2018; Schwager et al. 2017). Twenty outgroup

118 species were including in the analysis, spanning tetrapulmonates, pseudoscorpions, harvestmen,
119 and three horseshoe crab genomes. All collecting, vouchering, and accession data are provided in
120 Supplementary Tables S1-4.

121 Orthologous loci were drawn from MCL clustering of 3564 orthogroups computed from a larger
122 analysis of Chelicerata and outgroup taxa (Ballesteros and Sharma 2019). Untrimmed alignments
123 were used to produce a hidden Markov profile using hmmerbuild from hmmer package v.3.2.1
124 (Mistry et al. 2013). Each proteome/transcriptome of the species of interest was then searched
125 (hmmersearch) for matches against the collection of profiles with an expectation threshold of $e <$
126 10^{-20} ; for cases with more than one hits per locus, the sequence with the best score was selected,
127 and the corresponding sequence appended to the locus FASTA file aggregating the putative
128 orthologous found in each species. Clustering of putative orthologs was tested by comparing
129 each orthogroup's constituent sequences with the proteome of the fruit fly *Drosophila*
130 *melanogaster*, and removing from orthogroups any sequences with mismatching functional
131 annotations. Finally, alignments and gene trees were visually inspected for evident paralogy.
132 Paralogous sequences were discarded; if the number of sequences retained post-trimmed was
133 below the minimum taxon occupancy threshold, the entire locus was discarded.

134 Three matrices were assembled with minimal taxon occupancy thresholds: Matrix 1 (at least 115
135 species), Matrix 2 (at least 109 species), and Matrix 3 (at least 103 species). Phylogenetic
136 inference of these concatenated matrices was computed with IQ-TREE v. 1.6 (Nguyen et al.
137 2014) implementing the best-fitting amino acid substitution model per partition (-spp;
138 Supplementary Text). As approaches based on concatenation can be prone to incorrect inference,
139 species trees were estimated using the coalescent method implemented in ASTRAL v.3 (Mirarab
140 and Warnow 2015), using the collection of orthologous gene trees as inputs. Analysis of the

PHYLOGENOMIC ANALYSIS OF SCORPIONS

141 smallest matrix (Matrix 3) was trialed using Phylobayes-mpi v. 1.7 (Lartillot and Philippe 2004)
142 with four independent chains under the CAT + GTR + Γ_4 model.

143

144 *Divergence Time Estimation*

145 Divergence time estimation was computed on Matrix 2 using the approximate likelihood
146 calculation as implemented in Codeml and MCMCTree (both part of the PAML v. 4.8 software
147 package (Yang 2007; dos Reis and Yang 2019). The ML tree inferred from Matrix 2 was used as
148 the input tree and was calibrated using fifteen fossil taxa (Supplementary Text, Supplementary
149 Table S5). Four Bayesian inference chains were run for 2.5 M post-burnin generations (burnin of
150 25000 generations); convergence diagnostics were assessed using inbuilt tools in MCMCTree
151 (Puttick 2019).

152

153 *Toxin Homology and Evolution*

154 Cysteine-stabilized α -helix and β -sheet fold (CS $\alpha\beta$), disulphide-directed beta-hairpin (DDH)
155 and Inhibitor cystine knot (ICK) homologs from scorpion venom were retrieved from the
156 complete dataset used in the scorpion phylogenetic analyses following our recent approaches
157 (Santibáñez-López et al. 2018; Santibáñez-López et al. 2019b), and from UniProt
158 (Supplementary Table S6). Gene trees were conducted using IQ-TREE for the entire dataset
159 (1,353 CS $\alpha\beta$ -ICK scorpion toxins, with 41 DDH scorpion toxins as outgroups), and for each of
160 the four main clades recovered: (a) sodium channel toxins (NaTx); (b) potassium channel toxins
161 (KTx); (c) chlorine channel toxins (CITx); and (d) calcins. Comparative analyses between the
162 subclades recovered within the NaTx included the search for repetitive motifs in their mature
163 peptide using Multiple Em for Motif Elicitation (Bailey et al. 2015). The mature peptide of the

164 two main clades within the NaTx (Aah2-like and Cn2-like) were separately analyzed using
165 CLANS clustering (Frickey and Lupas 2004). Estimation of minimum ages for mammal-specific
166 toxins was derived from molecular dating performed herein, using the ages of the most inclusive
167 clades of taxa as minimum age estimates for gene age (phylostratigraphic bracketing).
168 Divergence times for scorpion predators such Herpestidae (Carnivora), Chiroptera, Eulipotyphla
169 and Rodentia were retrieved from a recent analysis of mammal diversification times (Upham et
170 al. 2019) and compared against the mammal-specific toxin origins.

171

172 RESULTS

173 *Phylogenomic analysis*

174 To infer the phylogenomic relationships of scorpions, we generated a dataset comprised of 99
175 venom gland transcriptomes (97 sequenced by us) and the genome of the buthid *Centruroides*
176 *sculpturatus* (Schwager et al. 2017), with 42 new assemblies sequenced herein (Supplementary
177 Tables S1-4). Inclusion of an older pyrosequenced genome of *Mesobuthus martensii* and Sanger-
178 sequenced EST libraries from other buthid species was trialed as separately from main analyses
179 to the degree of missing data incurred by these terminals. Outgroup taxa consisted of 20
180 transcriptomes spanning seven non-scorpion chelicerate orders. Inference of orthology followed
181 a previously established pipeline, drawing upon groupings of MCL clusters established in a
182 previous analysis of chelicerate relationships, followed by phylogenetically informed
183 identification of orthologous groups (Ballesteros and Hormiga 2016; Ballesteros and Sharma
184 2019). Three phylogenomic matrices were constructed spanning 192 to 660 loci (53,333 to
185 185631 aligned amino acid sites), using taxon minimum occupancy thresholds of 103, 109, and

PHYLOGENOMIC ANALYSIS OF SCORPIONS

186 115 terminals per locus (88.0%, 91.4%, and 94.1% complete, respectively). We refer to these
187 henceforth as Matrices 1-3, respectively.

188 Maximum likelihood analyses of concatenated datasets using IQ-TREE v. 1.6 consistently
189 recovered with maximal nodal support the monophyly of scorpions and the established basal
190 split between Buthida (consisting of Buthidae, Chaerilidae, and Pseudochactidae) and Iurida (the
191 remaining scorpion families) (Sharma et al. 2015). Matrices 1 and 2 recovered support for
192 *Lychas variegatus* as the sister group of the remaining buthids (87-95% bootstrap frequency),
193 whereas Matrix 3 recovered a nested placement of *L. variegatus* with weak nodal support (72%;
194 Supplementary Figs. S1-S3). All three datasets recovered with maximal support four major
195 clades within the buthids, consisting of (a) the largely Palearctic “*Buthus* group” (*Androctonus*,
196 *Birulatus*, *Buthus*, *Buthacus*, *Compsobuthus*, *Hottentotta*, *Leiurus*, and *Orthochirus*, with
197 *Mesobuthus* recovered in this group in supplementary analyses), (b) *Ananteris* and *Babycurus*,
198 (c) the “*Uroplectes* group” (*Grosphus*, *Parabuthus* and *Uroplectes*), and (d) the “*Tityus* group”
199 (*Centruroides*, *Heteroctenus*, *Jaguajir*, *Physoctonus*, *Rhopalurus*, *Tityus*, and *Troglorhopalurus*).
200 Relationships between these groups across all supermatrix analyses supported the sister group
201 relationship of the *Tityus* and *Uroplectes* groups, with this lineage in turn sister group to
202 (*Ananteris* + *Babycurus*). Relationships within Iurida closely reflected tree topologies reported in
203 our previous works and are not discussed further herein (Sharma et al. 2015; Santibáñez-López
204 et al. 2019a, b; Santibáñez-López et al. 2020).
205 In addition to maximum likelihood tree reconstruction based on a supermatrix approach, we
206 inferred relationships using multispecies coalescent (MSC) tree reconstruction using ASTRAL
207 v.2 for the constituent gene trees of Matrices 1-3. Higher-level relationships of scorpions were

208 largely congruent across analyses, excepting the placement of *Lychas*, which was unsupported
209 (posterior probability [PP] = 0.47-0.80 across analyses; Supplementary Fig. S4).
210 While we trialed the use of Phylobayes-mpi to estimate phylogenomic relationships under the
211 CAT + GTR + Γ_4 model, Bayesian inference chains consistently failed to converge (*maxdiff* =
212 1.0) after 10,596 -11,574 cycles of computational effort across four chains, equaling six months
213 of continuous wall-clock computation. The resulting tree topology of the chain with the highest
214 log-likelihood nevertheless reflected the same relationships as the maximum likelihood
215 inferences (Supplementary Fig. S5).

216

217 *Molecular dating*

218 Phylogenomic estimation of divergence times was inferred using a node dating approach with
219 MCMCTree on Matrix 2, implementing a likelihood approximation of branch lengths using a
220 multivariate normal distribution (Yang 2007; dos Reis and Yang 2019). Fossils used to inform
221 the dating consisted of six ingroup and nine outgroup node calibrations. All calibrations were
222 implemented as soft minimum and soft maximum ages. Identity of fossil taxa and
223 implementation of temporal ranges as calibrations are provided in the Supplementary Text, and
224 Supplementary Table S5. Divergence times were inferred under two models of rate evolution, a
225 correlated rates model and an independent rates model. Both clock models recovered comparable
226 inferences of basal scorpion diversification, with a split between Buthida and Iurida dating to the
227 Carboniferous-Permian (291-320 Mya; highest posterior density [HPD] interval: 247-352 Mya)
228 (Fig. 2). The diversification of Buthidae was estimated to span the end-Jurassic to the Early
229 Cretaceous (95% HPD: 105-161 Mya). Divergences within the major buthid clades fell within

PHYLOGENOMIC ANALYSIS OF SCORPIONS

230 the Late Cretaceous to Paleogene, excepting the *Buthus* group, wherein most divergences were
231 estimated to occur in the Neogene.

232

233 *Phylostratigraphic bracketing of venom gene family origins*

234 Putative cysteine-stabilized α -helix and β -sheet fold (CS $\alpha\beta$) and Inhibitor cystine knot (ICK)
235 venom components were identified in the 100 terminals and from the UniProt databases. Within
236 these peptides, the most relevant are those affecting the ion channels, such as the sodium channel
237 toxins (NaTx), potassium channel toxins (KTx), chlorotoxins (CITx), and the ryanodine receptor
238 ligands (Calcins). Using gene tree reconstructions, we documented 1,353 CS $\alpha\beta$ -ICK peptide
239 homologs spanning 151 scorpion species in 19 families (from which 48% were buthids;
240 Supplementary Table S6). Maximum likelihood analysis of the CS $\alpha\beta$ -ICK matrix and 41
241 disulphide-directed beta-hairpin (DDH) sequences as outgroups (399 amino acid sites) recovered
242 a gene tree subdivided into three major clades: (a) NaTx; (b) the calcins; and (c) the KTx
243 including the nested CITx (Supplementary Fig. S6-7). Our results reveal that calcins are
244 phylogenetically restricted to iurids (Supplementary Figs. S6, S8), whereas CITx are restricted to
245 a subset of Old World buthids (Supplementary Figs. S6, S9). Within KTx, ML analysis
246 recovered the presence of nine clusters, corresponding to (a) α KTx; (b) β KTx; (c) ϵ KTx (unique
247 to buthids); (d) κ KTx (unique to iurids); (e) λ KTx (unique to buthids); (f, g) Scorpine-like type 1
248 and type 2; and (h, i) Kunitz-type 1 and type 2 (Supplementary Fig. S7).
249 Within NaTx, mammal-specific toxins were recovered as six clusters within two major clades:
250 (a) the Aah2-like; and (b) the Cn2-like clades (Figs. 3A-B). The Aah2-like gene family was
251 found exclusively in Buthida (mostly from Old World Buthids, Supplementary Fig. S10-11) and
252 encodes for peptides with arthropod affinity, insect and mammal affinity, and mammal-affinity

253 (Figs. 3B). In contrast, within the Cn2-like gene family, we found one cluster restricted to Iurida,
254 and six clusters exclusively in Buthida (Supplementary Figs. S11-15). Among these, two clusters
255 with mammal-specific targets were found uniquely in the genera *Centruroides* and *Tityus*
256 (Supplementary Figs. S14-15). Our search for motifs with MEME v. 5.1 (Bailey et al. 2015)
257 showed a short non-unique motif of 10 amino acids (GXXCWCXXLPD) for members of the
258 Aah2-like gene family, and no specific motif for members of the Cn2-like gene family. More
259 specifically, no conserved or repetitive motifs were found in the mammal-specific toxins of
260 either the Aah2 or Cn2 gene families (Supplementary Fig. S16). Given the short sequence length
261 of NaTx, we also assessed patterns of relatedness between the 661 mature peptide sequence from
262 our dataset using CLANS clustering (Frickey and Lupas 2004). Consistent with the gene tree
263 analyses, this approach recovered seven major groups of peptides with mammal-specific affinity
264 (Fig. 3C).

265 To infer the age of mammal-specific genes, we employed a phylostratigraphic bracketing
266 approach, using the stem age of the most inclusive buthid clade containing a given mammal-
267 specific toxin as a proxy for the minimum estimated gene age. Stem ages of *Centruroides*,
268 *Tityus*, *Parabuthus*, and the node uniting *Hottentotta* with the remaining *Buthus* group scorpions
269 thus implied the earliest diversification of mammal-specific toxins in a temporal window
270 spanning 20-83 Mya (Fig. 2, inset).

271

272 DISCUSSION

273 *Phylogenetic Relationships Within Buthidae Validate Morphology-Based Systematics*

274 The systematic history of scorpions was previously largely dominated by morphological
275 analyses, with marked contention between competing interpretations of homologies and cladistic

PHYLOGENOMIC ANALYSIS OF SCORPIONS

276 practices (Sharma et al. 2015). Phylogenomic analyses initially refuted much of the traditional
277 understanding of scorpion higher-level systematics, with analyses showing that morphological
278 characters used to delimit families and superfamilial taxa are highly homoplastic and/or
279 uninformative (Sharma et al. 2015). While scorpion classification was subsequently revised to
280 bring taxonomic groupings into accordance with phylogenomic outcomes, a paucity of molecular
281 datasets addressing higher-level buthid relationships has remained a hurdle for understanding the
282 evolution of this asymmetrically large family.

283 Interestingly, this work support for generic groupings that were previously defined on the basis
284 of positions of pedipalp trichobothria (sensory setae). Intensive surveys of trichobothrial position
285 by Fet et al. (2005) were previously used to delimit six major groups of scorpions (the *Buthus*,
286 *Ananteris*, *Tityus*, *Charmus*, *Isometrus*, and *Uroplectes* groups), with the Palearctic *Buthus* groups
287 comprising the sister lineage of the remaining Buthidae. Our results accord precisely with the
288 morphological conception of the *Buthus* group, the *Uroplectes* group, and the *Tityus* group, with
289 some analyses additionally recovering the basally branching placement of the *Buthus* group,
290 albeit with weak support (Supplementary Figs. S1, S4).

291 Placements of *Lychas*, *Ananteris*, and *Babycurus* do not accord exactly with the traditional
292 morphological grouping (Supplementary Fig. S17). It was previously thought that *Lychas* fell
293 within a poorly resolved “*Ananteris* group”, whereas *Babycurus* was thought to inhabit another
294 group that includes *Isometrus* (the *Isometrus* group), a genus not sampled herein. However, the
295 Paleotropical genera *Lychas* and *Isometrus* are likely closely related, as they are distinguishable
296 mainly by the condition of the leg tibial spur (Koch, 1977). A previous phylogenetic study also
297 recovered *Isometrus* as a close relative of *Lychas*, though with poor genomic representation of
298 both taxa (Santibáñez-López et al. 2018).

299 Another group not sampled herein using transcriptomic data is the enigmatic *Charmus* group,
300 which is geographically restricted to the Indian subcontinent, southeast Asia, and parts of East
301 Africa. While molecular data are uncommon for this lineage, a separate analysis we performed—
302 in which we combined Matrix 2 with available Sanger-sequenced data for some buthid lineages
303 as well as a pyrosequenced genome for *Mesobuthus martensii*—also revealed that *Charmus* is
304 part of the group that includes the Malagasy endemic genus *Grospus* and the southern African
305 genera *Parabuthus* and *Uroplectes* (Supplementary Fig. S17).
306 Our results thus differed with morphological delimitations of buthid relationships only with
307 respect to the composition of the two groups that were also poorly resolved by the morphological
308 data in that study (the *Isometrus* and *Ananteris* groups; Fet et al. 2005). These outcomes
309 vindicate the phylogenetic utility of trichobothrial arrangement in establishing shallow-level
310 relationships within scorpion families.

311

312 *Contemporaneous Diversification of Buthid Mammal-Specific Toxins and Scorpion Mammalian*
313 *Predators*

314 Previous genomic resources sampling scorpion venom diversity have focused on a handful of
315 medically significant species, such as *Leiurus quinquestriatus* (Egyptian death stalker scorpion)
316 and *Centruroides sculpturatus* (Arizona bark scorpion) (Veiseh et al. 2007; Borges and Graham
317 2014; Schwager et al. 2017). We focused herein on comparative analyses sampling venom gene
318 expression broadly across scorpion phylogeny, toward characterizing the evolutionary dynamics
319 that precipitated scorpion toxicity to mammals. We also endeavored to increase high-quality
320 transcriptomic resources for species renowned for their toxicity in the genera *Androctonus*,
321 *Centruroides*, and *Tityus*. Through these datasets, we discovered that certain classes of toxins are

PHYLOGENOMIC ANALYSIS OF SCORPIONS

322 phylogenetically restricted; as examples, chlorotoxins occur only in Old World Buthidae;
323 scorpines occur in all families except Buthidae; and calcins are restricted to Iurida
324 (Supplementary Fig. S6).

325 Phylogenomic dating recovered a surprisingly young age for the basal diversification of Buthidae
326 in the Late Mesozoic. We inferred 21 independent origins of sodium channel toxins with known
327 specificity for mammalian targets spanning five separate buthid lineages, with three of these
328 occurring in genera sampled with multiple terminals. Phylostratigraphic bracketing of these
329 toxins' origins recovered age estimates broadly overlapping the basal diversification dates of
330 several mammal orders that include scorpion predators (Fig. 2). These mammal-specific buthid
331 NaTx were nested within a cluster of NaTx that target arthropod tissues across all scorpions (Fig.
332 3).

333 These results are consistent with an evolutionary arms race, wherein mammal-specific toxins
334 were derived in scorpions from insect-specific ancestral peptides, reflecting the derivation of an
335 anti-predator defensive adaptation from peptides previously used to target prey (Fig. 4). Similar
336 dynamics have recently been revealed in the highly venomous Australian funnel-web spiders,
337 wherein δ -hexatoxins exhibit high evolutionary conservation, reflecting a defensive role for
338 deterring vertebrate predators (Herzig et al. 2020). However, in scorpions, molecular signatures
339 of selection revealed no consistent pattern of amino acid sequence evolution across groups of
340 mammal-specific toxins, consistent with the inference of independent evolutionary origins of
341 anti-vertebrate defensive peptides (Supplementary Fig. S14). As an extension of this arms race,
342 counter-adaptations to scorpion venom are known to occur in some scorpion predators. For
343 example, the grasshopper mouse *Onychomys torridus* exhibits reduced sensitivity to pain caused
344 by the sting of the Arizona bark scorpion *Centruroides sculpturatus* (Rowe et al. 2013). The

345 mechanism of this counter-adaptation was shown to be amino acid variants of a voltage-gated
346 Na^+ channel in *O. torridus* that have evolved to selectively bind *C. sculpturatus* toxins, thereby
347 blocking action potential propagation. Parallel evolution of resistance to the venom of *C.*
348 *sculpturatus* via modification of Na^+ ion channels has also been suggested in the bat *Antrozous*
349 *pallidus* (Hopp et al. 2017). Comparable molecular dynamics underlying the evolution of
350 resistance to snake venom have evolved at least four times in mammal predators of snakes, such
351 as mongooses and honey badgers (Barchan et al. 1992; Drabeck et al. 2015; Holding et al. 2016).
352 Given the biodiversity of Buthidae and the molecular diversity of their venoms, broader,
353 phylogenetically informed surveys of venom gland transcriptomics may uncover additional
354 origins of mammal-specific toxin function in scorpions, as well as improve the precision of
355 molecular divergence time estimation. Venom gland transcriptomic databases sampling poorly
356 studied species may also offer additional targets for beneficial biomedical application that are not
357 represented among established scorpion biomedical research programs.

358

359

ACKNOWLEDGMENTS

360 We are indebted to Brandon Myers for providing some of the specimens analyzed in this study.
361 Steve Goodman kindly provided tissue of *Grosphus grandidieri* from Madagascar. Sequencing
362 was performed at the UW-Madison Biotechnology Center. Access to computing nodes for
363 intensive tasks was provided by the Center for High Throughput Computing (CHTC) and
364 facilitated by Christina Koch; and by the Bioinformatics Resource Center (BRC) of the
365 University of Wisconsin–Madison. Specimens in Israel were collected under permits 2017/41718
366 and 2018/42037, issued by the Israel National Parks Authority. Fieldwork in Israel was
367 supported by a National Geographic Society Expeditions Council grant no. NGS-271R-18 to

PHYLOGENOMIC ANALYSIS OF SCORPIONS

368 J.A.B. Fieldwork in Australia was supported by intramural UW-Madison funds. This work was
369 supported by National Science Foundation (grant no. IOS-1552610) to P.P.S.; FAPESP (BIOTA,
370 2013/50297-0), National Science Foundation (DOB- 1343578) to R.P.R.

371

372 REFERENCES

373 Ballesteros J.A., Hormiga G. 2016. A New Orthology Assessment Method for Phylogenomic
374 Data: Unrooted Phylogenetic Orthology. *Mol. Biol. Evol.* 33:2117–2134.

375 Ballesteros J.A., Sharma P.P. 2019. A Critical Appraisal of the Placement of Xiphosura
376 (Chelicerata) with Account of Known Sources of Phylogenetic Error. *Syst. Biol.* 68:896–
377 917.

378 Bailey T.L., Johnson J., Grant C.E., Noble W.S. 2015. The MEME Suite. *Nucleic Acids Res*
379 43:W39–W49.

380 Barchan D., Kachalsky S., Neumann D., Vogel Z., Ovadia M., Kochva E., Fuchs S. 1992. How
381 the mongoose can fight the snake: The binding site of the mongoose acetylcholine
382 receptor. *Proc. Natl. Acad. Sci. USA* 89:7717–7721.

383 Borges A., Graham M.R. 2014. Phylogenetics of Scorpions of Medical Importance. In: *Venom*
384 *Genomics and Proteomics* (ed: Gopalakrishnakone P, Calvete JJ). Springer Press,
385 Dordrecht, Germany, pp 81–104.

386 Cao Z., Yu Y., Wu Y., Hao P., Di Z., He Y., Chen Z., Yang W., Shen Z., He X., Sheng J., Xu X.,
387 Pan B., Feng J., Yang X., Hong W., Zhao W., Li Z., Huang K., Li T., Kong Y., Liu H.,
388 Jiang D., Zhang B., Hu J., Hu Y., Wang B., Dai J., Yuan B., Feng Y., Huang W., Xing
389 X., Zhao G., Li X., Li Y., Li W. 2013. The genome of *Mesobuthus martensii* reveals a
390 unique adaptation model of arthropods. *Nat. Commun.* 4:2602.

391 Casewell N.R., Wüster W., Vonk F.J., Harrison R.A., Fry B.G. 2013. Complex cocktails: the
392 evolutionary novelty of venoms. *Trends Ecol. Evol.* 28:219–229.

393 Díaz-Perlas C., Varese M., Guardiola S., García J., Sánchez-Navarro M., Giralt E., Teixidó M.
394 2018. From venoms to BBB-shuttles. MiniCTX3: a molecular vector derived from
395 scorpion venom. *Chem. Commun.* 54:12738–12741.

396 Drabeck D.H., Dean A.M., Jansa S.A. 2015. Why the honey badger don't care: Convergent
397 evolution of venom-targeted nicotinic acetylcholine receptors in mammals that survive
398 venomous snake bites. *Toxicon* 99:68–72.

399 dos Reis M., Yang Z. 2019. Bayesian Molecular Clock Dating Using Genome-Scale Datasets. In:
400 Anisimova M. (eds) *Evolutionary Genomics. Methods in Molecular Biology*, vol 1910.
401 Humana, New York, NY, pp. 309-330.

402 Fet V., Gantenbein B., Gromov A.V., Lowe G., Lourenço W.R. 2003. The first molecular
403 phylogeny of Buthidae (Scorpiones). *Euscorpius* 4:1–10.

404 Fet V., Soleglad M, Lowe G. 2005. A new trichobothrial character for the high-level systematics
405 of Buthoidea (Scorpiones: Buthida). *Euscorpius* 23:1–40.

406 Frickey T., Lupas A. 2004. CLANS: a Java application for visualizing protein families based on
407 pairwise similarity. *Bioinformatics* 20:3702–3704.

408 He Y., Zhao R., Di Z., Li Z., Xu X., Hong W., Wu Y., Zhao H., Li W., Cao Z. 2013. Molecular
409 diversity of Chaerilidae venom peptides reveals the dynamic evolution of scorpion
410 venom components from Buthidae to non-Buthidae. *J. Proteomics* 89:1–14.

411 Herzig V., Sunagar K., Wilson D.T.R., Pineda S.S., Israel M.R., Dutertre S., McFarland B.S.,
412 Undheim E.A.B., Hodgson W.C., Alewood P.F., Lewis R.J., Bosmans F., Vetter I., King

PHYLOGENOMIC ANALYSIS OF SCORPIONS

413 G.F., Fry B.G. 2020. Australian funnel-web spiders evolved human-lethal δ -hexatoxins
414 for defense against vertebrate predators. Proc. Natl. Acad. Sci. USA 117: 24920-24928.

415 Holding M.L., Drabeck D.H., Jansa S.A., Gibbs H.L. 2016. Venom Resistance as a Model for
416 Understanding the Molecular Basis of Complex Coevolutionary Adaptations. Integr.
417 Comp. Biol. 56:1032–1043.

418 Hopp B.H., Arvinson R.S., Adams M.E., Razak K.A. 2017. Arizona bark scorpion venom
419 resistance in the pallid bat, *Antrozous pallidus*. PLoS ONE 12:e0183215.

420 King G.F., Hardy M.C. 2013. Spider-Venom Peptides: Structure, Pharmacology, and Potential
421 for Control of Insect Pests. Ann. Rev. Entomol. 58:475–96.

422 Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the
423 amino- acid replacement process. Mol. Biol. Evol. 21:1095–1109.

424 Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent- based species tree estimation with many
425 hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.

426 Mistry J., Finn R.D., Eddy S.R., Bateman A., Punta M. 2013. Challenges in Homology Search:
427 HMMER3 and Convergent Evolution of Coiled-Coil Regions. Nucleic Acids Res.
428 41:e121.

429 Niermann C., Tate T.G., Suto A.L., Barajas R., While H.A., Guswiler O.D., Secor S.M., Rowe
430 A.H., Rowe M.P. 2020 Defensive Venoms: Is Pain Sufficient for Predator Deterrence?
431 Toxins 12:260.

432 Nguyen L.T., Schmidt H.A., Haeseler A., Minh B.Q. 2014. IQ-TREE: a fast and effective
433 stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol.
434 32:268–274.

435 Ojanguren-Affilastro A.A., Adilardi R.S., Mattoni C.I., Ramírez M.J., Ceccarelli F.S. 2017.

436 Dated phylogenetic studies of the southernmost American buthids (Scorpiones;

437 Buthidae). *Mol. Phylogenet. Evol.* 110:39–49.

438 Puttick M.N. 2019. MCMCTreeR: functions to prepare MCMCTree analyses and visualize

439 posterior ages on trees. *Bioinformatics* 35(24):5321–5322.

440 Rapôso C. 2017 Scorpion and spider venoms in cancer treatment: state of the art, challenges, and

441 perspectives. *J. Clin. Transl. Res.* 3:233–249.

442 Rowe A.H., Xiao Y., Rowe M.P., Cummins T.R., Zakon H.H. 2013. Voltage-gated sodium

443 channel in grasshopper mice defends against bark scorpion toxin. *Science* 342:441–446.

444 Santibáñez-López C.E., Kriebel R., Ballesteros J.A., Rush N., Witter Z., Williams J., Janies

445 D.A., Sharma P.P. 2018. Integration of phylogenomics and molecular modeling reveals

446 lineage-specific diversification of toxins in scorpions. *PeerJ* 6:e5902–5923.

447 Santibáñez-López C.E., González-Santillán E., Monod L., Sharma P.P. 2019a. Phylogenomics

448 facilitates stable scorpion systematics: Reassessing the relationships of Vaejovidae and a

449 new higher-level classification of Scorpiones (Arachnida). *Mol. Phylogenet. Evol.*

450 135:22–30.

451 Santibáñez-López C.E., Graham M.R., Sharma P.P., Ortiz E., Possani L.D. 2019b. Hadrurid

452 Scorpion Toxins: Evolutionary Conservation and Selective Pressures. *Toxins* 11:637.

453 Santibáñez-López C.E., Ojanguren-Affilastro A.A., Sharma P.P. 2020. Another one bites the

454 dust: taxonomic sampling of a key genus in phylogenomic datasets reveals more non-

455 monophyletic groups in traditional scorpion classification. *Invertebr. Syst.* 34:133–143.

456 Santos M.S., Silva C.G.L., Neto B.S., Júnior C.R.P., Lopes V.H.G., Júnior A.G.T., Bezerra D.A.,

457 Luna J.V.C.P., Cordeiro J.B., Júnior J.G., Lima M.A.P. 2016. Clinical and

PHYLOGENOMIC ANALYSIS OF SCORPIONS

458 epidemiological aspects of scorpionism in the world: a systematic review. *Wild Environ.*
459 *Med.* 27:504–518.

460 Schwager E.E., Sharma P.P., Clarke T., Leite D.J., Wierschin T., Pechmann M., Akiyama-Oda
461 Y., Esposito L., Bechsgaard J., Bilde T., Buffry A.D., Chao H., Dinh H., Doddapaneni
462 H., Dungan S., Eibner C., Extavour C.G., Funch P., Garb J., Gonzalez L.B., Gonzalez
463 V.L., Griffiths-Jones, S., Han Y., Hayashi C., Hilbrant M., Hughes D.S.T., Janssen R.,
464 Lee S.L., Maeso I., Murali S.C., Muzny D.M., Nunes da Fonseca R., Paese C.L.B., Qu J.,
465 Roshaugen M., Schomburg C., Schönauer A., Stollewerk A., Torres-Oliva M., Turetzek
466 N., Vanthournout B., Werren J.H., Wolf C., Worley K.C., Bucher G., Gibbs R.A.,
467 Coddington J., Oda H., Stanke M., Ayoub N.A., Prpic N.M., Flot J.F., Posnien N.,
468 Richards S., McGregor A.P. 2017. The house spider genome reveals an ancient whole-
469 genome duplication during arachnid evolution. *BMC Biol.* 15:62.

470 Sharma P.P., Fernández R., Esposito L.A., González-Santillán E., Monod L. 2015.
471 Phylogenomic resolution of scorpions reveals multilevel discordance with morphological
472 phylogenetic signal. *Proc. R. Soc. B* 282:20142953.

473 Sharma P.P., Baker C.M., Cosgrove J.G., Johnson J.E., Oberski J.T., Raven R.J., Harvey M.S.,
474 Boyer S.L., Giribet G. 2018. A revised dated phylogeny of scorpions: Phylogenomic
475 support for ancient divergence of the temperate Gondwanan family Bothriuridae. *Mol.*
476 *Phylogen. Evol.* 122:37–45.

477 Sissom W.D. 1990. Sytematics, Biogeography, and Paleontology. In: *The Biology of Scorpions*
478 (ed: Polis GA). Stanford University Press, Stanford, CA, US, pp 31–80.

479 Suranse V., Sawant N.S., Paripatydar S.V., Krutha K., Paingankar M.S., Padhye A.D.,

480 Bastawade D.B., Dahanukar N. 2017. First molecular phylogeny of scorpions of the

481 family Buthidae from India. *Mitochondrial DNA A* 28:606–611.

482 Tropea G., Onnis C. 2020. A remarkable discovery of a new scorpion genus and species from

483 Sardinia (Scorpiones: Chactoidea: Belisariidae). *Riv. Aracnol. Ital.* 26:3–25.

484 Upham N.S., Esselstyn J.A., Jetz W. 2019. Inferring the mammal tree: Species-level sets of

485 phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol.* 17:

486 e3000494.

487 Veiseh M., Gabikian P., Bahrami S.B., Veiseh O., Zhang M., Hackman R.C., Ravanpay A.C.,

488 Stroud M.R., Kusama Y., Hansen S.J., Kwok D., Munoz N.M., Sze R.W., Grady W.M.,

489 Greenberg N.M., Ellebogen R.G., Olson J.M. 2007. Tumor Paint: A Chlorotoxin:Cy5.5

490 Bioconjugate for Intraoperative Visualization of Cancer Foci. *Cancer Res.* 67:6882–6888.

491 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*

492 24:1586–1591.

493

494

Figure Legends

Figure 1. Exemplars of scorpion diversity in the two major parvorders. Top row, buthid species. Bottom row, Iurida species. Photos: *Buthus israelensis* and *Androctonus crassicauda* (R. Livne); *Buthacus leptochelys*, *Palaeocheloctonus pauliani* and *Opistophthalmus carinatus* (J. Ove Rein); *Centruroides meisei* and *Tityus serrulatus* (B. Myers); *Belisarius xambeui* (G. Giribet); *Hadrurus obscurus* and *Anuroctonus bajae* (C. Santibáñez-López). All photographs published with permission.

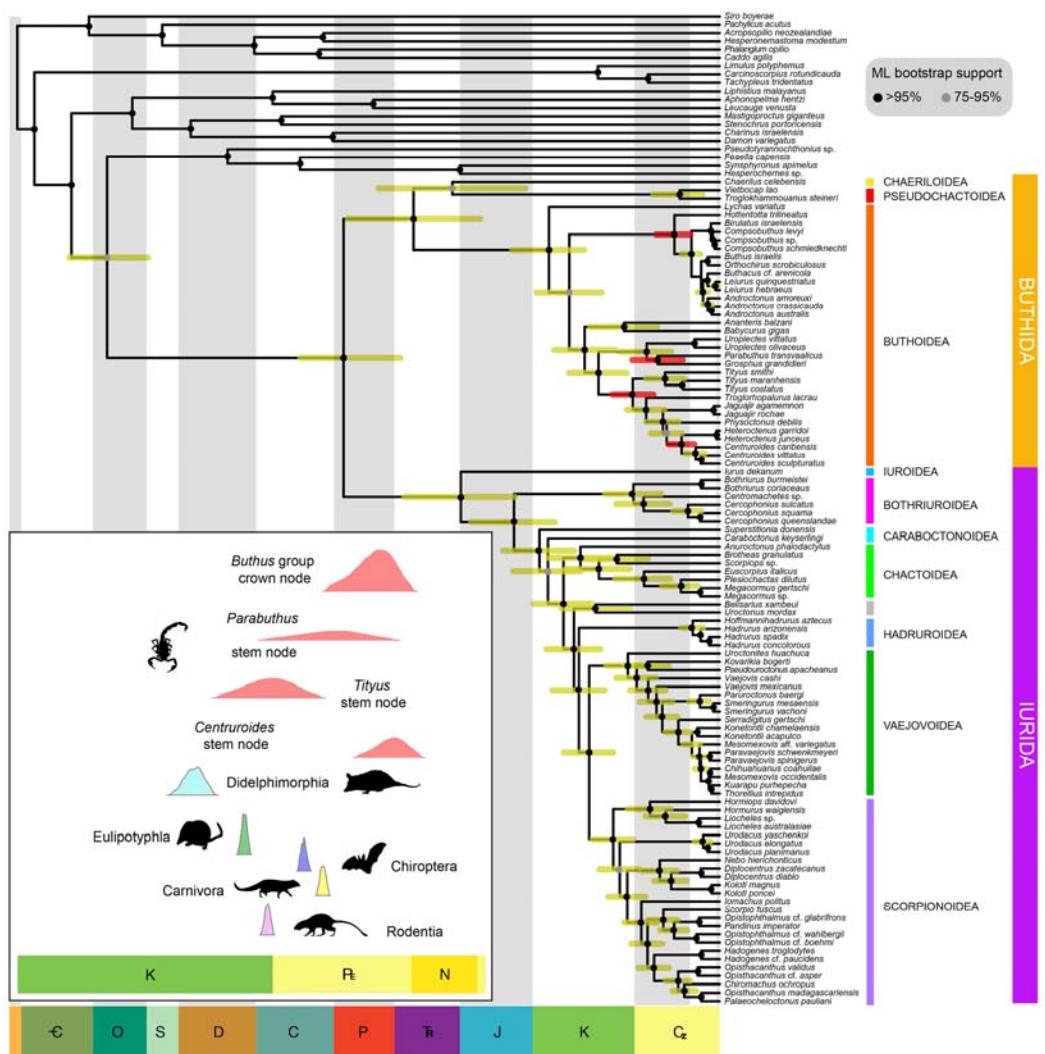


Figure 2. Chronogram of scorpion relationships derived from maximum-likelihood analysis of 192 loci (53,333 amino acid sites). Node ages are computed in a time-calibrated analysis using 15 fossil calibrations. Blue bars depict 95% credibility intervals of node ages, whereas red bars depict 95% credibility intervals for the most toxic buthids. Inset: Posterior distribution of node ages corresponding to mammal-specific toxin origins (in red), compared to the posterior distribution of the four major mammal orders that include scorpion predators. Mammal order ages from (43).

PHYLOGENOMIC ANALYSIS OF SCORPIONS

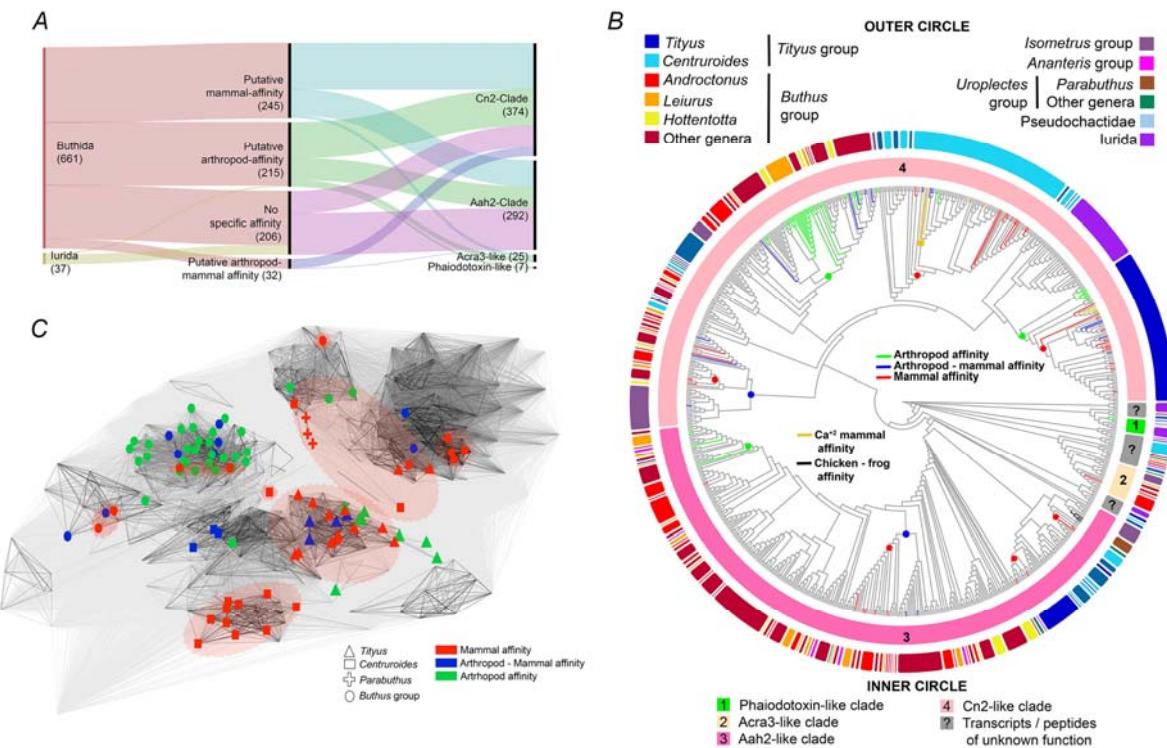


Figure 3. Evolutionary analyses of the sodium channel toxin family (NaTx). (A) Alluvial plot summarizing the affinity of the NaTx found in scorpion venom (center) of each parvorder (left), and their subtype (right). Numbers represent the total transcripts and/or peptides found in our transcriptomic analyses and UniProt. (B) NaTx affinity (inner bars) plotted onto the NaTx gene tree subdivided into the seven subclades (four well known plotted onto the alluvial plot in panel A). In gray: transcript clades with unknown function. Circles in nodes represent the putative ancestral function as a result of the most parsimonious ancestral state reconstruction. (C) Three-dimensional CLANS clustering of the mature peptide amino acid sequence. Colors for NaTx affinity and lines indicating pairwise similarity. Red dotted lines show seven distinctive clusters that include at least one peptide with known mammal affinity.

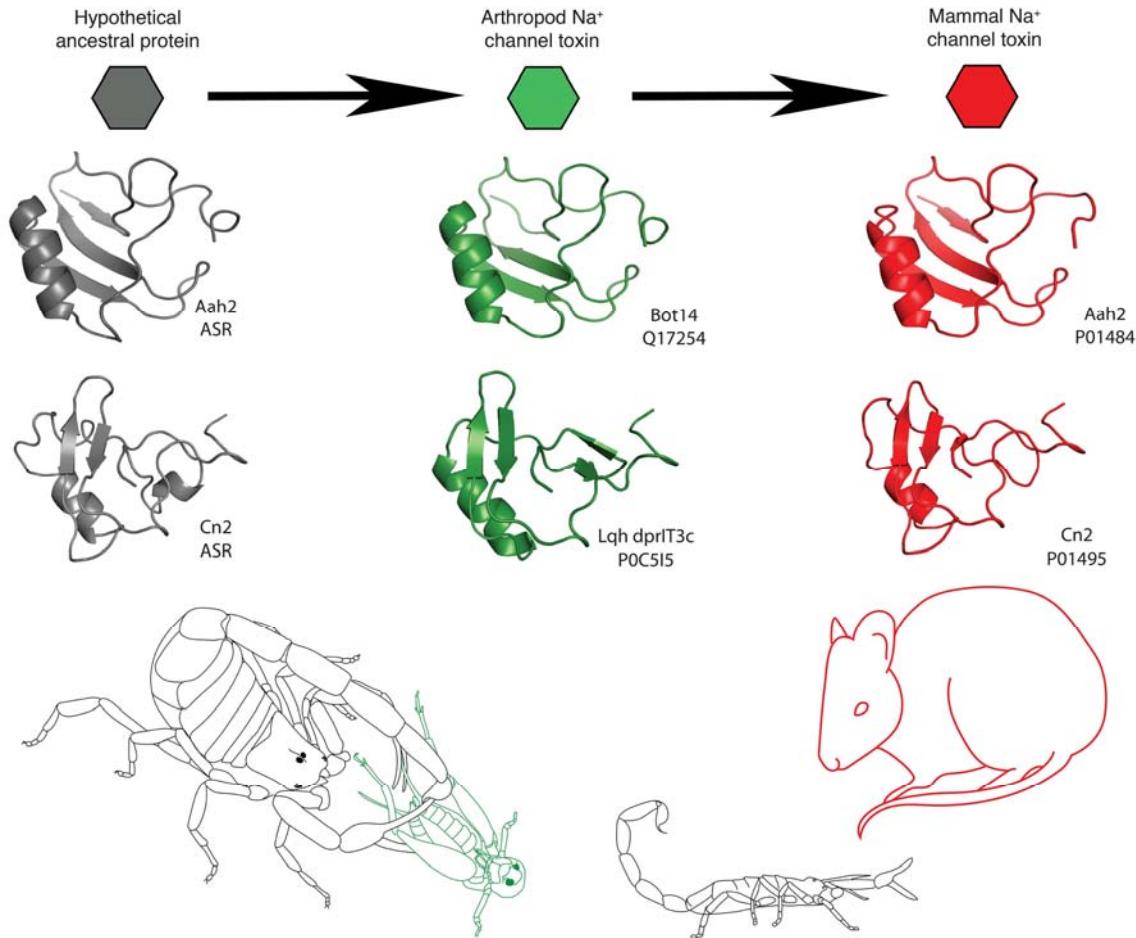


Figure 4. Secondary structures and proposed evolutionary pathway of scorpion Na^+ channel toxins from the Aah (top model) and Cn2 (bottom model) clades. Our results suggest mammal specific toxins are derived from insect specific ones, which in turn evolved from ancestral molecules of unknown physiological activity. Mammal specific toxins from *A. australis* (pdb file 1ptx) and *C. noxius* (pdb file 1cn2) are shown in red; the model for the insect specific toxins from *B. tunetanus* (Q17254) and *L. hebraeus* (POC515) are shown in green; and, the hypothetical ancestral molecules of unknown function are shown in gray. The structures of these ancestral proteins are derived from the primary sequences estimated using Ancestral State Reconstruction.