



HAL
open science

A phylogeny-aware approach reveals unexpected venom components in divergent lineages of cone snails

Alexander E. Fedosov, Paul Zaharias, Nicolas Puillandre

► To cite this version:

Alexander E. Fedosov, Paul Zaharias, Nicolas Puillandre. A phylogeny-aware approach reveals unexpected venom components in divergent lineages of cone snails. *Proceedings of the Royal Society B: Biological Sciences*, 2021, 288 (1954), 10.1098/rspb.2021.1017 . hal-03280348

HAL Id: hal-03280348

<https://hal.science/hal-03280348>

Submitted on 7 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

PROCEEDINGS OF THE ROYAL SOCIETY B

BIOLOGICAL SCIENCES

A phylogeny-aware approach reveals unexpected venom components in divergent lineages of cone snails.

Journal:	<i>Proceedings B</i>
Manuscript ID	RSPB-2021-1017.R1
Article Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Fedosov, Alexander; Russian Academy of Sciences; Muséum National d'Histoire Naturelle, Institut Systématique Evolution Biodiversité (ISYEB) Zaharias, Paul; Muséum National d'Histoire Naturelle, Institut Systématique Evolution Biodiversité (ISYEB); University of Illinois Urbana, Department of Computer Science Puillandre, Nicolas; Muséum National d'Histoire Naturelle, Institut Systématique Evolution Biodiversité (ISYEB)
Subject:	Bioinformatics < BIOLOGY, Evolution < BIOLOGY, Genomics < BIOLOGY
Keywords:	conotoxins, venom evolution, transcriptomics, Conidae, Conasprella, Pygmaeonus
Proceedings B category:	Genetics & Genomics

SCHOLARONE™
Manuscripts

Author-supplied statements

Relevant information will appear here if provided.

Ethics

Does your article include research that required ethical approval or permits?:

This article does not present research with ethical considerations

Statement (if applicable):

CUST_IF_YES_ETHICS :No data available.

Data

It is a condition of publication that data, code and materials supporting your paper are made publicly available. Does your paper present new data?:

Yes

Statement (if applicable):

Data accessibility. The transcriptomic sequencing data are deposited in the NCBI SRA database, under the Bioproject PRJNA735765. Sequences of the predicted toxins are provided in supplementary data (Table S2, Figures S1 – S8), Python scripts are available at https://github.com/Hyperdiverseproject/Divergent_Conidae.

Conflict of interest

I/We declare we have no competing interests

Statement (if applicable):

CUST_STATE_CONFLICT :No data available.

Authors' contributions

CUST_AUTHOR_CONTRIBUTIONS_QUESTION :No data available.

Statement (if applicable):

CUST_AUTHOR_CONTRIBUTIONS_TEXT :No data available.

1 **A phylogeny-aware approach reveals unexpected venom components in**
2 **divergent lineages of cone snails.**

3

4 Alexander Fedosov^{1,2}, Paul Zaharias^{2,3}, Nicolas Puillandre²

5 ¹ – A.N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, Leninsky
6 prospect 33, 119071 Moscow, Russian Federation, fedosovalexander@gmail.com

7 ² – Institut Systématique Evolution Biodiversité (ISYEB), Muséum National d’Histoire Naturelle,
8 CNRS, Sorbonne Université, EPHE, Université des Antilles, 57 rue Cuvier, CP 26, 75005 Paris,
9 France.

10 ³ – Department of Computer Science; University of Illinois Urbana-Champaign, Urbana IL 61801,
11 USA.

12

13 **Abstract**

14 Marine gastropods of the genus *Conus* are renowned for their remarkable diversity and deadly
15 venoms. While *Conus* venoms are increasingly well studied for their biomedical applications, we
16 know surprisingly little about venom composition in other lineages of Conidae. We performed
17 comprehensive venom transcriptomic profiling for *Conasprella coriolisi* and *Pygmaeonus*
18 *trillii*, first time for both respective genera. We complemented reference-based transcriptome
19 annotation by a *de novo* toxin prediction guided by phylogeny, which involved transcriptomic
20 data on two additional ‘divergent’ cone-snail lineages, *Profundiconus*, and *Californiconus*. We
21 identified toxin clusters (SSCs) shared among all or some of the four analysed genera based on
22 the identity of the signal region – a molecular tag present in toxins. In total 116 and 98 putative
23 toxins represent 29 and 28 toxin gene superfamilies in *Conasprella* and *Pygmaeonus*,
24 respectively; about quarter of these only found by semi-manual annotation of the SSCs. Two
25 rare gene superfamilies, originally identified from fish-hunting cone-snails, were detected
26 outside *Conus* rather unexpectedly, so we further investigated their distribution across Conidae
27 radiation. We demonstrate that both these, in fact, are ubiquitous in Conidae, sometimes with
28 extremely high expression. Our findings demonstrate how a phylogeny-aware approach
29 circumvents methodological caveats of the similarity-based transcriptome annotation.

30 Keywords: conotoxins, venom evolution, phylogeny, Conidae, *Conasprella*, *Pygmaeonus*

31 **Introduction**

32 Marine gastropods of the genus *Conus* are renowned for their remarkable diversity [1,2] and
33 complex hunting strategies enabled by elaborated and deadly venoms. *Conus* venoms comprise
34 highly-diversified neuro-peptides (conotoxins), hormones, and small molecules in species-
35 specific combinations that are suited to the biology of the prey and associated to particular
36 hunting strategies [3–7]. Currently, *Conus* venoms are being studied at an ever-increasing rate
37 because of their potential to be developed as drug leads. This capitalizes on their ability to
38 modulate or disrupt the functioning of ion-channels and receptors in the nervous system of
39 prey or potential predators, including vertebrates [8,9]. However, *Conus* venoms are equally
40 interesting from an evolutionary biology prospective. Generally, each toxin constitutes an
41 adaptive trait that possesses a single function, and can be easily quantified and remapped onto
42 the genome [10,11]. These properties, magnified by the impressive species diversity in *Conus*,
43 and by the documented complexity of venom in each species [1], make *Conus* venoms an ideal
44 model for studying drivers and dynamics of molecular evolution.

45 Eight genera (herein referred to as *divergent Conidae*) are currently included in the family
46 Conidae, in addition to *Conus* [12]. Among them, recent phylogenetic studies on the family
47 [13,14] demonstrate that the genera *Conasprella*, *Californiconus*, *Pygmaeonus*, and *Lilliconus*
48 form a separate lineage, sister to *Conus*, whereas *Profundiconus* is found to be the earliest
49 diverging lineage of the family (Fig. 1a). Current knowledge of venom composition among these
50 taxa is highly skewed [15]. While a wealth of data is available for *Conus* species and for the
51 single species of *Californiconus*, *C. californicus* [5,16], only fragmentary data has been published
52 for *Conasprella* [17,18], the first and only analysis of *Profundiconus* venom was published only
53 recently [19], and virtually no data exist for other divergent Conidae. To fill this gap, we carried
54 out the first comprehensive transcriptomic analyses for *Conasprella* and *Pygmaeonus* and we
55 present the results here. Our results will foster analyses of the apparition and diversification of
56 the venom component across the Conidae radiation. This is an important milestone on the way
57 to understanding cone-snail venom evolution.

58 Venoms of divergent Conidae have remained poorly studied for a reason. Unlike *Conus*, these
59 genera are much less speciose, most of their species are rare and either dwell in deep-water
60 (most *Conasprella*, *Profundiconus*), or have very restricted distribution (*Lilliconus*), or are very
61 small (*Pygmaeonus* and *Lilliconus*). These factors complicate sampling, and recovery of even
62 one live specimen suitable for venom profiling is generally a stroke of luck, and it was the case

63 with *Conasprella coriolisi* (Moolenbeek & Richard, 1995) and *Pygmaeonus traillii* (A. Adams,
64 1855) analysed here. This limited sampling posed a challenge to corroborating sets of predicted
65 venom transcripts. Divergent lineages are expected to possess divergent venom components
66 when compared to *Conus* venoms. Consequently, they are more difficult to identify by
67 conventional approaches in peptide annotation based on sequence similarity and structural
68 features [20,21], and increasingly rely on the *de novo* annotation. With only one specimen per
69 species available, accuracy of the *de novo* toxin prediction cannot be cross-validated by data
70 from independently sequenced conspecific specimens, and predicted venom components also
71 cannot be verified by means of proteogenomics [20]. So, to provide more robustness to the *de*
72 *novo* transcript annotation, we tested a phylogeny-aware approach. This approach first helped
73 to identify the divergent lineages in which one would expect to find divergent venom
74 components. Next, the phylogeny was used to identify related taxa among which the clusters of
75 transcripts predicted as venom components can be cross-validated. Finally, a phylogenetic
76 approach was also legitimate in tackling the diversity of venom peptides, where the sequence
77 of the signal region can serve as a proxy for precursor classification into gene superfamilies
78 [22,23]. In essence, the same principles make up the theoretical framework of Concerted Toxin
79 Discovery [24,25], which, however, has never been convincingly performed within one study.
80 Here we show that by applying this approach, a large fraction of the venom transcript diversity
81 overlooked by reference-based annotation can be identified. Furthermore, we demonstrate
82 that some of these novel clusters are also present, and may be quite diversified, in *Conus*. Yet
83 they have been barely noticed thus far because of the methodological caveats of the similarity-
84 based transcriptome annotation. Finally, we discuss the impact of such previously undetected
85 venom components on hypotheses related to the evolution of the cone snails and their toxins.

86 **Materials and methods**

87 *Specimen collection*

88 The specimen MNHN-IM-2013-47769, *Pygmaeonus traillii* hereafter *Pygmaeonus*, was
89 sampled in shallow waters off New Ireland during the KAVIENG 2014 expedition
90 (expeditions.mnhn.fr). It was photographed and dissected alive. The venom gland was
91 immediately suspended in RNAlater solution (Thermo Fisher Scientific, Waltham, MA, USA),
92 stored overnight at room temperature, and subsequently at -20°C. The specimen MNHN-IM-
93 2013-66001, *Conasprella coriolisi*, hereafter *Conasprella*, was collected by dredging at depths of
94 270-275 meters during the KANACONO expedition (doi 10.17600/16003900;
95 expeditions.mnhn.fr) off New Caledonia, west of the Isle of Pines. It was kept in chilled
96 seawater, and dissected alive upon arrival to the onshore laboratory. The venom gland was also
97 preserved in RNAlater and stored at -20°C.

98

99 *RNA extraction and sequencing*

100 RNA was extracted from venom glands of *Pygmaeonus* and *Conasprella* using the TRIzol
101 reagent (Thermo Fisher Scientific) following the protocol provided by the manufacturer.
102 Bioanalyzer traces were used to assess total RNA quality and determine suitability for
103 sequencing. The cDNA libraries were prepared and sequenced either at the New York Genome
104 Center (*Conasprella*) or at the Vincent J. Coates Genomics Sequencing Laboratory at UC
105 Berkeley (*Pygmaeonus*). In New York, libraries were prepared using the automated polyA
106 RNAseq library prep protocol and sequenced with Illumina HiSeq 4000 with 150-bp paired-end
107 reads, resulting in the acquisition of 15,0298,52 150-bp paired-end reads. In Berkeley, the KAPA
108 Stranded mRNA-Seq kit was used to synthesize cDNA, ligate adapters using TruSeq HT adapters
109 and barcode samples, and sequenced on the Illumina HiSeq 4000 system, resulting in the
110 acquisition of 30,063,937 100-bp paired-end reads

111

112 *Transcriptome assembly*

113 Adaptor removal and quality trimming of the *Conasprella* and *Pygmaeonus* raw reads were
114 performed using Trimmomatic v.0.36 [26] with the following parameters: ILLUMINACLIP option
115 enabled, seed mismatch threshold = 2, palindrome clip threshold = 40, simple clip threshold of

116 15; SLIDING WINDOW option enabled, window size = 4, quality threshold = 15; MINLEN = 36;
117 LEADING = 3; TRAILING = 3. The reads were then assembled using Trinity v2.11.0 [27] with the
118 kmer size set to 31, which performs best to assemble venom gland transcriptomes of *Conus*
119 [28,29]. The assembly metrics were checked using the TrinityStats.pl module. The same
120 parameters were used to trim and assemble raw read data on *Profundiconus neocaledonicus*
121 [19], hereafter *Profundiconus*, and *Californiconus californicus* [5], hereafter *Californiconus*
122 (Table S1). To quantify the abundance of the predicted transcripts we used the function rsem-
123 calculate-expression [30], with bowtie2 [31] mapper to map the trimmed reads on the
124 assembly. Transcripts-per-kilobase-million (tpm) values were used, as they are recognized as
125 the most appropriate metrics of expression levels [5,23].

126

127 *Identification of putative conotoxin precursors*

128 We applied three approaches to identify potential toxin transcripts in the assembled
129 transcriptomes. First, we conducted a direct BLASTx search of the *Conasprella* and
130 *Pygmaeonus* assemblies against an in-house toxin database. This database was obtained by
131 combining all entries with the keyword 'toxin' from UniProt with all entries from ConoServer
132 [32], and supplemented by lists of putative gastropod toxins of the tonnoidean *Charonia*
133 *tritonis* [33], buccinoideans *Cumia reticulata* [34] and *Hemifusus tuba* [35], non-conid
134 conoideans *Clavus canalicularis* and *C. davidgilmouri* [36], and cone-snails *Profundiconus* spp.
135 [19], *Conus ermineus* [23], *Conus magus* [37], *Conus tribblei* [38], *Conus praecellens* [29], *Conus*
136 *betulinus* [39], and *Conus litteratus* [40] plus the 15 species of *Conus* and *Californiconus*,
137 analysed by Phuong et al.[5]. Then alignments of the contigs that produced reliable hits (BLAST
138 PID >0.55, with aligned length no less than half of the best matching database entry, and with
139 no stop codons) were parsed from the XML output by the Python script1 and checked visually.
140 After removal of flanking regions, these predicted transcripts were combined in dataset 1,
141 comprising toxins with high sequence identity to known animal toxins, primarily, conotoxins.
142 Subsequently, we performed Coding DNA Sequence (CDS) prediction, using ORFfinder [41].
143 Only CDSs comprising 35 or more amino acid residues, and starting with either 'ATG' or
144 alternative initiation codons were output. We ran SignalP v5.0 [42] to identify a subset of CDSs
145 with signal region prediction (D value > 0.7), and further filtered this subset to remove CDSs
146 with a transmembrane domain (identified by Phobius v1.01 [43]). Then we removed redundant

147 CDSs derived from predicted alternative isoforms of the same transcript: CDSs showing less
148 than two AA-residue divergence were removed to keep only the CDS corresponding to the most
149 highly expressed isoform (in-house Python script²). The resulting catalog of CDSs was used for
150 structure-based search via HMMER v3.2.1 [44] against the Pfam database [45]. The CDSs with
151 HMMER hits were then sorted, based on the relevance of the HMMER annotations to the
152 venom functions. These annotated CDSs made up dataset 2. In general, three broad classes
153 were recognized: toxins (t), hormones (h), and enzymes and other peptides with known or
154 proposed function in envenomation (p).

155 Datasets 1 and 2 contained sequences with detectable sequence identity and/or with structural
156 similarity to known venom peptides. But, we expected from the divergence between *Conus* and
157 the analysed divergent Conidae that highly divergent clusters of venom components were
158 lacking from these datasets. As we had only one specimen per species analysed, we employed a
159 phylogenetic approach to *de novo* toxin identification. The scope of the search was defined to
160 include all lineages of Conidae outside the genus *Conus*, i.e., the genera *Profundiconus*,
161 *Conasprella*, *Californiconus*, and *Pygmaeonus* (Fig. 1a). First, the non-identical CDSs, containing
162 a signal sequence, but no transmembrane domains were recovered from the reassembled
163 datasets of *Californiconus californicus* and *Profundiconus neocaledonicus*, following same
164 methodology, as for *Conasprella* and *Pygmaeonus*. The trimmed reads were then remapped
165 on these CDSs and coverage-per-base was calculated, as a measure of reliability of predicted
166 CDSs. Those CDSs with a smallest per-base coverage value below 3 were removed from the
167 dataset. The signal sequences of the thus filtered CDSs (numbering in total 16,906) were pooled
168 into a single file and clustered using CD-Hit v4.8.1 [46]. The signal sequence is the most
169 conserved region of a conotoxin precursor and is widely used for conotoxin classification [47],
170 and phylogenetic clustering of gene superfamilies [22]. Therefore, we considered the recovered
171 signal sequence-based clusters (hereafter, SSC) as potential gene (super)families of secreted
172 peptides. The three alternative identity thresholds of 0.6, 0.65, and 0.7 used for clustering
173 correspond to the range of signal sequence PID in most canonical gene superfamilies based on
174 ConoServer [47]. The alternative clustering schemes were evaluated based on the already
175 annotated transcripts from the dataset 1, to make sure that transcripts representing distinct
176 toxin gene superfamilies, on no occasion end up in a single SSC. The set of 13,616 SSCs obtained
177 with the identity threshold of 0.65 was found to best separate known gene superfamilies, so it
178 was selected for the subsequent analyses. All the SSCs containing three or more transcripts,

179 represented in at least two genera of divergent Conidae, and showing moderately-high
180 (100<tpm<1,000) or high (tpm > 1,000) expression levels in at least one genus were identified
181 (in-house Python script3). These SSCs were aligned separately using MAFFT v7.475 [48] with G-
182 INS-1 strategy and 'unalignlevel' parameter set to 0.2. The cleavage sites were predicted by
183 ConoPrec [47], and the Cys-patterns were identified by in-house Python script4. When
184 screening such clusters, the following conditions were checked: i) predicted signal sequence
185 lacking long repeats of one or two residues, such as 'LLLLLLLLL', 'LSLSLSLSLSLS' or
186 'VSVSVSVSVSVSV', ii) complete precursor not exceeding 200 AA, iii) mature region comprising
187 over 20 AA, iv) consistent alignment features within each SSC. Some identified clusters might,
188 however, correspond to transcripts of house-keeping genes, including transcripts translated
189 into the wrong frame [37]. To filter these out, we used BLASTx (E-value 10⁻⁵) to search the
190 nucleotide sequences of the SSCs against the SwissProt manually curated database [49]. The
191 clusters that did not return a match were aligned to the best-match ConoServer entry using the
192 built-in amino acid search tool, and either ascribed to known venom peptide gene
193 superfamilies, or designated as novel gene superfamilies. The transcripts identified by the
194 analysis of SSCs and lacking from datasets 1 and 2 formed dataset 3. The final lists of venom
195 peptides were compiled for *Conasprella* and *Pygmaeonus* by combining datasets 1, 2 and 3.

196

197 *Analysis of distribution of novel gene superfamilies in species of Conus*

198 To determine whether the novel gene superfamilies identified from the SSCs in divergent
199 Conidae are also present in *Conus* species, we first reassembled 15 *Conus* transcriptomes (Table
200 S1), and remapped trimmed reads to the resulting assemblies, using the same methodology as
201 for the divergent Conidae datasets. We then ran CAP3 (with default parameters) followed by
202 CD-Hit (PID 99%) to reduce assembly redundancy. The clustered assemblies were used in
203 BLASTx against the database of novel gene superfamilies from dataset 3. To roughly estimate
204 the contribution of the novel gene superfamilies to the toxin expression in each *Conus* species,
205 one known highly expressed conotoxin gene superfamily was selected for each species to serve
206 as a reference (Table S1). All available sequences of this gene superfamily (specifically including
207 sequences identified in the original study) were added to the new gene superfamily database.
208 The BLAST results (E-value of -10) were first sorted by in-house Python script5 in the following
209 manner: i) the query transcripts were assigned to the reference superfamily if the PID exceeded
210 85% while the aligned length constituted no less than 0.7 of the length of the best-matching

211 entry from the BLAST database, and tpm expression level exceeded 5; ii) query transcripts were
212 provisionally assigned to a novel gene superfamily if PID exceeded 30%, while aligned length
213 was no less than 0.6 of the length of the best matching sequence in the database, and the tpm
214 exceeded 5. Transcripts that fulfilled these conditions were aligned by gene superfamily, and
215 then each alignment was screened to remove erroneously assigned transcripts. Then
216 expression levels were summed up for each gene superfamily, and relative expression (in
217 percent) calculated. Based on these data we performed a principal component analysis (PCA) to
218 evaluate the degree of venom composition similarities among the analysed Conidae. The PCA
219 diagram was constructed with PAST v4.06 [50], using the variance-covariance method.

220

221

222 **Results**

223 *a. Venom composition in Conasprella and Pygmaeconus*

224 Direct similarity search with BLASTx identified 96 *Conasprella* transcripts with high similarity
225 (BLAST PID above 55%) to known *Conus* venom components included in the in-house toxin
226 database. Among these, 80 were counterparts of *Conus* venom peptides referable to 17 gene
227 superfamilies (Table S2). The search of predicted CDS against the Pfam-A HMM domain
228 database revealed a diversity of additional transcripts with proposed functions in venom. Most
229 numerous among them were transcripts bearing Von Willebrand factor domains, and various
230 peptidases (M, C, S), both with typically low expression levels (tpm <40). A total of 81 predicted
231 *Pygmaeconus* venom transcripts were identified by BLASTx. Of these only 44 were counterparts
232 of the *Conus* venom peptides and represented 15 gene superfamilies. Among other revealed
233 components, most notable were diversified transcripts with high similarity to neuropeptides:
234 APWGamide, cerebrin, elevenin, FFamide, FMRamide, FxRIamide, LASGLVamide-4, LFRamide-
235 2, NdWFamide-2, Wwamide, all with low expression (tpm <20), except NdWFamide-2 (tpm 120)
236 – Table S2. HMMER analysis predicted a diversity of additional peptides with previously
237 suggested functions in venom: astacin, peptidases C, M, S, and trypsin-like, chitinase, CAP, ShK
238 and peptides containing Von Willebrand-like domains.

239 The 13,616 SSCs were filtered to select only those comprising three or more predicted
240 transcripts, found in at least two species of divergent conid genera (Fig. 1b), and highly
241 expressed in at least one genus. The 90 SSCs that fulfilled these criteria were manually curated

242 to exclude clusters comprising transcripts that did not show features of toxins, leading to a final
243 set of 71 SSCs. Of these, 37 clusters contained CDSs identified by direct BLASTx from
244 *Conasprella* and/or *Pygmaeonus*, and already assigned to known gene superfamilies. Another
245 22 SSCs comprised transcripts assigned to known toxin gene superfamilies but missed by
246 BLASTx. These clusters added 25 (23.6%) and 19 (29.7%) new transcripts of known gene
247 superfamilies to the catalogs of *Conasprella* and *Pygmaeonus*, respectively. Among them,
248 seven and five known conotoxin gene superfamilies in *Conasprella* and *Pygmaeonus*,
249 respectively, were only identified from these SSCs, and were thus lacking in datasets 1 and 2.
250 Seven more SSCs are diversified in the divergent Conidae genera, are highly expressed in at
251 least one studied transcriptome, and demonstrate canonical conotoxin precursor structure
252 [15], but do not show similarity to any established conotoxin superfamily. These are designated
253 as new conotoxin superfamilies DivCon 1-7. The transcripts of known and of newly designated
254 gene superfamilies of conotoxins constitute the datasets 1 and 3. In *Conasprella*, datasets 1 and
255 3 contain 74 and 24 transcripts respectively, whereas in *Pygmaeonus* – 40 and 47 transcripts
256 respectively.

257 The final catalogs comprise 170 and 190 venom components classified in 29 and 28 gene
258 superfamilies in *Conasprella* and *Pygmaeonus*, respectively (Table S2). Of them, 116 and 98
259 transcripts from *Conasprella* and *Pygmaeonus*, respectively, represent known or new
260 conopeptide gene superfamilies, classified into four groups: i) 'canonical' gene superfamilies
261 and common classes of *Conus* venom peptides, such as conkunitzin, conodipine, conophysin,
262 conoporin, ii) 'divergent' gene superfamilies, largely known from *Californiconus californicus*
263 [5,16], plus the recently identified very taxonomically restricted gene superfamilies New-Geo-1
264 [51], and Pmag02 [37] iii) novel gene superfamilies, sharing structural properties of conotoxins,
265 iv) putative conotoxins (Fig. 2). The latter group comprises unrelated transcripts with structural
266 similarity to known conotoxins detected by HMMER, which are not assigned to any gene
267 superfamily. A total of 86 predicted *Conasprella* transcripts were assigned to 16 'canonical'
268 gene superfamilies, with dominating P- (18 transcripts), M- (11), O2- (9), and I2- (8) gene
269 superfamilies (Fig. 2a). A total of 19 transcripts are identified in six 'divergent' gene
270 superfamilies accounting for 20.3% of the summed toxin expression, and seven predicted
271 transcripts in four novel gene superfamilies account for only 2.45% of the summed toxin
272 expression. Notably fewer transcripts of the 'canonical' gene superfamilies (47) are identified in
273 *Pygmaeonus*, and almost half of them (21 transcripts) represent the T- superfamily, followed

274 by O1- (6), L- (3) and O3- (3). Among the seven 'divergent' gene superfamilies identified in
275 *Pygmaeonus*, the Divergent-MSTLGMTLL (8 transcripts) is the most diversified and is by far the
276 most highly expressed (18.6% of the summed toxin expression). Novel gene superfamilies in
277 *Pygmaeonus* are represented by 28 predicted transcripts, that contribute 15.2% to the
278 summed toxin expression.

279

280 *b. Novel gene superfamilies identified through clustering of the signal region*

281 Seven SSCs, comprising in total 55 putative toxins of divergent Conidae are designated as novel
282 gene superfamilies (Supplementary figures 1-7, Table S3). All the predicted transcripts of
283 DivCon2 are cysteine-free, and DivCon3, DivCon5, and DivCon7 show conserved arrangement of
284 cys residues. The remaining three gene superfamilies vary in arrangement of Cys residues in the
285 mature peptide region, and members of DivCon1 and DivCon4 also display highly divergent pre-
286 and mature peptide regions. However, variations in the length and cys pattern are also found in
287 many 'canonical' conotoxin gene superfamilies, (i.e. A-, I2-, M-, O1-), and is reflected in
288 diversified functions of the included gene families (see e.g. [37]).

289 Among the novel toxin gene superfamilies, ConDiv3 is notable for the peculiar sequence of its
290 six members, bearing an Arg-Phe-Gly motif (RF-amide) C-terminally. ConDiv3 is only detected in
291 the transcriptomes of the *Californiconus* and *Pygmaeonus* clade (Fig. 1a), and in the former is
292 represented by a single low expression transcript (tpm 9.34). Of five transcripts identified in
293 *Pygmaeonus*, three show high expression, with tpm values exceeding 1,000-1,500. These three
294 precursors Pyg6, Pyg9 and Pyg11 form a distinct cluster (Fig. 3) and are distinctive in that they
295 possess a pre-region (underlined in Figure 3) and have an internal cleavage site within the
296 predicted mature peptide region.

297

298 *c. Diversity and expression of novel gene superfamilies in Conus*

299 We hypothesized that members of the novel gene superfamilies DivCon 1-7 may also be
300 present in *Conus*, but overlooked, since their published transcriptome annotations mainly relied
301 on similarity-based search (BLAST). Furthermore, the discovery of the gene superfamilies New-
302 Geo-1 and Pmag02 in divergent Conidae requires corroboration. We therefore evaluated the
303 distribution of these gene superfamilies in 15 species of *Conus* representative of both its

304 phylogenetic diversity (12 different subgenera) and the known dietary guilds (worm-, fish- and
305 mollusk-hunters).

306 Counterparts of DivCon2 are identified in transcriptomes of six *Conus* species, with a single
307 transcript in each species, usually with moderately-high expression levels ($100 < \text{tpm} < 1,000$).

308 Additional DivCon2 members are revealed in what Li et al. [21] referred to as the ‘putative
309 MTKLL’ gene superfamily in *Conus lenavati*, *C. characteristicus* and *C. betulinus* (Supplementary
310 figure 2). Similarly, incomplete precursors, but ones obviously closely related to those in the
311 DivCon7 superfamily, are detected in *Conus arenatus* and *C. sponsalis*. The precursor Im20.1 of
312 *C. imperialis* [52] is also clearly referable to the ConDiv7 superfamily (Supplementary figure 6).

313 The gene superfamilies New-Geo-1 and Pmag02 are present in all, and DivCon6 – in almost all
314 the analysed *Conus* species. Both the Pmag02 and the New-Geo-1 superfamilies are
315 represented by multiple transcripts per species. Even after removal of the minor isoforms that
316 are less than 2 AA residues divergent from the closest major isoform of the same species, final
317 datasets of Pmag02 and the New-Geo-1 comprise 34 and 83 sequences, respectively. The New-
318 Geo-1 superfamily reaches highest relative expression in *Conus ebraeus*, *C. sponsalis*, *C. textile*
319 and *C. tribblei*, and contributes about 1% of the total toxin expression in each of these species
320 (Fig. 4a, table S4). Pmag02, in general shows even higher expression, contributing 1-3% to the
321 summed expression in most *Conus* species, but with a maximum of notable 15% in *C.*
322 *marmoreus*.

323 Discussion

324 The venoms in *Conasprella* and *Pygmaeonus* differ notably from each other in terms of
325 dominant venom gene superfamilies: P- and O3- in *Conasprella*, versus T- and L- in
326 *Pygmaeonus*. PCA analysis of the data in table S4 suggests that the *Pygmaeonus* venom is
327 highly divergent from other Conidae venoms (Fig. 4b). This can be explained by the
328 phylogenetic distinctiveness of *Pygmaeonus*, but also by the small size of the animal compared
329 to all other Conidae included in the analysis. *Pygmaeonus* venom evolution might have been
330 driven by adaptation to an uncommon niche among Conidae, and thus to a different spectrum
331 of interactions from those in larger Conidae. Further studies on the feeding biology and diet of
332 both *Conasprella* and *Pygmaeonus* are needed to corroborate this hypothesis.

333 The 98 and 58 transcripts of known gene superfamilies identified for *Conasprella* and
334 *Pygmaeonus* respectively are well within the diversity range reported in the single-
335 transcriptome studies on *Conus*. These numbers slightly exceed those reported for
336 *Profundiconus neocaledonicus* (55 -[19]), but a much higher diversity is reported for
337 *Californiconus* (185 - [5]). In part, this can be explained by biological factors, such as dietary
338 breadth, varying among taxa, with the most diverse diet found in *Californiconus* [5]. Despite our
339 efforts, we believe that the venom diversity reported for *Conasprella* and *Pygmaeonus* is, to
340 some extent, an underestimate, resulting from the limited data available to us. Analyses based
341 on a single transcriptome typically report fewer toxins, and discrepancies may sometimes be
342 striking. For example, 53 toxins were identified from *C. (Pionoconus) consors* [53] as opposed to
343 a total of 232 toxins in three separately sequenced specimens of the closely related *C.*
344 (*Pionoconus*) *magus* [37]. Furthermore, in our de novo CDSs annotation, we prioritized
345 reliability of the toxin identification, and so we used rather stringent filtering criteria. The N-
346 terminally incomplete CDSs were ignored, as well as those with low probability of signal region
347 prediction (D-value < 0.7), low expression, less than three predicted transcripts, or exclusive to
348 one taxon. Nevertheless, in both *Conasprella* and *Pygmaeonus*, a quarter to almost one third
349 of the known gene superfamily members could only be identified from the annotation of
350 predicted CDSs, but not from search against reference databases. Most likely, this is partly a
351 result of the high PID value we used to limit the output of the initial BLASTx step, and by
352 relaxing it, we might have been able to identify more toxins at the first step of annotation.
353 However, when we tried relaxing the PID or BLASTx e-values, it led to huge outputs with
354 increasing proportions of false positives, and their manual curation was not feasible. An array of

355 algorithms, known as machine learning and recently developed into the automated pipeline
356 *ConusPipe* [21], offers yet another way to optimize toxin identification. This tool showed
357 excellent performance when trained on *Conus* datasets and applied to the identification of
358 *Conus* toxins. Nevertheless, we found it methodologically incorrect to use a training set of
359 sequences derived from *Conus*, and then apply it to datasets of notably divergent taxa. Still,
360 most of the parameters used by *ConusPipe*, were either set explicitly, or checked at the stage of
361 SSC screening. Despite this semi-manual procedure allowing us to improve recovery of toxin
362 sequences, additional transcriptomic and proteomic data, including on other species of
363 divergent Conidae, will be important to corroborate our findings.

364 Of particular interest is the gene superfamily DivCon3 identified by the annotation of the SSCs,
365 a likely innovation of the *Californiconus-Pygmaeconus* subclade of Conidae. Due to the cleavage
366 sites within the mature region (monobasic in Pyg6 and dibasic in Pyg9 and Pyg11), we
367 hypothesize that the final peptide products of these three precursors are 13-14 AA-long
368 oligopeptides bearing a RF-amide motif C-terminally. Because of both the presence of a C-
369 terminal RF-amide and the very small size, the predicted cleavage products of the ConDiv3 are
370 similar to conorfamides [20,54,55]. A pronounced physiological effect was demonstrated for
371 the conorfamide CNF-Vc1 from *C. victoriae*. In mice it elicits increase of intracellular calcium
372 levels in the dorsal root ganglia and causes nearly complete muscle paralysis [20]. A similar
373 pharmacology may characterize DivCon3 members, and the high expression of these transcripts
374 in the venom gland of *Pygmaeconus* implies their functional significance. Further functional
375 studies on these oligopeptides are necessary to identify their molecular targets. In contrast, the
376 signal region of ConDiv3 does not show any similarity to that of conorfamides, and mature
377 peptide regions of ConDiv3 bear two conservatively arranged Cys residues, whereas known
378 conorfamides are cysteine-free. This suggests that DivCon3 and conorfamides of *Conus* are
379 likely convergently-evolved venom components and constitute yet another remarkable
380 parallelism in the molecular evolution of toxins in Conidae.

381 Despite the fact that research on the chemical structure and pharmacological properties of
382 conopeptides commenced over four decades ago [56], the complexity of *Conus* venoms may
383 still be greatly underestimated. Recent studies have demonstrated that defense-invoked
384 venoms may differ in composition from predation-invoked venoms [3]. Likewise, some *Conus*
385 feeding strategies involve the release of a subset of venom components directly into the water
386 to alter the behavior of their prey prior to injection of a killing shot of venom. The

387 physiologically active components in this subset may be as exotic as specialized insulins, or
388 small molecules mimicking the natural pheromones of the prey [7,57]. This suggests that there
389 may be a great diversity of venom components, or specific enzymes involved in the biosynthesis
390 of these components that are still not identified. The reason for this is largely methodological -
391 most venom analyses utilize similarity-based searches, as they primarily target canonical
392 conotoxin gene superfamilies, and (at best) peptides of similar structural properties.

393 Our phylogeny-aware approach on a subset of *Conus* species, in which specific divergent and/or
394 taxonomically restricted venom components are sought out in different lineages of cone snails,
395 revealed a previously uncharacterized diversity of putative toxins even in what might seem to
396 be well-annotated transcriptomes. Remarkably, New-Geo-1 and Pmag02 appear to be
397 ubiquitous in *Conus* as well as in the divergent Conidae genera. This case shows how an
398 inaccurate picture of venom components distribution across the Conidae evolutionary tree can
399 bias research hypotheses related to toxin evolution. The New-Geo-1 and Pmag02 superfamilies
400 were previously known to be highly expressed in the fish-hunting subgenera *Gastridium* and
401 *Pionoconus* respectively, and so might be interpreted as specific adaptations to piscivory. If this
402 were correct, the very high expression levels of New-Geo-1 and Pmag02 in non-piscivorous
403 *Profundiconus* and *Pygmaeonus*, respectively, could suggest that these components are a part
404 of the defensive venom targeting fish predators. However, recently, transcripts referable to
405 Pmag02 were also identified in the vermivorous *Conus* lineages from West Africa [6]. Finally, as
406 a phylogenetically more representative picture of New-Geo-1 and Pmag02 distribution in
407 Conidae emerges with our study, any support evades for the hypothesis that these gene
408 superfamilies are at all related to piscivory.

409 On the one hand, we emphasize a need for thorough and accurate annotation of transcriptomic
410 data, even if it requires laborious tasks that cannot be fully automated. On the other hand, we
411 must admit that further studies that are solely based on -OMICs data are deemed to remain
412 somewhat incremental, because they are unable to produce functional data. Lacking such data,
413 the roles of various venom components remain unclear, and with it the benefits that a
414 particular taxon acquires by evolving them. Major breakthroughs in understanding drivers of
415 Conidae venom evolution should thus be guided by a knowledge of feeding ecology of different
416 species of Conidae and require functional assays alongside venom profiling. A major challenge
417 along the way is the further elaboration of existing methodologies to overcome common

418 shortages of research samples and, increasingly, improving behavior documentation practices
419 to eventually analyse molecular data within an ecological context.

420

421 **Data accessibility.** The transcriptomic sequencing data are deposited in the NCBI SRA database,
422 under the Bioproject PRJNA735765. Sequences of the predicted toxins are provided in
423 supplementary data (Table S2, Figures S1 – S8), Python scripts are available at
424 https://github.com/Hyperdiverseproject/Divergent_Conidae.

425

426 **Acknowledgements**

427 The material was collected during the KANACONO expedition in New Caledonia (convention
428 MNHN-Provence Sud, APA_NCPS_2016_012; PI N. Puillandre and S. Samadi) and the KAVIENG
429 2014 expedition in Papua-New-Guinea (endorsed by the New Ireland Provincial Administration
430 and operated under a Memorandum of Understanding with the University of Papua New
431 Guinea; PI P. Bouchet and J. Kinch), as part of the Our Planet Reviewed and the Tropical Deep-
432 Sea Benthos programs, organized jointly by the *Muséum national d'Histoire naturelle* (MNHN),
433 Pro-Natura International (PNI) and the *Institut de Recherche pour le Développement* (IRD). The
434 organizers acknowledge funding from the Total Foundation, the *Laboratoire d'Excellence*
435 *Diversités Biologiques et Culturelles* (LabEx BCDiv, ANR-10-LABX-0003-BCDiv), the *Programme*
436 *Investissement d'Avenir* (ANR-11-IDEX-0004-02), the *Fonds Pacifique*, CNRS' *Institut Ecologie et*
437 *Environnement* (INEE) and the project CONOTAX, funded by the French *Agence Nationale de la*
438 *Recherche – France* (ANR-13-JSV7-0013-01). These expeditions operated under the regulations
439 then in force in the countries in question and satisfy the conditions set by the Nagoya Protocol
440 for access to genetic resources. We thank Laetitia Aznar-Cormano for her help with the RNA
441 extraction of the *Pygmaeconus* sample. We are thankful to Giulia Fassio (University of Roma)
442 for providing intermediate data files on *Profundiconus*, and to Helena Safavi-Hemami
443 (University of Utah) for her comments on the novel gene superfamilies. We thankful to Claudia
444 Ratti for her comments on the manuscript. The present study was supported by the Russian
445 Science Foundation, grant N 19-74-10020 to AF, and by the European Research Council (ERC)
446 under the European Union's Horizon 2020 research and innovation program (grant agreement
447 No.865101) to NP.

448 **References**

- 449 1. Olivera BM, Showers Corneli P, Watkins M, Fedosov A. 2014 Biodiversity of Cone Snails and
450 Other Venomous Marine Gastropods: Evolutionary Success Through Neuropharmacology. *Annual*
451 *reviews Animal Biosciences* **2**, 487–513.
- 452 2. Puillandre P, Duda TF, Meyer C, Olivera BM, Bouchet P. 2014 One, four or 100 genera? A new
453 classification of the cone snails. *Journal of Molluscan Studies* **81**, 1–23.
- 454 3. Dutertre S *et al.* 2014 Evolution of separate predation- and defence-evoked venoms in
455 carnivorous cone snails. *Nature communications* **3521**, 1–9.
- 456 4. Olivera BM, Seger J, Horvath MP, Fedosov AE. 2015 Prey-capture Strategies of Fish-hunting Cone
457 Snails: Behavior, Neurobiology and Evolution. *Brain, Behavior and Evolution* **86**, 58–74.
- 458 5. Phuong MA, Mahardika GN, Alfaro ME. 2016 Dietary breadth is positively correlated with venom
459 complexity in cone snails. *BMC Genomics* **17**, 1–15.
- 460 6. Abalde S, Tenorio MJ, Afonso CML, Zardoya R. 2020 Comparative transcriptomics of the venoms
461 of continental and insular radiations of West African cones. *Proc. R. Soc. B.* **287**, 20200794.
462 (doi:10.1098/rspb.2020.0794)
- 463 7. Torres JP *et al.* 2021 Small-molecule mimicry hunting strategy in the imperial cone snail, *Conus*
464 *imperialis*. *Sci. Adv.* **7**, eabf2704. (doi:10.1126/sciadv.abf2704)
- 465 8. Prashanth JR, Brust A, Alewood PF, Dutertre S, Lewis RJ. 2014 Cone snail venomics: from novel
466 biology to novel therapeutics. *Future Med Chem.* **6**, 1659–75.
- 467 9. Safavi-Hemami H, Brogan SE, Olivera BM. 2019 Pain therapeutics from cone snail venoms: From
468 Ziconotide to novel non-opioid pathways. *Journal of Proteomics* **190**, 12–20.
- 469 10. Casewell NR, Wüster W, Vonk FJ, Harrison RA, Fry BG. 2013 Complex cocktails: the evolutionary
470 novelty of venoms. *Trends in Ecology & Evolution* **28**, 219–229. (doi:10.1016/j.tree.2012.10.020)
- 471 11. Zancolli G *et al.* 2019 When one phenotype is not enough: divergent evolutionary trajectories
472 govern venom variation in a widespread rattlesnake species. *Proceedings of the Royal Society B:*
473 *Biological Sciences* **286**, 20182735. (doi:10.1098/rspb.2018.2735)
- 474 12. MolluscaBase eds. In press. Conidae J. Fleming, 1822. *MolluscaBase*.
- 475 13. Uribe JE, Puillandre N, Zardoya R. 2017 Beyond Conus: Phylogenetic relationships of Conidae
476 based on complete mitochondrial genomes. *Molecular Phylogenetics and Evolution* **107**, 142–151.
- 477 14. Phuong MA, Alfaro ME, Mahardika GN, Marwoto RM, Prabowo RE, von Rintelen T, Vogt PWH,
478 Hendricks JR, Puillandre N. 2019 Lack of Signal for the Impact of Conotoxin Gene Diversity on Speciation
479 Rates in Cone Snails. *Systematic Biology* **68**, 781–796.
- 480 15. Puillandre N, Fedosov AE, Kantor YI. 2016 Systematics and Evolution of the Conoidea. In
481 *Evolution of Venomous Animals and Their Toxins* (ed P Gopalakrishnakone), pp. 1–32. Springer.
- 482 16. Biggs JS *et al.* 2010 Evolution of *Conus* peptide toxins: Analysis of *Conus californicus* Reeve,
483 1844. *Molecular Phylogenetics and Evolution* **56**, 1–12. (doi:10.1016/j.ympev.2010.03.029)
- 484 17. Aguilar MB, López-Vera E, Ortiz E, Becerril B, Possani LD, Olivera BM, Heimer de la Cotera EP.
485 2005 A Novel Conotoxin from *Conus delessertii* with Posttranslationally Modified Lysine Residues [†].
486 *Biochemistry* **44**, 11130–11136. (doi:10.1021/bi050518l)

- 487 18. Figueroa-Montiel A, Bernáldez J, Jiménez S, Ueberhide B, González L, Licea-Navarro A. 2018
488 Antimycobacterial Activity: A New Pharmacological Target for Conotoxins Found in the First Reported
489 Conotoxin from *Conasprella ximenes*. *Toxins* **10**, 51. (doi:10.3390/toxins10020051)
- 490 19. Fassio G, Modica MV, Mary L, Zaharias P, Fedosov AE, Gorson J, Kantor YI, Holford M, Puillandre
491 N. 2019 Venom Diversity and Evolution in the Most Divergent Cone Snail Genus *Profundiconus*. *Toxins*
492 **11**, 623. (doi:10.3390/toxins11110623)
- 493 20. Robinson SD, Safavi-Hemami H, Raghuraman S, Imperial JS, Papenfuss AT, Teichert RW, Purcell
494 AW, Olivera BM, Norton RS. 2015 Discovery by proteogenomics and characterization of an RF-amide
495 neuropeptide from cone snail venom. *Journal of Proteomics* **114**, 38–47.
496 (doi:10.1016/j.jprot.2014.11.003)
- 497 21. Li Q, Watkins M, Robinson SD, Safavi-Hemami H, Yandell M. 2018 Discovery of Novel Conotoxin
498 Candidates Using Machine Learning. *Toxins* **10**, 503. (doi:10.3390/toxins10120503)
- 499 22. Puillandre N, Watkins M, Olivera BM. 2010 Evolution of *Conus* peptide genes: duplication and
500 positive selection in the A-superfamily. *Journal of Molecular Evolution* **70**, 190–202.
- 501 23. Abalde S, Tenorio MJ, Afonso CML, Zardoya R. 2018 Conotoxin Diversity in *Chelyconus ermineus*
502 (Born, 1778) and the Convergent Origin of Piscivory in the Atlantic and Indo-Pacific Cones. *Genome*
503 *Biology and Evolution* **10**, 2643–2662.
- 504 24. Olivera BM, Teichert RW. 2007 Diversity of the neurotoxic *Conus* peptides: a model for
505 concerted pharmacological discovery. *Molecular Interventions* **7**, 253–262.
- 506 25. Puillandre N, Holford M. 2010 The Terebridae and teretoxins: Combining phylogeny and
507 anatomy for concerted discovery of bioactive compounds. *BMC Chemical Biology* **10**, 1–12.
- 508 26. Bolger AM, Lohse M, Usadel B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence
509 data. *Bioinformatics* **30**, 2114–2120. (doi:10.1093/bioinformatics/btu170)
- 510 27. Grabherr MG *et al.* 2011 Full-length transcriptome assembly from RNA-Seq data without a
511 reference genome. *Nat Biotechnol* **29**, 644–652. (doi:10.1038/nbt.1883)
- 512 28. Robinson SD, Safavi-Hemami H, McIntosh LD, Purcell AW, Norton RS, Papenfuss AT. 2014
513 Diversity of Conotoxin Gene Superfamilies in the Venomous Snail, *Conus victoriae*. *PLoS ONE* **9**, e87648.
514 (doi:10.1371/journal.pone.0087648)
- 515 29. Li Q *et al.* 2017 Divergence of the venom exogene repertoire in two sister species of *Turriconus*.
516 *Genome Biology and Evolution* **9**, 2211–2225.
- 517 30. Li B, Dewey CN. 2011 RSEM: accurate transcript quantification from RNA-Seq data with or
518 without a reference genome. , 16.
- 519 31. Langmead B, Salzberg SL. 2012 Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–
520 359. (doi:10.1038/nmeth.1923)
- 521 32. Kaas Q, Westermann J-C, Halai R, Wang CKL, Craik DJ. 2008 ConoServer, a database for
522 conopeptide sequences and structures. *Bioinformatics* **24**, 445–446.
523 (doi:10.1093/bioinformatics/btm596)
- 524 33. Bose U, Wang T, Zhao M, Motti CA, Hall MR, Cummins SF. 2017 Multiomics analysis of the giant
525 triton snail salivary gland, a crown-of-thorns starfish predator. *Scientific reports* **7**, 1–14.
- 526 34. Modica MV, Lombardo F, Franchini P, Oliverio M. 2015 The venomous cocktail of the vampire
527 snail *Colubraria reticulata* (Mollusca, Gastropoda). *BMC Genomics* **16**: 441, 1–21.

- 528 35. Li R, Bekaert M, Wu L, Mu C, Song W, Migaud H, Wang C. 2019 Transcriptomic Analysis of
529 Marine Gastropod *Hemifusus tuba* Provides Novel Insights into Conotoxin Genes. *Marine Drugs* **17**, 466.
530 (doi:10.3390/md17080466)
- 531 36. Lu A *et al.* 2020 Transcriptomic Profiling Reveals Extraordinary Diversity of Venom Peptides in
532 Unexplored Predatory Gastropods of the Genus *Clavus*. *Genome Biology and Evolution* **12**, 684–700.
533 (doi:10.1093/gbe/evaa083)
- 534 37. Pardos-Blas, Irisarri, Abalde, Tenorio, Zardoya. 2019 Conotoxin Diversity in the Venom Gland
535 Transcriptome of the Magician’s Cone, *Pionoconus magus*. *Marine Drugs* **17**, 553.
536 (doi:10.3390/md17100553)
- 537 38. Barghi N, Concepcion GP, Olivera BM, Lluisma AO. 2015 Comparison of the Venom Peptides and
538 Their Expression in Closely Related *Conus* Species: Insights into Adaptive Post-speciation Evolution of
539 *Conus* Exogenomes. *Genome Biology and Evolution* **7**, 1797–1814.
- 540 39. Peng C *et al.* 2016 High-throughput identification of novel conotoxins from the Chinese tubular
541 cone snail (*Conus betulinus*) by multi-transcriptome sequencing. *GigaSci* **5**, 17. (doi:10.1186/s13742-
542 016-0122-9)
- 543 40. Li X, Chen W, Zhangsun D, Luo S. 2020 Diversity of Conopeptides and Their Precursor Genes of
544 *Conus litteratus*. *Marine Drugs* **18**, 464. (doi:10.3390/md18090464)
- 545 41. Wheeler DL. 2003 Database resources of the National Center for Biotechnology. *Nucleic Acids*
546 *Research* **31**, 28–33. (doi:10.1093/nar/gkg033)
- 547 42. Nielsen H. 2017 Predicting Secretory Proteins with SignalP. In *Protein Function Prediction:*
548 *Methods and Protocols* (ed D Kihara), pp. 59–73. New York, NY: Springer. (doi:10.1007/978-1-4939-
549 7015-5_6)
- 550 43. Käll L, Krogh A, Sonnhammer ELL. 2007 Advantages of combined transmembrane topology and
551 signal peptide prediction—the Phobius web server. *Nucleic Acids Res* **35**, W429–W432.
552 (doi:10.1093/nar/gkm256)
- 553 44. Finn RD, Clements J, Eddy SR. 2011 HMMER web server: interactive sequence similarity
554 searching. *Nucleic Acids Research* **39**, W29–W37. (doi:10.1093/nar/gkr367)
- 555 45. Mistry J *et al.* 2021 Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**,
556 D412–D419. (doi:10.1093/nar/gkaa913)
- 557 46. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012 CD-HIT: accelerated for clustering the next-generation
558 sequencing data. *Bioinformatics* **28**, 3150–3152. (doi:10.1093/bioinformatics/bts565)
- 559 47. Kaas Q, Yu R, Jin A-H, Dutertre S, Craik DJ. 2012 ConoServer: updated content, knowledge, and
560 discovery tools in the conopeptide database. *Nucleic Acids Research* **40**, D325–D330.
561 (doi:10.1093/nar/gkr886)
- 562 48. Katoh K, Standley DM. 2013 MAFFT Multiple Sequence Alignment Software Version 7:
563 Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780.
564 (doi:10.1093/molbev/mst010)
- 565 49. Bairoch A, Apweiler R. 1996 The SWISS-PROT Protein Sequence Data Bank and Its New
566 Supplement TREMBL. *Nucleic Acids Research* **24**, 21–25. (doi:10.1093/nar/24.1.21)
- 567 50. Hammer Ø, Harper DAT, Ryan PD. 2001 PAST: Paleontological statistics software package for
568 education and data analyses. *Paleontol. Electron.* **4**, 1–9.

- 569 51. Dutt M, Dutertre S, Jin A-H, Lavergne V, Alewood P, Lewis R. 2019 Venomics Reveals Venom
570 Complexity of the Piscivorous Cone Snail, *Conus tulipa*. *Marine Drugs* **17**, 71. (doi:10.3390/md17010071)
- 571 52. Jin A-H, Dutertre S, Dutt M, Lavergne V, Jones A, Lewis RJ, Alewood PF. 2019 Transcriptomic-
572 Proteomic Correlation in the Predation-Evoked Venom of the Cone Snail, *Conus imperialis*. *Mar Drugs*
573 **17**. (doi:10.3390/md17030177)
- 574 53. Terrat Y, Biass D, Dutertre S, Favreau P, Remm M, Stöcklin R, Piquemal D, Ducancel F. 2012 High-
575 resolution picture of a venom gland transcriptome: case study with the marine snail *Conus consors*.
576 *Toxicon* **59**, 34–46. (doi:10.1016/j.toxicon.2011.10.001)
- 577 54. Aguilar MB, Luna-Ramírez KS, Echeverría D, Falcón A, Olivera BM, Heimer de la Cotera EP, Maillo
578 M. 2008 Conorfamide-Sr2, a gamma-carboxyglutamate-containing FMRFamide-related peptide from the
579 venom of *Conus spurius* with activity in mice and mollusks. *Peptides* **29**, 186–195.
580 (doi:10.1016/j.peptides.2007.09.022)
- 581 55. Lebbe EKM, Tytgat J. 2016 In the picture: disulfide-poor conopeptides, a class of
582 pharmacologically interesting compounds. *J Venom Anim Toxins Incl Trop Dis* **22**, 30.
583 (doi:10.1186/s40409-016-0083-6)
- 584 56. Edean R, Parish G, Gyr P. 1974 Pharmacology of the venom of *Conus geographus*. *Toxicon* **12**,
585 131–138. (doi:10.1016/0041-0101(74)90236-0)
- 586 57. Safavi-Hemami H *et al.* 2015 Specialized insulin is used for chemical warfare by fish-hunting cone
587 snails. *PNAS* **112**, 1743–1748.
- 588 58. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013 MEGA6: Molecular Evolutionary
589 Genetics Analysis Version 6.0. *Molecular Biology and Evolution* , 2725–2729.
- 590

591 **Figure captions**

592 Figure 1. A. phylogenetic tree of the family Conidae (after Phuong et al. 2019 - [14]). B. Venn
593 diagram showing numbers of SSCs shared by divergent Conidae genera. Shells of sequenced
594 specimens (*Profundiconus*, *Pygmaeonus*), or conspecific to sequenced ones (*Conasprella*,
595 *Californiconus*), shown above (not to scale). In the Venn diagram intersections: regular font –
596 total cluster count; in bold – number of clusters with at least one transcript of moderately high
597 or high expression (tpm>100).

598 Figure 2. A. Counts of the identified transcripts by gene superfamily for *Conasprella coriolisi*
599 (left) and *Pygmaeonus traillii* (right). B. log10 transformed relative expression levels of the
600 conotoxin gene superfamilies in *Conasprella coriolisi* (left) and *Pygmaeonus traillii* (right).

601 Figure 3. Alignment of the DivConN3 gene superfamily. The Neighbor-Joining phylogenetic tree
602 (obtained with MEGA v6 [58]) on the left based on the complete precursor sequences. Signal
603 sequences shown in blue and pro-region underlined green; bold letters correspond to Cys-
604 residues and predicted cleavage sites within mature peptide region; the orange box marks C-
605 terminal RF-amide motif.

606 Figure 4. A. Heat map of the gene superfamily expression in the four species of the divergent
607 Conidae and 15 species of *Conus*. Pmag02 and New-Geo-1 are highlighted in red and green,
608 respectively. B. PCA diagram of the divergent Conidae (stars) and 15 species of *Conus* (dots)
609 based on gene superfamily expression data (Table S4). Principal components 1 and 2 account
610 for a total of 42.27% of the observed variation. The position of *Pygmaeonus* at the extreme of
611 PC1 is mainly due to the contribution of L-, and next N- / Divergent-MSTLGMTLL- gene
612 superfamilies. The placement of *Conasprella* is largely explained by the contributions of con-
613 ikot-ikot- and M- gene superfamilies.

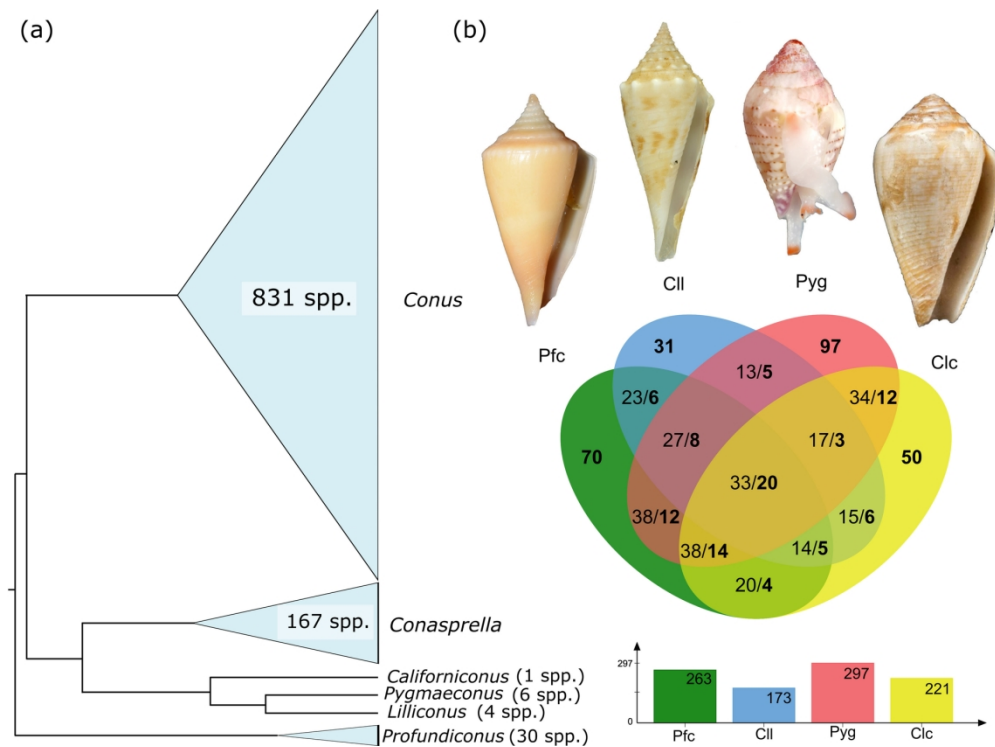


Figure 1. A. phylogenetic tree of the family Conidae (after Phuong et al. 2019 - [14]). B. Venn diagram showing numbers of SSCs shared by divergent Conidae genera. Shells of sequenced specimens (*Profundiconus*, *Pygmaeonus*), or conspecific to sequenced ones (*Conasprella*, *Californiconus*), shown above (not to scale). In the Venn diagram intersections: regular font – total cluster count; in bold – number of clusters with at least one transcript of moderately high or high expression ($tpm > 100$).

169x125mm (300 x 300 DPI)

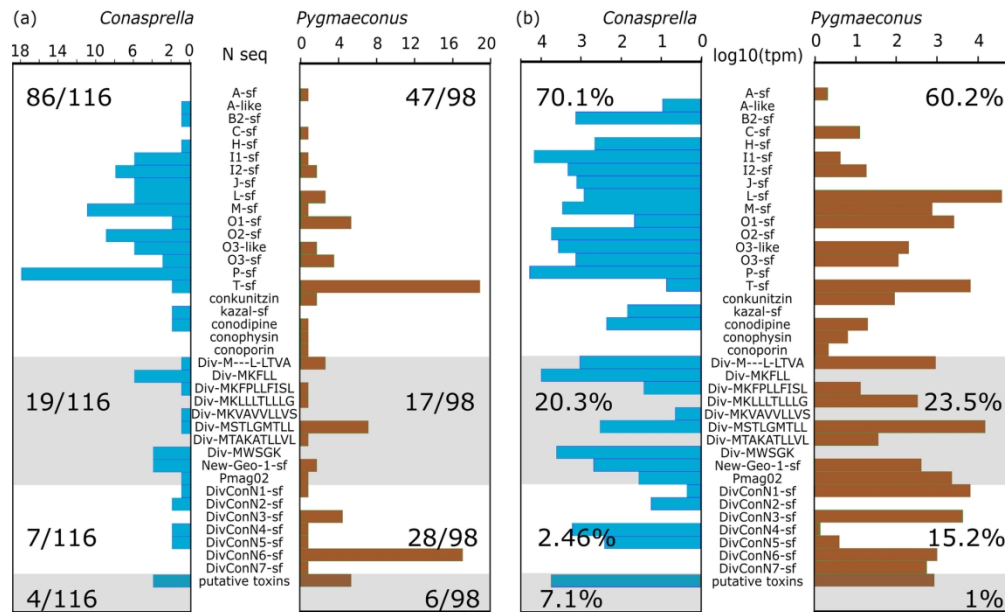


Figure 2. A. Counts of the identified transcripts by gene superfamily for *Conasprella coriolisi* (left) and *Pygmaeonus traillii* (right). B. log₁₀ transformed relative expression levels of the conotoxin gene superfamilies in *Conasprella coriolisi* (left) and *Pygmaeonus traillii* (right).

169x104mm (300 x 300 DPI)

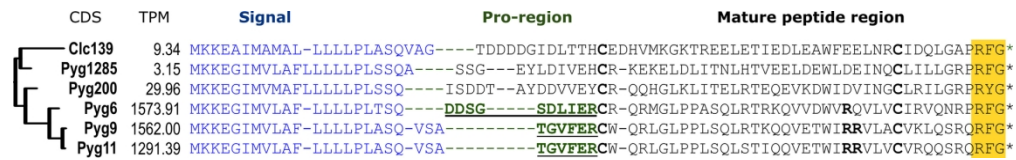


Figure 3. Alignment of the DivConN3 gene superfamily. The Neighbor-Joining phylogenetic tree (obtained with MEGA v6 [58]) on the left based on the complete precursor sequences. Signal sequences shown in blue and pro-region underlined green; bold letters correspond to Cys-residues and predicted cleavage sites within mature peptide region; the orange box marks C-terminal RF-amide motif.

169x24mm (300 x 300 DPI)

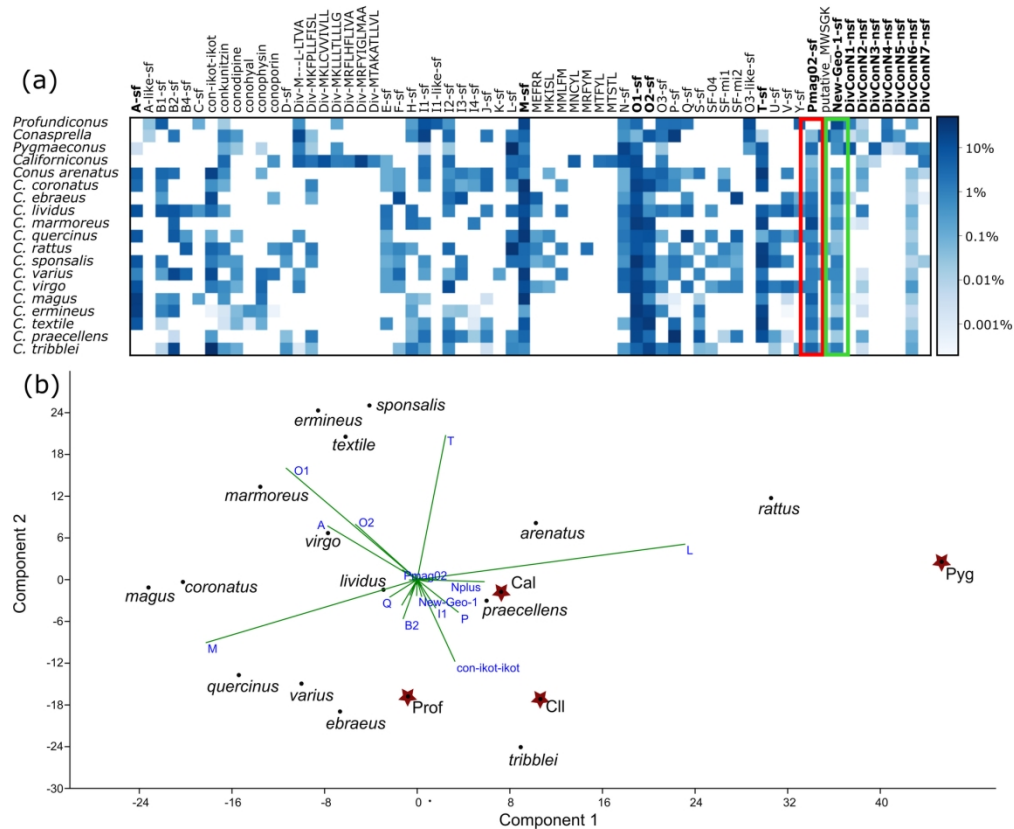


Figure 4. A. Heat map of the gene superfamily expression in the four species of the divergent Conidae and 15 species of *Conus*. Pmag02 and New-Geo-1 are highlighted in red and green, respectively. B. PCA diagram of the divergent Conidae (stars) and 15 species of *Conus* (dots) based on gene superfamily expression data (Table S4). Principal components 1 and 2 account for a total of 42.27% of the observed variation. The position of *Pygmaeonus* at the extreme of PC1 is mainly due to the contribution of L-, and next N- / Divergent-MSTLGMTLL- gene superfamilies. The placement of *Conasprella* is largely explained by the contributions of con-ikot-ikot- and M- gene superfamilies.

169x140mm (300 x 300 DPI)