


# Phylogenetic analyses suggest centipede venom arsenals were repeatedly stocked by horizontal gene transfer

Eivind A. B. Undheim<sup>1,2,3</sup>✉ & Ronald A. Jenner<sup>4</sup> ✉

Venoms have evolved over a hundred times in animals. Venom toxins are thought to evolve mostly by recruitment of endogenous proteins with physiological functions. Here we report phylogenetic analyses of venom proteome-annotated venom gland transcriptome data, assisted by genomic analyses, to show that centipede venoms have recruited at least five gene families from bacterial and fungal donors, involving at least eight horizontal gene transfer events. These results establish centipedes as currently the only known animals with venoms used in predation and defence that contain multiple gene families derived from horizontal gene transfer. The results also provide the first evidence for the implication of horizontal gene transfer in the evolutionary origin of venom in an animal lineage. Three of the bacterial gene families encode virulence factors, suggesting that horizontal gene transfer can provide a fast track channel for the evolution of novelty by the exaptation of bacterial weapons into animal venoms.

<sup>1</sup>Centre for Biodiversity Dynamics, Department of Biology, NTNU, Trondheim, Norway. <sup>2</sup>Centre for Ecological and Evolutionary Synthesis, Department of Bioscience, University of Oslo, Blindern, Oslo, Norway. <sup>3</sup>Centre for Advanced Imaging, University of Queensland, St Lucia, QLD, Australia. <sup>4</sup>Department of Life Sciences, Natural History Museum, London, UK. ✉email: [e.a.undheim@ibv.uio.no](mailto:e.a.undheim@ibv.uio.no); [r.jenner@nhm.ac.uk](mailto:r.jenner@nhm.ac.uk)

**H**orizontal gene transfer (HGT) between kingdoms and domains of life has contributed to the evolution of a diversity of novel adaptive traits in animals, including the ability of bdelloid rotifers to withstand desiccation, the ability of springtails to feed on decaying organic matter, and the ability of plant-parasitic nematodes to degrade plant cell walls<sup>1–7</sup>. HGT has also contributed to the evolution of venom, one of the most convergently evolved animal adaptations. Venoms are complex, typically proteinaceous, secretions that are used primarily for predation and defence by a wide phylogenetic range of animals. However, although animal venoms have evolved at least a hundred times independently<sup>8</sup>, the contribution of HGT to the evolution of venom arsenals has so far been shown to be minor.

HGT is a well-supported hypothesis for only three gene families present in arthropod and cnidarian venoms. Phylogenetic analyses, in some cases supported by genomic information, strongly suggest that bacteria were the source of type D phospholipases found in the venoms of sicariid spiders, scorpions, and ticks<sup>9</sup>, and of pore-forming toxins expressed in the venom glands of ticks as well as gland cells in the digestive system of cnidarians, although it is debated whether these should be considered part of the venom system or not<sup>10</sup>. Similarly, glycoside family 19 chitinases found in the venom of chalcidoid parasitoid wasps were probably transferred from parasitic fungi<sup>11</sup>. Other potential cases of HGT contributing to insect venoms currently lack phylogenetic support<sup>12–14</sup>, while the direction of HGT of neurotoxic  $\alpha$ -latrotoxins present in the venom of theridiid spiders and bacteria remains uncertain<sup>15</sup>. Although HGT is currently not considered to be a major mechanism of venom evolution, venoms are nevertheless a promising research area given the existence of many tens of thousands of mostly unstudied venomous animal species. Many venoms also contain a substantial number of proteins with few or no known metazoan homologues<sup>16–21</sup>, and these may include HGT candidates.

One venomous lineage that contains a large diversity of unassignable venom proteins<sup>22,23</sup> is centipedes (Chilopoda). Centipedes are one of the oldest terrestrial venomous lineages, with a fossil record going back 418 million years<sup>24</sup>. Living species belong to five orders: Scutigermorpha (long-legged house centipedes), Lithobiomorpha (stone centipedes), Geophilomorpha (long-bodied earth centipedes), Scolopendromorpha (the most familiar centipedes, including large tropical species), and Craterostigmomorpha (two species from Tasmania and New Zealand). All of these have complex venoms that are used for predation and defence. While most of the protein families contained in centipede venoms were recruited from gene families that are widespread in animals, others have few or no metazoan homologues. This pattern suggests that the evolutionary origins of several centipede venom toxins could lie outside the animal kingdom.

We show that multiple HGTs have stocked centipede venom arsenals throughout their evolution. Phylogenetic analyses of venom gland transcriptome and venom proteome data assisted by genomic analyses identified seven gene families encoding centipede venom proteins and peptides that were horizontally transferred between bacteria, fungi, oomycetes, and centipedes. Our analyses reveal between 10 and 12 HGT events. At least eight HGTs involved five gene families that transferred from bacteria and fungi into centipede venoms, whereas the direction of two or three HGTs between centipedes and fungi and oomycetes remain uncertain. Three of the protein families in bacterial donor taxa are virulence factors involved in pathogenicity, suggesting that centipedes have repurposed bacterial weapons as venom components involved in predation and/or defence. Our findings suggest that HGT can be an important factor shaping the evolution of animal venoms.

## Results and discussion

**Overall support for HGT.** Several methods are available for identifying HGT<sup>25</sup>. A combination of phylogenetic analyses of candidate HGT gene families including both potential donor and host sequences, and confirming their presence in host genomes is considered to be the most robust method. We used this approach to identify putative HGTs from non-metazoan sources into centipede venoms. Table 1 summarizes the support for all inferred HGTs that have contributed to centipede venom arsenals. The robustly supported phylogenetic nesting of clades of centipede sequences within paraphyletic backbones of non-metazoan donor sequences supports HGT for five of the seven gene families:  $\beta$ -pore-forming toxin ( $\beta$ -PFTx), centipede peptidylarginine deiminase (centiPAD), protein with a domain of unknown function (DUF3472), pesticidal crystal protein domain-containing protein-like protein (PCPDP-like), and uncharacterized protein family 5 (unchar05). The phylogenetic nesting of centipede geotoxin 2 (GEOTX02) within fungal sequences is less well supported, while the centipede sequences for uncharacterized protein family 16 (unchar16) group in a clade that is sister to a clade of oomycete sequences. Furthermore, by confirming that five of the genes map to protein-coding genes with introns in the genome of the geophilomorph centipede *Strigamia maritima*, which is the only published centipede genome<sup>26</sup>, we show that they are bona fide centipede genes rather than the result of contamination or symbionts. Importantly, a recent study<sup>27</sup> that examined the presence of contamination in the genome of *S. maritima* confirms that none of our HGT candidates map to the only genomic scaffold for which there are signs of contamination (scaffold JH431684; C. M. Francois, pers. comm.).

We bolster our conclusions about HGT with three ancillary criteria. First, all seven putative HGT gene families are present in both centipede venom gland transcriptomes and milked venom proteomes, which argues against them being accidental contamination. Second, each putative HGT gene is consistently expressed in the venom glands of multiple species collected from disparate geographic locations and habitats, which would not be expected if the sequences derived from local contaminants. Third, putative HGT sequences from different centipede species that are contaminants would be expected to group with related non-centipede sequences in different places in gene trees, rather than cluster together in a single clade. The strong clustering of the centipede sequences into well-supported clades in our gene trees, and the lack of the haphazard interleaving of putative donor and centipede sequences in any of our trees strongly suggest that the putative HGT genes are bona fide centipede sequences. Fulfilment of these ancillary criteria in addition to the phylogenetic nesting of the centipede sequences within paraphyletic groups of donor sequences, and the presence of five of the seven genes in the genome of *S. maritima*, further decreases the probability that our results are due to contamination or symbionts. Below we will discuss the full support for our conclusions for each of the genes, and the possibility that the genes that could not be checked against the *S. maritima* genome (centiPAD and PCPDP-like) could be due to symbionts.

**Bacterial pore-forming toxins transferred twice into centipedes.** Centipede  $\beta$ -PFTxs were recruited into the ancestral centipede venom proteome, with subsequent losses from craterostigmomorph and geophilomorph venoms<sup>23</sup>. This gene family belongs to the bacterial aerolysin-like  $\beta$ -pore-forming toxin superfamily, which Moran et al.<sup>10</sup> showed was transferred at least six times from bacteria to eukaryotes, including animals. We did not specifically design our phylogenetic dataset to provide a precise estimate of when and where all non-centipede HGTs occurred, but our findings agree with and extend their results.

**Table 1 Summary of gene families horizontally transferred into centipede venoms.**

Gene	HGT source	Number of HGT events <sup>a</sup>	Phylogenetic location of HGT	Phylogenetic location of recruitment into venom	Mapped to <i>Strigamia maritima</i> genome <sup>b</sup>
$\beta$ -PFTx	Bacteria	2 (1)	Arthropoda or Chilopoda; within Lithobiomorpha	Chilopoda	SMAR004242, SMAR004243, SMAR012417
centiPAD	Bacteria	2	Within Scutigermorpha; within Lithobiomorpha	Within Scutigermorpha; within Lithobiomorpha	n/a
DUF3472	Bacteria	1 or 2 (1)	In the stem of Pleurostigmophora or Amalpighiata <sup>c</sup> ; or in Epimorpha and within Lithobiomorpha	Within Scolopendromorpha	SMAR002991, SMAR002992, SMAR002993, SMAR008653
GEOTX02	Fungi <sup>d</sup>	1 or 2	Geophilomorpha	Geophilomorpha	(group 1: SMAR012843, SMAR003678, SMAR004759); (group 2: SMAR012429, SMAR005429); (group 3: SMAR014279; (group 4: SMAR009615, SMAR004692, SMAR001285, SMAR007268, SMAR006394, SMAR009617, SMAR010233)
PCPDP-like	Bacteria	1	Lithobiomorpha	Lithobiomorpha	n/a
unchar05	Fungi	2	Geophilomorpha, within Lithobiomorpha	Geophilomorpha	SMAR002275, SMAR004333, SMAR005016, SMAR002277, SMAR015613
unchar16	Oomycetes <sup>d</sup>	1	Unknown	Craterostigmomorpha	SMAR001399, SMAR001400

n/a The absence of these genes from the genome of *S. maritima* is uninformative because the HGT events happened elsewhere in the tree.  
<sup>a</sup>The number in parentheses shows the number of times the gene was recruited into the venom proteome if that differs from the number of HGT events<sup>23</sup>.  
<sup>b</sup>The identity of all paralogous loci is given. All are protein-coding loci with introns. Different paralog groups are indicated in parentheses.  
<sup>c</sup>Due to uncertainty about centipede phylogeny<sup>52</sup> we cannot distinguish between a single HGT into Pleurostigmophora (non-scutigeromorph centipedes), followed by a loss in Craterostigmomorpha, or a HGT into Amalpighiata (Lithobiomorpha + Epimorpha). Both these hypotheses suggest a loss in hemicopid lithobiomorphs.  
<sup>d</sup>The direction of transfer is ambiguous.

Although the structure of the gene tree is complex (Fig. 1; see Supplementary Fig. 1 for full tree), it shows that centipede  $\beta$ -PFTxs transferred twice from bacteria, once into the stem lineage of centipedes or arthropods (upper clade with 94% bootstrap support in Fig. 1), and once into the lithobiomorph lineage (located in the lower clade). This inference is supported by tree topology tests, which strongly reject monophyly of centipede  $\beta$ -PFTxs (see Supplementary Data 1). The structure of the tree, especially the complex interleaving pattern of bacterial, fungal, plant, and animal sequences in the lower clade of Fig. 1, suggests a complex history of multiple HGTs from bacteria to eukaryotes as well as losses of  $\beta$ -PFTx. For instance, an early transfer of  $\beta$ -PFTx into the arthropod stem lineage implies that it was lost in non-centipede myriapods and pancrustaceans, according to the current consensus on arthropod phylogeny<sup>28</sup>. However, the pronounced phylogenetic disjunction of the non-centipede animal sequences, and the lack of species from phyla with a strong representation in our custom (see “Methods”) and public sequence databases, such as arthropods, molluscs and nematodes, suggest that multiple HGTs have occurred from bacteria to animals. This interpretation is supported by tree topology tests that reject animal monophyly (see Supplementary Data 1).

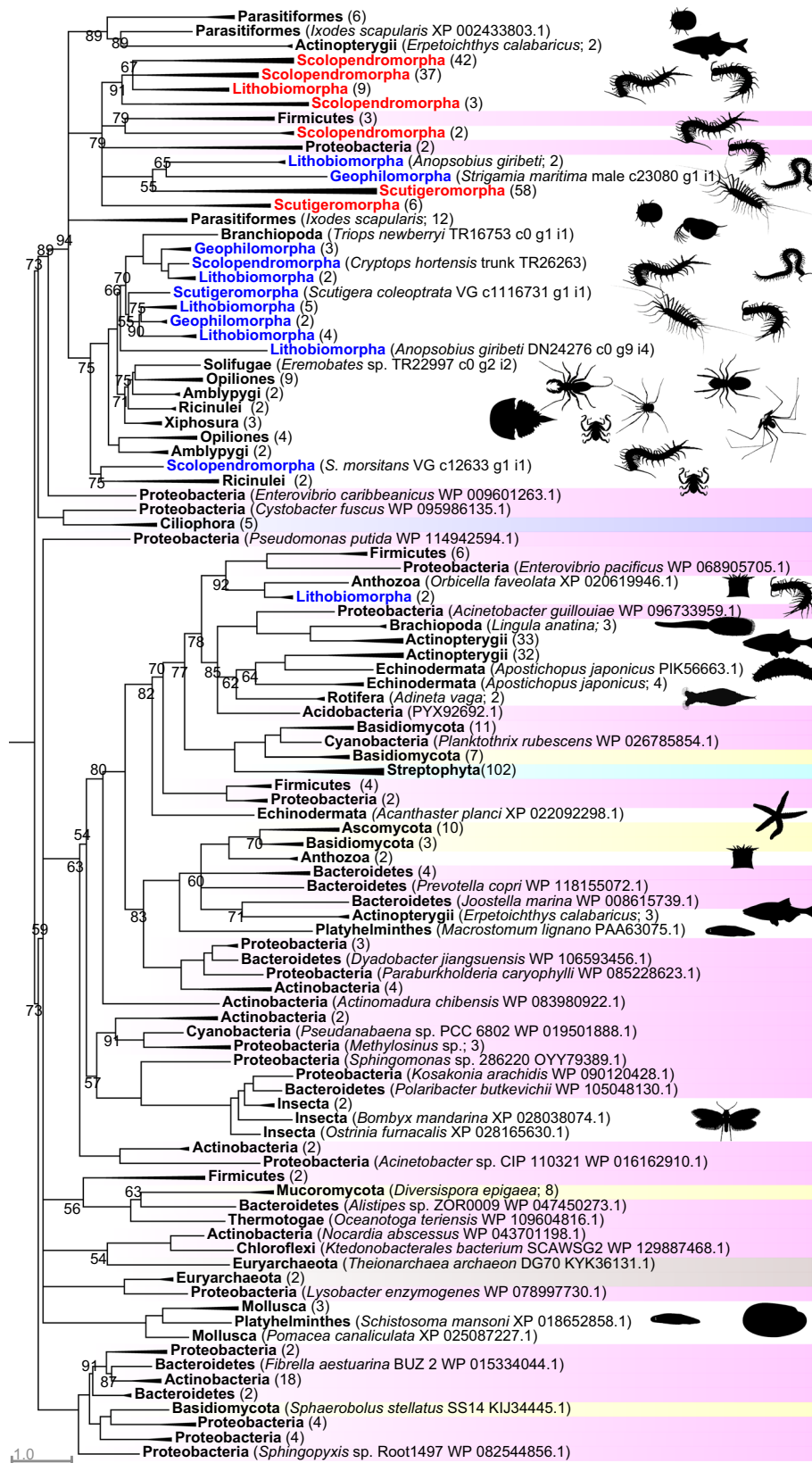
The  $\beta$ -PFTxs of *S. maritima* map to three protein-coding paralogous genomic loci with introns (see Table 1). The phylogenetic distribution of these paralogs in three sub-clades of centipede sequences in the upper clade of Fig. 1 shows that the duplications that produced them happened early in the evolution of centipedes. However,  $\beta$ -PFTx and the other three protein families that were recruited into the ancestral centipede venom are absent from the venom proteome of *S. maritima*, which shows that streamlining of venom arsenals occurs alongside the recruitment and diversification of new components<sup>23</sup>.

The  $\beta$ -PFTxs produced by bacteria are virulence factors that contribute to pathogenicity by the lysing of host cells<sup>29</sup>. Interestingly, although they are not expressed in their tentacle

venom, cnidarian  $\beta$ -PFTxs, which were horizontally transferred independently from those found in the venoms of centipedes and arachnids, are secreted into the pharynx and gut and aid digestion by disintegrating prey tissues, although their paralytic activity may also assist in prey immobilisation<sup>10,30,31</sup>. There is no experimental data for the role of  $\beta$ -PFTxs in centipedes, but they are believed to be at least in part responsible for the cytolytic activities of centipede venoms by the formation of transmembrane pores<sup>32</sup>. The great diversity of  $\beta$ -PFTx transcripts expressed in centipede venom proteomes, and the abundance of their expression<sup>22,23,33,34</sup>, suggest that  $\beta$ -PFTx likely plays important roles in prey immobilisation and processing.

#### Bacterial exotoxins probable source of PCPDP-like proteins.

We previously detected proteins with a pesticidal crystal protein domain (InterPro accession IPR036716) in the venom of *Lithobius forficatus*<sup>23</sup>. Homologous sequences are also present in transcriptomes of other centipedes from both lithobiomorph families (Lithobiidae: *L. forficatus*, *E. cavernicolus*; Henicopidae: *A. giribeti*, *P. validus*). All centipede PCPDP-like sequences cluster together in a strongly supported clade that is embedded in a paraphyletic backbone of bacterial PCPDP sequences (Fig. 2; see Supplementary Fig. 2 for full tree). The tree also shows that PCPDP-like proteins were independently transferred into beetles, a cnidarian and a tardigrade. This is supported by topological tree tests that strongly reject metazoan monophyly (see Supporting Data 1). The clade of centipede sequences includes species collected from the UK, Europe, North America, New Zealand, and Australia, and contains no interleaved bacterial sequences. This strongly suggests that the PCPDP-like sequences are bona fide centipede sequences rather than bacterial contaminants. Although on current evidence we cannot categorically reject the possibility that PCPDP-like protein is produced by symbionts, further evidence against this conclusion is that the centipede



sequences are very distinct from their nearest bacterial relatives (see below), which is reflected by the relatively long branch leading to the centipede clade. Lastly, a morphological study of the venom system of *L. forcipatus* found no evidence for bacterial symbionts in the venom producing and secreting tissues<sup>35</sup>.

The role of PCPDP-like proteins in centipede venom remains unknown, but our results suggest they evolved from bacterial insecticidal pore-forming toxins. The most intensely studied bacterial PCPDPs are pore-forming insecticidal endotoxins known as Cry toxins or  $\delta$ -endotoxins, which are used widely in



**Fig. 1 A maximum likelihood tree of  $\beta$ -PFTx sequences shows two clades of centipede  $\beta$ -PFTx sequences nested within a paraphyletic backbone of bacterial sequences.** The tree shows that the centipede  $\beta$ -PFTxs originated from at least two bacterial HGTs, one along the centipede or arthropod stem lineage (represented by the clade at the top of the tree with 94% bootstrap support), and one within the lithobiomorph lineage (represented by the clade of two lithobiomorph sequences lower down the tree). Centipede sequences are coloured blue (present in transcriptomes) and red (present in transcriptomes and venom proteomes). Highlighted sequences are Bacteria (pink), Euryarchaeota (brown), Protozoa (purple), Fungi (yellow), and Streptophyta (cyan). Metazoan sequences are not highlighted. Collapsed clades have the number of included sequences indicated in parentheses. For the uncollapsed tree see Supplementary Fig. 1. The tree was reconstructed using the WAG + R7 model and is displayed as midpoint rooted. Bootstrap support values are shown for each clade, and clades with support <50% are collapsed into polytomies. Clades without bootstrap values have >95% support. Non-centipede images are sourced from Phylopic ([www.phylopic.org](http://www.phylopic.org); credit for the Opiliones image is with Gareth Monger: <https://creativecommons.org/licenses/by/3.0/>).

GM crops<sup>36–39</sup>. They are produced by *Bacillus* species in the *B. cereus* group<sup>40,41</sup>, especially *B. thuringiensis*, the entomopathogenic bacterium from which they were first described, and which feeds on the insects killed by the toxin<sup>42</sup>. Cry toxins consist of three conserved domains: an N-terminal domain of  $\alpha$ -helices that is thought to be responsible for insertion into the cell membrane and pore formation, plus a middle and a C-terminal domain comprising  $\beta$ -sheets that are involved in receptor interactions, and which may confer host-specific toxicity<sup>37,43,44</sup>. Cry toxins are not secreted, but released as parasporal crystalline bodies through lysis of the spore-forming bacterial cell. The Cry toxin genes are located on plasmids, and plasmid transfer may explain why three-domain Cry proteins or genes have been found in several bacterial species outside the *B. cereus* group<sup>37,41</sup>.

In addition to three-domain Cry proteins our tree also contains sequences from a broad range of bacterial phyla that only contain a single Cry toxin domain, which in all cases is the pore-forming N-terminal domain. The centipede and other eukaryotic PCPDP-like sequences likewise only contain this N-terminal domain. A hint of how centipedes may have repurposed an insecticidal bacterial toxin into a venom protein is suggested by the most closely related bacterial sequences. All bacterial sequences that group together with the centipede sequences in the clade at the top of Fig. 2 also only contain the pore-forming N-terminal domain, and like the centipede sequences include a signal peptide region. This suggests that the bacterial proteins are exotoxins that are secreted from cells, like the centipede PCPDP-like proteins. Unlike the centipede sequences, the bacterial sequences in this clade also contain C-terminal cell wall-binding repeats (InterPro accession IPR018337), and/or a ricin B lectin domain (InterPro accession IPR000772). Cell wall-binding and ricin domains could help bind such putative exotoxins to bacterial or eukaryotic host cells, enabling the N-terminal perforating domain's cytolytic action. The centipede PCPDP-like sequences may derive from such putative bacterial exotoxins, followed by loss of these target-binding domains. Alternatively, the centipede proteins may derive from a bacterial endotoxin, either a non-secreted single-Cry-toxin-domain protein, or a true three-domain Cry toxin, by adding a signal peptide. The low sequence similarity of the bacterial and centipede sequences makes it impossible to distinguish these possibilities. However, it is unlikely that only the N-terminal domain was transferred from bacteria and joined to a native centipede sequence because BLAST searches of the C-terminal region of the PCPDP-like sequences against centipede transcriptomes and the genome of *S. maritima* produce no hits.

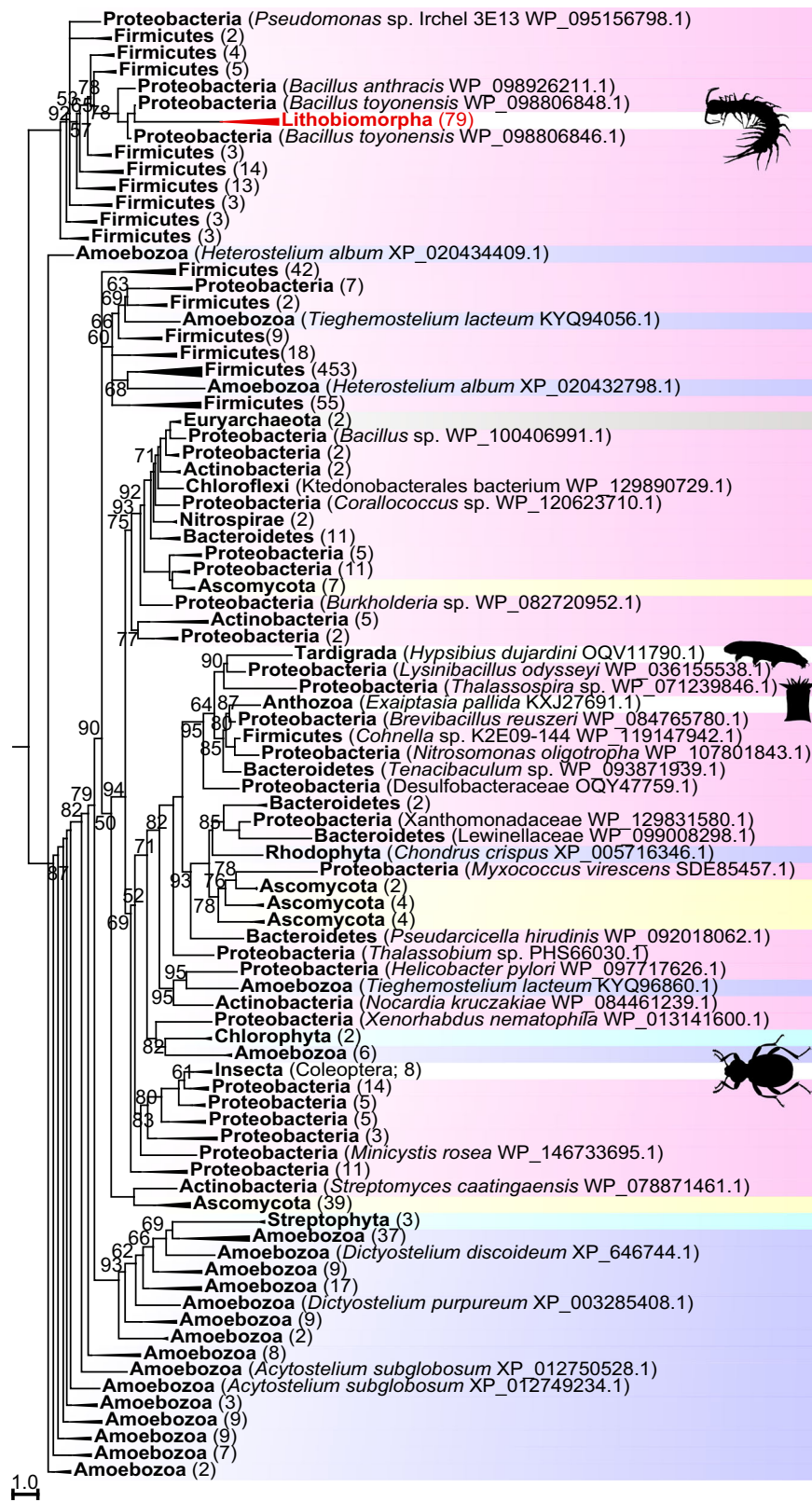
**Two bacterial HGTs of centiPADs.** We previously detected the enzyme peptidylarginine deiminase (PAD) in the venoms of two distantly related centipede species, *Thereuopoda longicornis* (order Scutigermorpha), and *Lithobius forficatus* (order Lithobiomorpha)<sup>22,23</sup>. Our phylogenetic analysis shows that these sequences are positioned in different parts of the tree, separated by many strongly supported nodes. Hence, centiPADs are the

result of two HGTs from different bacterial phyla. *T. longicornis* centiPAD derives from Gammaproteobacteria, while *L. forficatus* centiPAD derives from Bacteroidetes (Fig. 3; see Supplementary Fig. 3 for full tree). The centiPAD sequences are deeply nested within a large tree of bacterial sequences, confirming that human and bacterial PADs are evolutionarily unrelated<sup>45,46</sup>. Interestingly, the nesting of four fungal branches and a sequence derived from the black garden ant *Lasius niger* within the paraphyletic backbone of bacterial sequences suggest that PAD was transferred multiple times from bacteria to other eukaryotes as well.

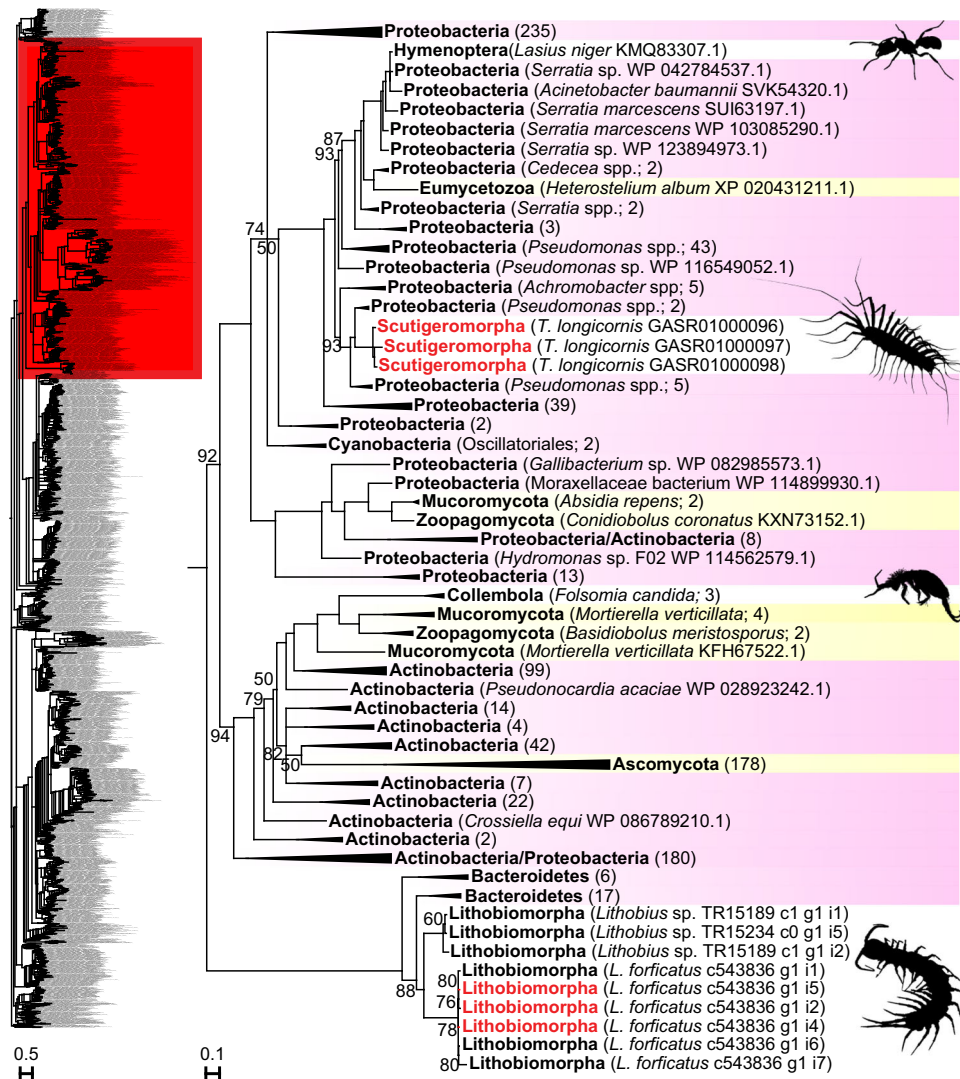
We cannot categorically reject the possibility that centiPADs are produced by bacterial symbionts, which, if true, would be the second example of an animal venom component being produced by bacteria<sup>47</sup>. However, the balance of evidence suggests that centiPADs are a bona fide centipede gene family. CentiPAD is a prominent component of the venom proteome of *T. longicornis*<sup>22</sup>, which is incompatible with it being due to accidental bacterial contamination. The sequences of *T. longicornis* can be up to 78% similar to the most closely related bacterial PAD sequences, but they share unique features that separate them from all bacterial sequences grouped in the same clade. Compared to related PAD sequences derived from the gammaproteobacterial genera *Pseudomonas*, *Cedecea*, *Aeromonas*, *Serratia*, *Stenotrophomonas*, and *Acinetobacter*, as well as the betaproteobacterial genera *Achromobacter*, *Paucibacter*, and *Undibacterium*, the centiPAD sequences uniquely have a Met593 and a single amino acid deletion at position 606 (see alignment in Supplementary Data 2). These distinctive differences further support the conclusion that the *T. longicornis* centiPADs are bona fide centipede sequences.

The *Lithobius* centiPAD sequences group together in a strongly supported clade without interleaving bacterial sequences. This clade groups sequences from specimens collected in the UK, continental Europe, and North America<sup>23,48,49</sup>. This strongly suggests that they are bona fide centipede sequences, a conclusion in line with the lack of evidence for microorganisms in the venom system of *L. forficatus*<sup>35</sup>. The European sequences (represented by UK sequences; an identical German sequence was excluded) form a sister clade to the American sequences. Because the latter were not determined to species by the original collectors<sup>48</sup>, it is unclear if they are *L. forficatus*, which was imported from Europe to North America some time before the end of the 19<sup>th</sup> century<sup>50</sup>. CentiPAD is absent from the transcriptomes of other lithobiomorph species: *Eupolybothrus cavernicolus*, *Paralamyctes validus*, and *Anopsobius giribeti*<sup>51,52</sup>. With the exception of *E. cavernicolus*, no venom glands were included in these transcriptomes, so these could be false negatives. However, the mean GC content of the UK centiPAD sequences is on the edge of the first quartile of all non-HGT venom protein sequences (0.385 vs. 0.384) from all centipede species analysed in our previous study<sup>23</sup> (see Supplementary Data 3), which suggests that the HGT probably occurred relatively recently.

A recent transfer is also likely for the *T. longicornis* centiPADs. The mean GC content of the three *T. longicornis* centiPAD sequences (0.588) is extremely skewed in the other direction and



**Fig. 2** A maximum likelihood tree of PCDP-like sequences shows a clade of centipede sequences nested within a paraphyletic backbone of bacterial sequences. It shows that the centipede sequences originated from a bacterial HGT into the lithobiomorph lineage. Centipede sequences are coloured red. Highlighted sequences are Bacteria (pink), Viridiplantae (cyan), Protozoa (purple), Euryarchaeota (brown), and Fungi (yellow). Metazoan sequences are not highlighted. Collapsed clades have the number of included sequences indicated in parentheses. For the uncollapsed tree see Supplementary Fig. 2. The tree was reconstructed using the VT + G4 model and is displayed as midpoint rooted. Bootstrap support values are shown for each clade, and clades with support <50% are collapsed into polytomies. Clades without bootstrap values have >95% support. Non-centipede images are sourced from Phylopic ([www.phylopic.org](http://www.phylopic.org)).



**Fig. 3** A maximum likelihood tree of PAD sequences shows two clades of centiPAD sequences nested within a paraphyletic backbone of bacterial sequences. The tree represents one clade nested within a larger tree (red highlight in inset) made up entirely of bacterial sequences. The tree shows that centiPADs originated from two bacterial HGTs, one within the lithobiomorph lineage, and one within the scutigermorph lineage. Centipede sequences are in black (present in transcriptomes) and red (present in transcriptomes and venom proteomes). Highlighted sequences are Bacteria (pink) and Fungi (yellow). Metazoan sequences are not highlighted. Collapsed clades have the number of included sequences indicated in parentheses. For the uncollapsed tree see Supplementary Fig. 3. The tree was reconstructed using the WAG + G4 model and is displayed as midpoint rooted. Bootstrap support values are shown for each clade, and clades with support <50% are collapsed into polytomies. Clades without bootstrap values have >95% support. Collembolan image was sourced from Phylopic ([www.phylopic.org](http://www.phylopic.org); credit for the Collembola image is with Birgit Lang: <https://creativecommons.org/licenses/by/3.0/>).

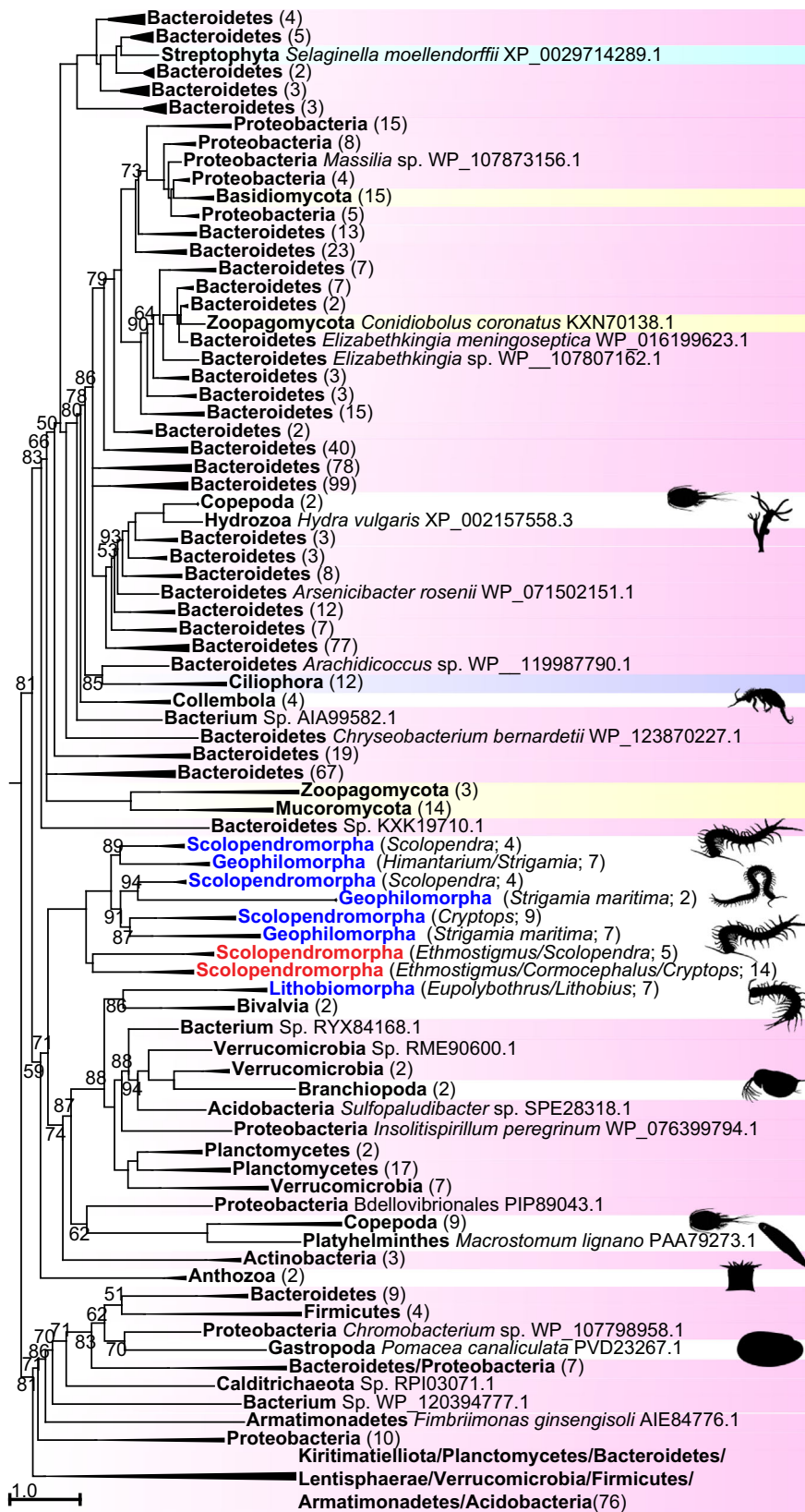
falls outside the 99th percentile (0.557) of all non-HGT centipede venom protein sequences. This skew and the sequence similarity of the centipede and bacterial sequences indicate that this HGT may have happened relatively recently. The absence of centiPAD sequences from the transcriptomes of other scutigermorphs (*Scutigera weberi*, *Sphendononema guilgingii*, and *Scutigera coleoptrata*)<sup>23,52</sup> provides further support for a relatively recent HGT. Since only the transcriptome of *S. coleoptrata* contains venom gland tissue the other two may be false negatives. We consider this unlikely, however, because they represent different scutigermorph families, while *S. coleoptrata* and *T. longicornis* belong to the family Scutigerae. The unique presence of centiPAD in *T. longicornis* therefore suggests that this gene was transferred after its lineage split off from that of *S. coleoptrata*, which is estimated to have happened by about 200 million years ago<sup>53</sup>.

Bacterial PAD converts peptidylarginine into citrulline residues, and the effects of this process have been most intensely investigated

for the pathogenic bacterium *Porphyromonas gingivalis*. *Porphyromonas* PAD (PPAD) is a major virulence factor that causes inflammatory gum disease, and is a risk factor for rheumatoid arthritis<sup>45,46,54,55</sup>. How PPAD contributes to pathogenicity is an active area of research, and it may include defusing the host's immune system and the formation of protective biofilms<sup>55,56</sup>. It is unknown what role centiPADs play in centipede venom but modulating the activity of other venom components through posttranslational modification is one possibility. The centiPAD sequences from both species have conserved the five catalytic residues responsible for PPAD's enzymatic activity (Asp1372, His2321, Asp2323, Asn2928, Cys4010 in the PAD alignment in Supplementary Data 5), but they have changed two residues that determine substrate specificity of bacterial PADs<sup>46</sup>.

**One or two bacterial HGTs of DUF3472-domain proteins.** Proteins with a domain of unknown function DUF3472 (InterPro





accession IPR021862) are found in the venom of several species of scolopendromorph centipedes, as well as in geophilomorph and lithobiomorph venom gland and non-venom gland transcriptomes<sup>23,33,34,57,58</sup>. In addition, many of the sequences have an N-terminal DUF5077 domain (InterPro accession

IPR031712). Our phylogenetic analysis places the centipede sequences into two clades separated by bacterial and metazoan sequences (Fig. 4; see Supplementary Fig. 4 for full tree). This suggests that DUF3472-domain proteins may have transferred twice from bacteria to centipedes, once into the lineage leading to



**Fig. 4 A maximum likelihood tree of sequences with DUF3472-domains shows two clades of centipede sequences nested within a paraphyletic backbone of bacterial sequences.** This suggests that the centipede sequences may have originated from two bacterial HGTs, one into the epimorph lineage, and one within the lithobiomorph lineage. However, tree topology tests cannot reject centipede monophyly (see Supplementary Data 1). Centipede sequences are coloured blue (present in transcriptomes) and red (present in transcriptomes and venom proteomes). Highlighted sequences are Bacteria (pink), Protozoa (purple), Streptophyta (cyan), and Fungi (yellow). Metazoan sequences are not highlighted. Collapsed clades have the number of included sequences indicated in parentheses. For the uncollapsed tree see Supplementary Fig. 4. The tree was reconstructed using the WAG + R10 model and is displayed as midpoint rooted. Bootstrap support values are shown for each clade, and clades with support <50% are collapsed into polytomies. Clades without bootstrap values have >95% support. Copepod image was sourced from Phylopic ([www.phylopic.org](http://www.phylopic.org); credit for the Collembola image is with Birgit Lang: <https://creativecommons.org/licenses/by/3.0/>).

Epimorpha (geophilomorphs and scolopendromorphs), and once into lithobiid lithobiomorphs. Topological tree tests cannot statistically reject centipede monophyly, but do reject metazoan monophyly (see Supplementary Data 1). This shows that DUF3472-domain proteins have been transferred from bacteria to animals multiple times, like  $\beta$ -PFTxs and PCPDP-like proteins. DUF3472-domain proteins from *S. maritima* map to four protein-coding genomic loci with introns (see Table 1), and the tree suggests that these and the multiple copies found in scolopendromorphs are the result of several rounds of gene duplication.

### Multiple HGTs between fungi, oomycetes and centipedes.

Centipedes not only express four gene families in their venoms that were horizontally transferred from bacteria, but also three gene families that find their nearest homologues in fungi and oomycetes (water molds). GEOTX02 is a peptide present in the venom of the geophilomorph *S. maritima*, and similar sequences with a corresponding cysteine framework are restricted to a few species of ascomycete fungi. The sequences exhibit two distinct cysteine patterns, with 8 or 10 cysteine residues in the mature domain of the peptide, with the latter being restricted to the top clade in the tree with 85% bootstrap support (Fig. 5a; see Supplementary Fig. 5 for full tree). The centipede sequences map to four paralogue groups of genes with introns in the genome of *S. maritima*, with the clade with 74% bootstrap support representing paralogue groups 1–3 and the collapsed clade of eleven *S. maritima* sequences representing paralogue group 4 (see Table 1). The tree suggests that the centipede sequences with the two different cysteine patterns may have resulted from two HGTs, although a tree topology test cannot reject centipede monophyly (see Supplementary Data 1), and the direction of these horizontal transfers remains uncertain. The ascomycetes included in the tree belong to two orders (Dothideomycetes and Sordariomycetes) and include species known to infect animals and plants. The transfers therefore possibly involved an arthropod-infecting ascomycete as either a donor or recipient of GEOTX02.

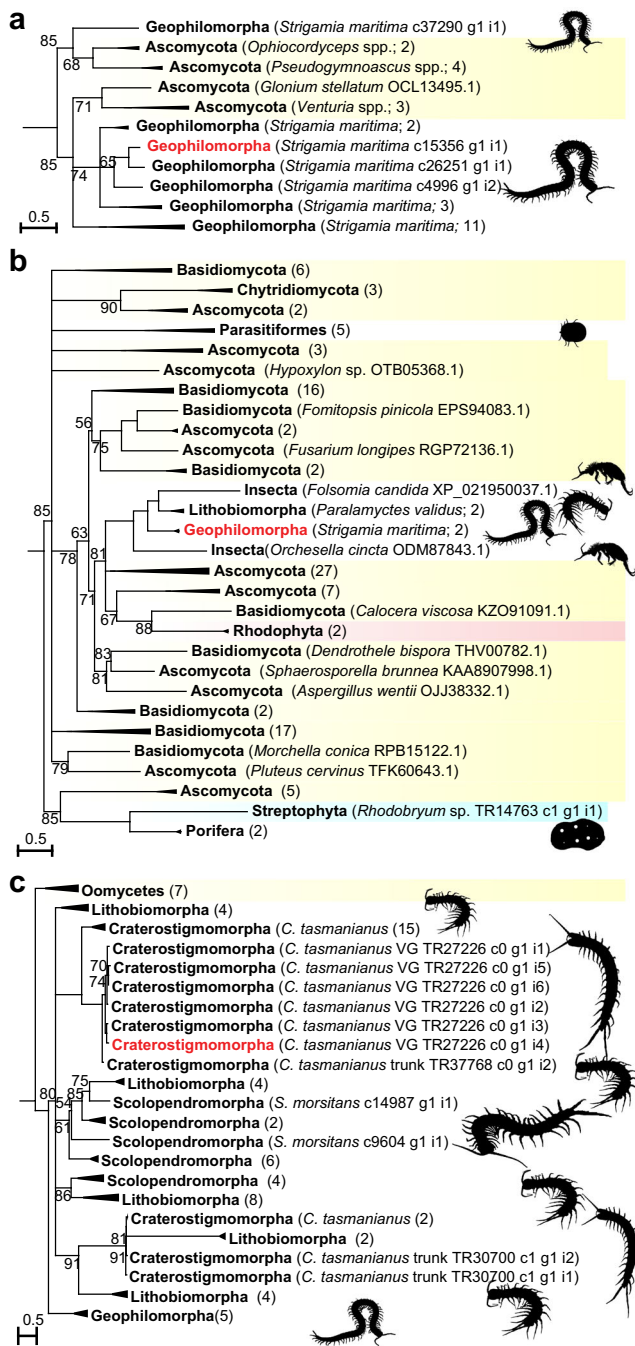
Unchar05 is another venom protein family that has been horizontally transferred between centipedes and fungi. Unchar05 is present in the venom of *S. maritima* but is also found in a trunk transcriptome of the lithobiomorph *Paralamyctes validus*. The two unchar05 transcripts identified in the venom proteome of *S. maritima* map to a protein-coding genomic locus with introns (SMAR002275), which is one of five paralogous loci (see Table 1), four of which are expressed as transcripts in the venom gland of *S. maritima*. Our phylogenetic analysis (Fig. 5b; see Supplementary Fig. 6 for full tree) shows that unchar05 was transferred into centipedes from fungal donors. The centipede sequences group in a clade with sequences from two species of springtails, *Folsomia candida* and *Orchesella cincta*, but neither the centipede nor the springtail sequences are monophyletic. This taxonomic interleaving of sequences and the phylogenetic disjunction between the centipede species suggest that unchar05 horizontally transferred

twice into centipedes. This may also be true for the springtails, where unchar05 homologues are found in at least two different families, and whose genomes contain hundreds of genes of HGT origin<sup>5,6</sup>. Moreover, the tree also contains a well-supported clade of mite sequences that includes species that have also previously been shown to have received horizontally transferred fungal genes<sup>3</sup>.

Although a tree topology test cannot reject metazoan monophyly in our tree (see Supplementary Data 1), we consider the alternative hypothesis of a single early HGT of unchar05 into animals followed by rampant losses to be less convincing. To explain the large phylogenetic disjunction of the sequences on various levels—within centipedes, within insects, and within animals—would require an immense amount of gene loss throughout the animal kingdom to leave just this handful of metazoan homologues, several of which represent taxa already known to be recipients of horizontally transferred genes.

The third gene family that has probably undergone eukaryotic HGT is Unchar16. It encodes cysteine-rich proteins found in the venom gland and non-venom gland transcriptomes of pleurostigmophoran (non-scutigeromorph) centipedes, as well as in the venom of the craterostigmomorph *Craterostigma tasmanianus*. Unchar16 maps to two protein-coding paralogous loci with introns in the genome of *S. maritima* (see Table 1). Our searches identified small secretory proteins from plant-parasitic oomycetes as homologues based upon sequence similarity and corresponding cysteine patterns. Unchar16 has undergone marked sequence evolution in centipedes, and all centipede sequences group in a well-supported clade when the tree is rooted with oomycetes (Fig. 5c; see Supplementary Fig. 7 for full tree). However, two different HGT scenarios may explain the data depending on how the tree is rooted.

Oomycetes originated at about the same time as centipedes, about 430 million years ago<sup>59</sup>, so a HGT between early oomycete and centipede lineages is possible if unchar16 was transferred from oomycetes into the stem lineage of pleurostigmophoran centipedes. However, the early evolutionary history of oomycetes and the taxonomic distribution of oomycete unchar16 homologues casts doubt on this scenario. Early diverging oomycete lineages are exclusively marine, with the exception of the genus *Haptoglossa*<sup>60,61</sup>. Moreover, the oomycete homologues of unchar16 that we identified belong to the predominantly terrestrial oomycete order Peronosporales, which is a lineage that evolved much later, in the early Mesozoic about 225–190 million years ago<sup>59</sup>. This suggests that unchar16 may have horizontally transferred much more recently from centipedes into the peronosporalean lineage of oomycetes—the reverse transfer would require independent HGTs into all four pleurostigmophoran centipede lineages. HGT is known to have contributed to the evolution of oomycete secretomes<sup>62,63</sup>, but which centipede lineage functioned as a donor of unchar16 in this scenario remains unclear given the lack of resolution in the tree. On the balance of available evidence, we prefer this second scenario, but hope that future research will shed further light on this tantalizing riddle.



### HGT is a potentially major mechanism of venom evolution.

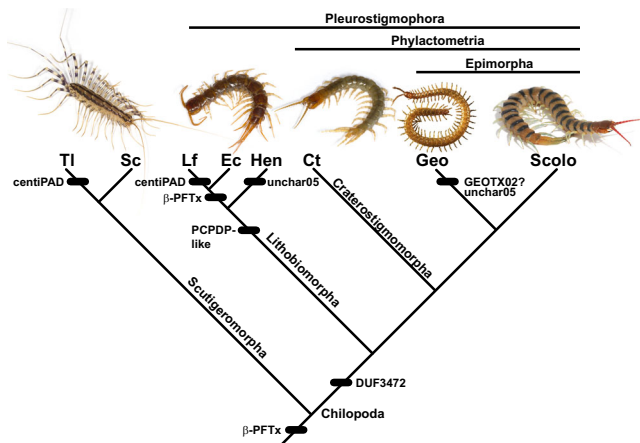
Our results suggest that HGT has been a key factor in the expansion and diversification of centipede venoms in all five orders throughout their evolutionary history (Fig. 6). Because genes were horizontally transferred from bacteria and fungi both deeply and repeatedly in the phylogeny of centipedes, we expect that the vast majority of centipede species produce venoms that include multiple horizontally transferred components. Because proteotranscriptomic venom profiles are currently available for only a small number of the more than 3,100 described species of centipedes, new insights into the full impact of HGT on centipede venom evolution are likely to emerge from future studies.

Our findings increase the number of animal venom protein families with well-supported HGT origins from three to at least eight, which increases the number of known HGT events stocking venom arsenals from five or six to at least thirteen. We show that

**Fig. 5** Maximum likelihood trees showing eukaryotic HGTs between fungi, oomycetes, and centipedes. **a** Tree of GEOTX02 homologues showing that the centipede sequences are distributed across two clades, and interleaved with ascomycete sequences. The direction and number of HGTs (one or two) is uncertain. The tree was reconstructed using the VT + I + G4 model and is midpoint rooted. For the uncollapsed tree see Supplementary Fig. 5. **b** Tree of unchar05 homologues showing the four centipede sequences grouping in a clade with two collembolan sequences, nested within a paraphyletic backbone of fungal sequences. The tree shows that the centipede sequences likely originated from two fungal HGTs, one into the geophilomorph lineage, and one within the lithobiomorph lineage. The tree was reconstructed using the WAG + R5 model and is midpoint rooted. For the uncollapsed tree see Supplementary Fig. 6. **c** Tree of unchar16 homologues showing a clade of centipede sequences that is the sister group to a clade of oomycete sequences. The direction of HGT is unclear. The tree was reconstructed using the VT + R3 model and is rooted with the oomycete sequences. For the uncollapsed tree see Supplementary Fig. 7. For each tree, bootstrap support values are shown for each clade and clades with support <50% are collapsed into polytomies. Clades without bootstrap values have >95% support. Centipede sequences are coloured black (present in transcriptomes) and red (present in transcriptomes and venom proteomes). Highlighted sequences are Fungi (yellow), Rhodophyta (reddish brown), and Streptophyta (cyan). Metazoan sequences are not highlighted. Collapsed clades have the number of included sequences indicated in parentheses. Non-centipede images are sourced from Phylopic ([www.phylopic.org](http://www.phylopic.org); credit for the Collembola images is with Birgit Lang: <https://creativecommons.org/licenses/by/3.0/>).

centipedes are the first known animals with venoms used for predation and defence that contain multiple gene families derived from HGT. It is likely that HGT contributions to venom evolution are a much more widespread phenomenon. More than a hundred animal lineages have evolved venoms<sup>8</sup>, and recent proteotranscriptomic studies of venoms from a wide range of taxa have identified substantial numbers of protein families with few or no known metazoan homologues (e.g.<sup>16–21</sup>). Such gene families are especially promising for identifying new HGT candidates, but this requires a targeted approach, like the one adopted here, that goes beyond the standard BLAST-based annotation pipelines commonly used in venom profiling studies.

Our findings expand the insights generated by previous research into how HGT can increase the adaptive versatility of organisms<sup>1,2</sup>. Our results suggest that HGT can allow a venomous lineage to reap the immediate adaptive benefits of genes evolved in unrelated lineages if the gene products are preadapted to a venom function. For instance, the incorporation of a cytolytic bacterial pore-forming toxin, such as  $\beta$ -PFTx, into the ancestral centipede venom may have conferred an immediate functional benefit, for example in prey immobilisation. In this scenario, the pore-forming activity of the bacterial protein is a preadaptation that would have allowed the protein to take on this function in the centipede venom without first having to evolve modifications to gain a venom function. This parallels, for example, the use of detoxifying enzymes by herbivorous arthropods that were horizontally transferred from, and similarly used, by bacterial and fungal donors<sup>64</sup>. The selective benefit of the horizontal transfer of  $\beta$ -PFTx into the earliest centipede venom could have been substantial because it is just one of two putative toxins that could have been involved in prey immobilization. The other three protein families that we reconstructed as present in the ancestral centipede venom are metalloprotease family M12A, glycoside hydrolase family 18, and centipede CAP1 (cysteine-rich secretory proteins, antigen 5 and pathogenesis-related protein family 1), which is the second putative venom toxin<sup>23</sup>. The recruitment of



**Fig. 6 Phylogenetic distribution of centipede venom gene families horizontally transferred from bacteria and fungi.** ‘?’ indicates uncertainty in the direction of transfer. Taxon abbreviations are as follows. TI: *Thereuopoda longicornis*; Sc: *Scutigera coleoptrata*; Lf: *Lithobius forficatus*; Ec: *Eupolybothrus cavernicolus*; Hen: Henicopidae; Ct: *Craterostigma tasmanianus*; Geo: Geophilomorpha; Scolo: Scolopendromorpha.

$\beta$ -PFTx into the ancestral Catepede venom represents the first known example of HGT contributing to the evolutionary origin of venom in a lineage. Horizontal transfer could therefore have been a crucial step in setting centipedes on the selective trajectory that eventually led to the complex venoms of modern species.

The fact that the centipede venom homologues of the three horizontally transferred bacterial virulence factors for which there are functional data have retained the structural domains involved in pore-formation ( $\beta$ -PFTx and PCPDP-like proteins), or conserved the catalytic sites involved in enzymatic action (centiPAD), is consistent with a continuity of function and adaptive value from donor to recipient taxa. Moreover, the gene duplications that have subsequently occurred in the genome-confirmed gene families underlines a commonly observed feature of the route to the functional consolidation and diversification of horizontally transferred genes<sup>2</sup>. Our results therefore show that HGT can provide a fast track channel for the evolution of novelty by the exaptation of bacterial weapons for new functions in animal venoms.

## Methods

**Initial identification of HGT candidates.** We used the transcriptomic and proteomic data from Undheim et al.<sup>22</sup> and Jenner et al.<sup>23</sup> to identify HGT candidates expressed in centipede venom glands and venoms. Manual inspection of BLAST results generated for these studies for more than 90 venom protein families yielded sixteen protein families with either non-metazoan hits, and/or few or no metazoan hits ( $\beta$ -PFTx, centiPAD, CHILOTX01, DUF3472, GEOTX02, LTHTX01, LTHTX03, PCPDP-like, SCTX01, SCTX02, SLPTX02, SLPTX04, SLPTX06, SLPTX30, unchar05, and unchar16). We performed a protein BLAST search of these HGT candidate sequences against a local version of the NCBI non-redundant (nr) database (downloaded from the NCBI FTP Server <ftp://ftp.ncbi.nlm.nih.gov/> on 5 June 2019) with BLAST version 2.4.0, and an *E*-value cut off of *e*-3. Significant hits against non-metazoan sequences were found for  $\beta$ -PFTx, centiPAD, DUF3472, GEOTX02, PCPDP-like, unchar05, unchar16, and SLPTX02. These BLAST results were submitted to the Alienness web server (<http://alienness.sophia.inra.fr/cgi/index.cgi>), which is a tool designed to detect HGT candidates<sup>65</sup>. Alienness calculates an Alien Index for each query sequence based on the *E*-values of the best BLAST hits to putative candidate donors (non-metazoan) and recipient (metazoan) taxa. The following taxa and taxon codes were excluded from the Alien Index calculations as self-hits for the different protein families that generated positive Alien Indices:  $\beta$ -PFTx: *Cormocephalus westwoodi* (1096223), *Ethmostigma rubripes* (62613), *Lithobius forficatus* (7552), *Scutigera coleoptrata* (29022), *Scolopendra alternans* (1329349), *Sco. morsitans* (943129), *Sco. subspinipes* (55038), *Thereuopoda longicornis* (353555), *Ixodes scapularis* (6945), *Limulus polyphemus* (6850), *Strigamia maritima* (126957), *Cryptops hortensis* (1268897), *Acuclavella merickeli* (703423), *Damon variegatus* (317683), *Cryptocellus becki* (1642531),

*Lithobius* (7551), centipedes (7540); centiPAD: *L. forficatus* (7552), *T. longicornis* (353555), centipedes: 7540; DUF3472: *S. maritima* (126957), *E. rubripes* (62613), *Himantarium gabrielis* (241672), *Sco. morsitans* (943129), *Sco. subspinipes* (55038), *C. westwoodi* (1096223), *Cryptops hortensis* (1268897), *L. forficatus* (7552), centipedes (7540); GEOTX02: *S. maritima* (126957), centipedes (7540); unchar05: *S. maritima* (126957), centipedes (7540); unchar16: *Craterostigma tasmanianus* (60162), *Sco. morsitans* (943129), *L. forficatus* (7552), *S. maritima* (126957), centipedes (7540); SLPTX02: centipedes (7540), *L. forficatus* (7552), *Scu. coleoptrata* (29022), *C. tasmanianus* (60162), *C. hortensis* (1268897), *H. gabrielis* (241672), *Lithobius* (7551), *Sco. Subspinipes* (55038), *E. rubripes* (62613). Results are summarized in Supplementary Data 7. The PCPDP-like gene family didn't generate any BLAST hits. However, because the sequences contain an insecticidal delta-endotoxin domain known only from bacteria we included this gene family in our analyses as well. SLPTX02 was dropped from further consideration because the broad phylogenetic distribution of homologues suggests an ancient origin of this protein family.

**Construction of phylogenetic datasets.** All analyses were performed on amino acid translations of the transcriptome sequences. We used HMMER v3.2.1 (<http://hmmerr.org>) with default settings to generate Hidden Markov Models for each of the seven venom protein families with possible non-metazoan origins, and retained all hits above HMMER's default inclusion threshold (per-sequence *E*-value of 0.01). Geneious version 11.1.5 (<https://www.geneious.com>) was used to construct alignments for training HMMER profiles, using the local paired iterative alignment method (L-INS-i) in MAFFT v7.450<sup>66</sup> (see Supplementary Data 5 and 6 for the alignments and profiles). We included in these alignments all the full-length centipede sequences that we generated for these gene families in our previous studies<sup>22,23</sup>. For  $\beta$ -PFTxs and the PCPDP-like gene family, we additionally included a selection of outgroup taxa, and the PCPDP-like alignment was limited to the N-terminal Cry toxin domain. We used the HMMER profiles to search against a local fasta version of the nr database (downloaded from the NCBI FTP Server <ftp://ftp.ncbi.nlm.nih.gov/> on 21 May 2019) for possible homologues of the centipede sequences. We also used these profiles to search a previously published<sup>67</sup> custom database of 155 de novo assembled and translated transcriptomes obtained from the NCBI Sequence Read Archive (SRA), representing 134 animal species, with 121 arthropod species including eight millipede whole body and eight centipede whole body or trunk transcriptomes, as well as seven species of fungi, plants, and choanoflagellates (see Supplemental Table S2 in Dash et al.<sup>67</sup>). This database was supplemented with assembled transcriptomes for the centipedes *Paralamyctes validus*, *Anopsobius giribeti*, *Scutigera weberi*, and *Sphendononema guildingii*<sup>52</sup>. Complementing these transcriptome-based sequence data we used the HMMER profiles to search for homologues in 25 metazoan Ensembl genomes (<http://ensemblgenomes.org>) representing these major lineages: Cnidaria: *Thelohanellus kitauei*, *Nematostella vectensis*; Placozoa: *Trichoplax adhaerens*; Ctenophora: *Mnemiopsis leidyi*; Deuterostomia: *Strongylocentrotus purpuratus*; Rotifera: *Adimeta vaga*; Brachiopoda: *Lingula anatina*; Mollusca: *Octopus bimaculoides*, *Crassostrea gigas*, *Lottia gigantea*; Annelida: *Capitella teleta*, *Helobdella robusta*; Nematoda: *Pristionchus pacificus*, *Caenorhabditis elegans*; Arthropoda, Arachnida: *Ixodes scapularis*, *Sarcoptes scabiei*, *Tetranychus urticae*, *Stegodyphus mimosarum*; Arthropoda, Pancrustacea: *Daphnia pulex*, *Lepeophtheirus salmonis*, *Folsomia candida*, *Nasonia vitripennis*, *Apis mellifera*, *Megaselia scalaris*, *Anopheles gambiae*.

Once a comprehensive list of homologues was generated, we removed identical sequences using CD-HIT v4.6<sup>68</sup>, and examined and filtered false positives using CLC Main WorkBench v7 (Qiagen, Aarhus, Denmark) and Geneious v11.1.5 (<https://www.geneious.com>). In the case of  $\beta$ -PFTx, we also filtered the non-chilopod sequences with CD-HIT to only include sequences with <95% sequence identity due to a large number of identified unique homologues (2164 sequences). To create datasets of manageable size for PAD, PCPDP, and DUF3472, while retaining a broad net for capturing putative donor taxa and sampling metazoan homologues, the identified homologues were first sorted to Kingdom and then filtered with CD-HIT to include only sequences with <90% (bacteria, fungi, protists, and viruses) or 70% sequence identity (non-myriapod animals, Archaea, and plants). Due to the large number of PAD homologues still retained by this approach (6716 sequences), we then removed all sequences with a pairwise distance to any chilopod sequence >0.5.

The remaining sequences were aligned using the local paired iterative alignment method (L-INS-i) in MAFFT v7.304b<sup>66</sup>. For the alignment of GEOTX02, we first aligned the structurally important conserved cysteines<sup>69</sup>, and then used the MAFFT regional alignment ruby script to align the pre-, inter-, and post-cysteine regions by local paired iterative alignment method as above. All alignments are included in Supplementary Data 4. We used InterProScan<sup>70</sup> as implemented in Geneious v11.1.5 (<https://www.geneious.com>) to generate protein domain annotations for all alignments (see Supplementary Data 8). The evolutionary history of each protein family was then reconstructed using a molecular phylogenetic approach. The most appropriate evolutionary model was determined using ModelFinder<sup>71</sup>, before using IQ-TREE v1.5.5<sup>72</sup> to reconstruct molecular phylogenies by maximum likelihood, and estimating branch support values by ultrafast bootstrap using 10,000 replicates<sup>73</sup>. Because taxonomic outgroups could not be designated we used midpoint rooting to root the trees. Trees were visualised in Archaeopteryx v0.9921<sup>74</sup>.



**Tree topology tests.** Likelihood ratio tests of constrained tree topologies were performed in IQ-TREE 2<sup>75</sup>, which implements several different tests. Each test compares support for the unconstrained optimal maximum likelihood tree with a tree that constrains the monophyly of selected taxa as a polytomy. Results are given in Supplementary Data 1. Mesquite 3.61 (build 927)<sup>76</sup> was used to build constraint topologies.

**Mapping of genes against the *Strigamia maritima* genome.** All sequences belonging to candidate HGT gene families present in the venom gland transcriptome and venom proteome of *S. maritima* were mapped against its genome using the TBLASTN search function with an E-value cut off of  $e^{-5}$  on the EnsemblMetazoa web portal at [http://metazoa.ensembl.org/Strigamia\\_maritima/Info/Index](http://metazoa.ensembl.org/Strigamia_maritima/Info/Index) (last accessed 1 April 2020).

**GC contents analyses.** We used CLC Main WorkBench v7 (Qiagen, Aarhus, Denmark) to calculate GC frequencies of all nucleotide sequences encoding centipede venom proteins and peptides published by Jenner et al.<sup>23</sup>. Descriptive statistics were calculated with GraphPad Prism v8.4.1 (GraphPad Software, La Jolla California USA, [www.graphpad.com](http://www.graphpad.com)), and are available as Supplementary Data 3.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The transcriptomic custom database used in this study is available in the NIRD Research Data Archive with identifier 10.11582/2020.00067 [[https://lantigena.com/II-XpFdcjUOuQ3kwVUGwNUCawa65ouPHcGAUUYz4\\_G8tW7vXlL81qj8DGsAVtk-Pln4FKNoqN6enY799zIGURLfFK78EEeGN7Vjv6rkUj6QgiCaGMuFn2wNUw-N3avmVFclTjxYAKWjK8PqF7hKgWurRu8L2F61L-640J09VwrlvwCQM](https://lantigena.com/II-XpFdcjUOuQ3kwVUGwNUCawa65ouPHcGAUUYz4_G8tW7vXlL81qj8DGsAVtk-Pln4FKNoqN6enY799zIGURLfFK78EEeGN7Vjv6rkUj6QgiCaGMuFn2wNUw-N3avmVFclTjxYAKWjK8PqF7hKgWurRu8L2F61L-640J09VwrlvwCQM)]. The transcriptome data from Undheim et al.<sup>22</sup> are available at the National Center for Biotechnology Information (NCBI) under bioprojects PRJNA200639, PRJNA200641, PRJNA200753, PRJNA200640, and PRJNA213032, while individually curated sequences are available in the Transcriptome Shotgun Assembly Sequence Database (<https://www.ncbi.nlm.nih.gov/nucleotide/>) as GASI01000001–GASI01000195, GASL01000001–GASL01000050, GASK01000001–GASK01000051, GASH01000001–GASH01000185, and GASR01000001–GASR01000119. Undheim et al.'s proteomic evidence are available as supplementary files associated with the original publication. The assembled transcriptomes from Jenner et al.<sup>23</sup> are available via the Natural History Museum's Data Portal (<https://data.nhm.ac.uk/dataset/evolution-of-centipede-venoms>; last accessed 30 June 2020).<sup>9</sup>), while the proteomic data are available in the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD013356. In addition, we used the following databases: NCBI non-redundant (nr) database (<https://www.ncbi.nlm.nih.gov/>), EnsemblMetazoa (<https://metazoa.ensembl.org/index.html>), and the databases in the InterPro Consortium (<https://www.ebi.ac.uk/interpro/>).

Received: 17 July 2020; Accepted: 12 January 2021;

Published online: 05 February 2021

## References

- Boto, L. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc. R. Soc. B* **281**, 20132450 (2014).
- Husnik, F. & McCutcheon, J. P. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.* **16**, 67–79 (2018).
- Dong, X. et al. Genomes of trombidid mites reveal novel predicted allergens and laterally transferred genes associated with secondary metabolism. *GigaScience* **7**, 1–33 (2018).
- Wybouw, N., Van Leeuwen, T. & Dermauw, W. A massive incorporation of microbial genes into the genome of *Tetranychus urticae*, a polyphagous arthropod herbivore. *Insect Mol. Biol.* **27**, 333–351 (2018).
- Faddeeva-Vakhrusheva, A. et al. Gene family evolution reflects adaptation to soil environmental stressors in the genome of the collembolan *Orchesella cincta*. *Genome Biol. Evol.* **8**, 2106–2117 (2016).
- Faddeeva-Vakhrusheva, A. et al. Coping with living in the soil: the genome of the parthenogenetic springtail *Folsomia candida*. *BMC Genomics* **18**, 493 (2017).
- Paganini, J. et al. Contribution of lateral gene transfers to the genome composition and parasitic ability of root-knot nematodes. *PLoS ONE* **7**, e50875 (2012).
- Schendel, V., Rash, L. D., Jenner, R. A. & Undheim, E. A. B. The diversity of venom: the importance of behavior and venom system morphology in understanding its ecology and evolution. *Toxins* **11**, 666 (2019).
- Cordes, M. H. J. & Binford, G. J. Evolutionary dynamics of origin and loss in the deep history of phospholipase D toxin genes. *BMC Evol. Biol.* **18**, 194 (2018).
- Moran, Y., Fredman, D., Szczesny, P., Grynberg, M. & Technau, U. Recurrent horizontal transfer of bacterial toxin genes to eukaryotes. *Mol. Biol. Evol.* **29**, 2223–2230 (2012).
- Martinson, E. O., Martinson, V. G., Edwards, R., Mrinalini & Werren, J. H. Laterally transferred gene recruited as a venom in parasitoid wasps. *Mol. Biol. Evol.* **33**, 1042–1052 (2016).
- Ribeiro, J. M. C. et al. An annotated catalogue of salivary gland transcripts in the adult female mosquito, *Aedes aegypti*. *BMC Genomics* **8**, 6 (2007).
- Cambier, S. et al. Gall wasp transcriptomes unravel potential effectors involved in molecular dialogues with oak and rose. *Front. Physiol.* **10**, 926 (2019).
- Alvarado, G. et al. Bioinformatic analysis suggests potential mechanisms underlying parasitoid venom evolution and function. *Genomics* **112**, 1096–1104 (2020).
- Gendreau, K. L. et al. House spider genome uncovers evolutionary shifts in the diversity and expression of black widow venom proteins associated with extreme toxicity. *BMC Genomics* **18**, 178 (2017).
- Madio, B., Undheim, E. A. B. & King, G. F. Revisiting venom of the sea anemone *Stichodactyla haddonii*: Omics techniques reveal the complete toxin arsenal of a well-studied sea anemone genus. *J. Proteom.* **166**, 83–92 (2017).
- Von Reumont, B. M., Undheim, E. A. B., Jauss, R.-T. & Jenner, R. A. Venomics of remiped crustaceans reveals novel peptide diversity and illuminates the venom's biological role. *Toxins* **9**, 234 (2017).
- Drukewitz, S. H., Bokelmann, L., Undheim, E. A. B. & von Reumont, B. M. Toxins from scratch? Diverse, multimodal gene origins in the predatory robber fly *Dasygogon diadema* indicate a dynamic venom evolution in dipteran insects. *GigaScience* **8**, 1–13 (2019).
- Walker, A. A. et al. Melt with this kiss: paralyzing and liquefying venom of the assassin bug *Pristhesancus plagipennis* (Hemiptera: Reduviidae). *Mol. Cell. Proteom.* **16**, 552–566 (2017).
- Özbek, R. et al. Proteo-transcriptomic characterization of the venom from the endoparasitoid wasp *Pimpla turionellae* with aspects on its biology and evolution. *Toxins* **11**, 721 (2019).
- Fingerhut, L. C. H. W. et al. Shotgun proteomics analysis of saliva and salivary gland tissue from the common octopus *Octopus vulgaris*. *J. Proteome Res.* **17**, 3866–3876 (2018).
- Undheim, E. A. B. et al. Clawing through evolution: toxin diversification and convergence in the ancient lineage Chilopoda (centipedes). *Mol. Biol. Evol.* **31**, 2124–2148 (2014).
- Jenner, R. A., Von Reumont, B. M., Campbell, L. I. & Undheim, E. A. B. Parallel evolution of complex centipede venoms revealed by comparative proteo-transcriptomic analyses. *Mol. Biol. Evol.* **36**, 2748–2763 (2019).
- Shear, W. A. & Edgecombe, G. D. The geological record and phylogeny of the Myriapoda. *Arthropod Struct. Dev.* **39**, 174–190 (2010).
- Ravenhall, M., Škunca, N., Lassalle, F. & Dessimo, C. Inferring horizontal gene transfer. *PLoS Comput. Biol.* **11**, e1004095 (2015).
- Chipman, A. D. et al. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. *PLoS Biol.* **12**, e1002005 (2014).
- Francois, C. M., Durand, F., Figuet, E. & Galtier, N. Prevalence and implications of contamination in public genomic resources: a case study of 43 reference arthropod assemblies. *G3* **10**, 721–730 (2020).
- Giribet, G. & Edgecombe, G. D. *The Invertebrate Tree of Life* (Princeton University Press 2020).
- Podobnik, M., Kisovec, M. & Anderluh, G., Molecular mechanism of pore formation by aerolysin-like proteins. *Phil. Trans. R. Soc. B* **372**, 20160209 (2017).
- Sher, D., Fishman, Y., Melamed-Book, N., ZHANG, M. & Zlotkin, E. Osmotically driven prey disintegration in the gastrovascular cavity of the green hydra by a pore-forming protein. *FASEB J.* **22**, 207–214 (2008).
- Sher, D. et al. Hydralysins, a new category of  $\beta$ -pore-forming toxins in Cnidaria. *J. Biol. Chem.* **280**, 22847–22855 (2005).
- Undheim, E. A. B., Fry, B. G. & King, G. F. Centipede venom: recent discoveries and current state of knowledge. *Toxins* **7**, 679–704 (2015).
- Ellsworth, S. A. et al. Convergent recruitment of adamalysin-like metalloproteases in the venom of the red bark centipede (*Scolopocryptops sexspinosus*). *Toxicon* **168**, 1–15 (2019).
- Smith, J. J. & Undheim, E. A. B. True lies: using proteomics to assess the accuracy of transcriptome-based venomomics in centipedes uncovers false positives and reveals startling intraspecific variation in *Scolopendra subspinipes*. *Toxins* **10**, 96 (2018).
- Rosenberg, J. & Hilken, G. Fine structural organization of the poison gland of *Lithobius forficatus* (Chilopoda, Lithobiomorpha). *Norw. J. Entomol.* **53**, 119–127 (2006).



36. Osman, G. E. H. et al. Bioinsecticide *Bacillus thuringiensis* a comprehensive review. *Egypt. J. Biol. Pest Co.* **25**, 271–288 (2015).
37. Palma, L., Muñoz, D., Berry, C., Murillo, J. & Caballero, P. *Bacillus thuringiensis* toxins: an overview of their biocidal activity. *Toxins* **6**, 3296–3325 (2014).
38. Schnepf, E. et al. *Bacillus thuringiensis* and its pesticidal crystal proteins. *Microbiol. Mol. Biol. Rev.* **62**, 775–806 (1998).
39. Peng, Q., Yu, Q. & Song, F. Expression of cry genes in *Bacillus thuringiensis* biotechnology. *Appl. Microbiol. Biot.* **103**, 1617–1626 (2019).
40. Baek, I., Lee, K., Goodfellow, M. & Chun, J. Comparative genomic and phylogenomic analyses clarify relationships within and between *Bacillus cereus* and *Bacillus thuringiensis*: proposal for the recognition of two *Bacillus thuringiensis* genomovars. *Front. Microbiol.* **10**, 1978 (2019).
41. Castillo-Esparza, J. F., Hernández-González, I. & Ibarra, J. E. Search for Cry proteins expressed by *Bacillus* spp. genomes, using hidden Markov model profiles. *3 Biotech* **9**, 13 (2019).
42. Costa Argôlo-Filho, R. & Lopes Loguercio, L. *Bacillus thuringiensis* is an environmental pathogen and host-specificity has developed as an adaptation to human-generated ecological niches. *Insects* **5**, 62–91 (2014).
43. Bravo, A. et al. Evolution of *Bacillus thuringiensis* Cry toxins insecticidal activity. *Microb. Biotechnol.* **6**, 17–26 (2013).
44. Wu, J.-W. et al. Adaptive evolution of cry genes in *Bacillus thuringiensis*: implications for their specificity determination. *Geno. Prot. Bioinfo.* **5**, 102–110 (2007).
45. Gully, N. et al. *Porphyromonas gingivalis* peptidylarginine deiminase, a key contributor in the pathogenesis of experimental periodontal disease and experimental arthritis. *PLoS ONE* **9**, e100838 (2014).
46. Montgomery, A. B. et al. Crystal structure of *Porphyromonas gingivalis* peptidylarginine deiminase: implications for autoimmunity in rheumatoid arthritis. *Ann. Rheum. Dis.* **75**, 1255–1261 (2015).
47. Yoshida, N. et al. Chaperonin turned insect toxin. *Nature* **411**, 44 (2001).
48. Brewer, M. S. & Bond, J. E. Ordinal-level phylogenomics of the arthropod class Diplopoda (millipedes) based on an analysis of 221 nuclear protein-coding loci generated using next-generation sequence analyses. *PLoS ONE* **8**, e79935 (2013).
49. Rehm, P., Meusemann, K., Borner, J., Misof, B. & Burmester, T. Phylogenetic position of Myriapoda revealed by 454 transcriptome sequencing. *Mol. Phylo. Evol.* **77**, 25–33 (2014).
50. Hickerson, C. M., Anthony, C. D. & Walton, B. M. Edge effects and intraguild predation in native and introduced centipedes: evidence from the field and from laboratory microcosms. *Oecologia* **146**, 110–119 (2005).
51. Stoev, P. et al. *Eupolybothrus cavernicolus* Komerički & Stoev sp. n. (Chilopoda: Lithobiomorpha: Lithobiidae): the first eukaryotic species description combining transcriptomic, DNA barcoding and micro-CT imaging data. *Biodivers. Data J.* **1**, e1013 (2013).
52. Fernández, R., Edgecombe, G. D. & Giribet, G. Exploring phylogenetic relationships within Myriapoda and the effects of matrix composition and occupancy on phylogenomic reconstruction. *Syst. Biol.* **65**, 871–889 (2016).
53. Giribet, G. & Edgecombe, G. E. Stable phylogenetic patterns in scutigermorph centipedes (Myriapoda: Chilopoda: Scutigermorpha): dating the diversification of an ancient lineage of terrestrial arthropods. *Invertebr. Syst.* **27**, 485–501 (2013).
54. Bereta, G. et al. Structure, function, and inhibition of a genomic/clinical variant of *Porphyromonas gingivalis* peptidylarginine deiminase. *Protein Sci.* **28**, 478–486 (2019).
55. Stobernack, T. et al. A secreted bacterial peptidylarginine deiminase can neutralize human innate immune defenses. *mBio* **9**, e01704–e01718 (2018).
56. Karkowska-Kuleta, J. et al. The activity of bacterial peptidylarginine deiminase is important during formation of dual-species biofilm by periodontal pathogen *Porphyromonas gingivalis* and opportunistic fungus *Candida albicans*. *Pathog. Dis.* **76**, 1–13 (2018).
57. Nystrom, G. S., Ward, M. J., Ellsworth, S. A. & Rokyta, D. R. Sex-based venom variation in the eastern bark centipede (*Hemiscolopendra marginata*). *Toxicon* **169**, 45–58 (2019).
58. Ward, M. J. & Rokyta, D. R. Venom-gland transcriptomics and venom proteomics of the giant Florida blue centipede, *Scolopendra viridis*. *Toxicon* **152**, 121–136 (2018).
59. Matari, N. H. & Blair, J. E. A multilocus timescale for oomycete evolution estimated under three distinct molecular clock models. *BMC Evol. Biol.* **14**, 101 (2014).
60. McCarthy, C. G. P. & Fitzpatrick, D. A. Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes. *mSphere* **2**, e00095–00017 (2017).
61. Thines, M. Phylogeny and evolution of plant pathogenic oomycetes—a global overview. *Eur. J. Plant Pathol.* **138**, 431–447 (2014).
62. McCarthy, C. G. P. & Fitzpatrick, D. A. Systematic search for evidence of interdomain horizontal gene transfer from prokaryotes to oomycete lineages. *mSphere* **1**, e00195–00116 (2016).
63. Savory, F., Leonard, G. & Richards, T. A. The role of horizontal gene transfer in the evolution of the Oomycetes. *PLoS Pathog.* **11**, e1004805 (2015).
64. Wybouw, N., Pauchet, Y., Heckel, D. G. & Van Leeuwen, T. Horizontal gene transfer contributes to the evolution of arthropod herbivory. *Genome Biol. Evol.* **8**, 1785–1801 (2016).
65. Rancurel, C., Legrand, L. & Danchin, E. G. J. Alienness: Rapid detection of candidate horizontal gene transfers across the Tree of Life. *Genes* **8**, 248 (2017).
66. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
67. Dash, T. S. et al. A centipede toxin family defines an ancient class of CSαβ defensins. *Structure* **27**, 315–326 (2019).
68. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
69. Undheim, E. A. B., Mobli, M. & King, G. F. Toxin structures as evolutionary tools: using conserved 3D folds to study the evolution of rapidly evolving peptides. *Bioessays* **38**, 539–548 (2016).
70. Quevillon, E. et al. InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
71. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
72. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
73. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
74. Han, M. V. & Zmasek, C. M. phyloXML: XML for evolutionary biology and comparative genomics. *Bioinformatics* **10**, 356 (2009).
75. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
76. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary analysis. <http://www.mesquiteproject.org> (2019).

## Acknowledgements

This work was supported by the Natural Environment Research Council (grant NE/I001530/1 to R.A.J.), the Australian Research Council (DECRA Fellowship DE160101142 to E.A.B.U., and the Discovery Grant DP160104025 to E.A.B.U. and R.A.J.), and the Norwegian Research Council (FRIPRO-YRT Fellowship no. 287462 to E.A.B.U.). R.A.J. is grateful to Luca Venturini and Nathan Kenny for providing bioinformatic assistance, and to Greg Edgecombe, Matt Clark, Pete Olson and Ana Riesgo for discussions. We thank Gonzalo Giribet and Ligia Benavides for supplying assembled transcriptomes for *Paralamyctes validus*, *Anopsobius giribeti*, *Scutigera weberi*, and *Sphendononema guil-dingii*. We would like to acknowledge the use of resources provided by UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway.

## Author contributions

E.A.B.U. and R.A.J. designed the research, performed the analyses, and wrote the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-21093-8>.

**Correspondence** and requests for materials should be addressed to E.A.B.U. or R.A.J.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021