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

Eivind A. B. Undheim1,2,3 & Ronald A. Jenner4

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.
Horizontal 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 walls17–21. 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 independently8, 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 ticks9, 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 not10. Similarly, glycoside family 19 chitinases found in the venom of chalcidoid parasitoid wasps were probably transferred from parasitic fungi11. Other potential cases of HGT contributing to insect venoms currently lack phylogenetic support12–14, while the direction of HGT of neurotoxic α-latrotoxins present in the venom of theridiid spiders and bacteria remains uncertain15. 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 homologues16–21, and these may include HGT candidates.

One venomous lineage that contains a large diversity of unassignable venom proteins,22,23 is centipedes (Chilopoda). Centipedes are one of the oldest terrestrial venomous lineages, with a fossil record going back 418 million years24. Living species belong to five orders: Scutigeromorpha (long-legged house centipedes), Lithobionomorpha (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 HGT25. 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: β-pore-forming toxin (β-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 genome26, we show that they are bona fide centipede genes rather than the result of contamination or symbionts. Importantly, a recent study27 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 β-PFTxs were recruited into the ancestral centipede venom proteome, with subsequent losses from craterostigmomorph and geophilomorph venom25. This gene family belongs to the bacterial aerolysin-like β-pore-forming toxin superfamily, which Moran et al.10 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 | Phylogenetic location of HGT | Phylogenetic location of recruitment into venom | Mapped to Strigamia maritima genome |
|------------|------------|-----------------------|-----------------------------|-----------------------------------------------|-----------------------------------|
| β-PFTx     | Bacteria   | 2 (1)                 | Arthropoda or Chilopoda; within Lithobiomorpha | Chilopoda                                     | SMAR004242, SMAR004243, SMAR012417 |
| centiPAD   | Bacteria   | 2                     | Within Scutigeromorpha; within Lithobiomorpha | Within Scutigeromorpha; within Lithobiomorpha | SMAR002991, SMAR002992, SMAR002993, SMAR008653 |
| DUF3472    | Bacteria   | 1 or 2 (1)            | In the stem of Pleurostigmophora or Amalpighiata; or in Epimorpha and within Lithobiomorpha | Within Sclopendromorpha                        | n/a                               |
| GEOTX02    | Fungi      | 1 or 2                | Geophilomorpha               | Geophilomorpha                                 | n/a                               |
| PCPDP-like | Bacteria   | 1                     | Lithobiomorpha               | Lithobiomorpha                                 | (group 1: SMAR012843, SMAR003678, SMAR004759); (group 2: SMAR012429, SMAR005429); (group 3: SMAR014279; (group 4: SMAR009615, SMAR004692, SMAR01285, SMAR007268, SMAR006394, SMAR009617, SMAR010233) |
| unchar05   | Fungi      | 2                     | Geophilomorpha, within Lithobiomorpha | Geophilomorpha                                 | n/a                               |
| unchar16   | Oomycetes  | 1                     | Unknown                      | Craterostigmomorpha                            | SMAR002275, SMAR004333, SMAR005016, SMAR002277, SMAR005613 |

n/a The absence of these genes from the genome of S. maritima is uninformative because the HGT events happened elsewhere in the tree.

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.

The identity of all paralogous loci is given. All are protein-coding loci with introns. Different paralog groups are indicated in parentheses.

Due to uncertainty about centipede phylogeny 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 hericapid lithobiomorphs.

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 β-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 β-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 β-PFTx. For instance, an early transfer of β-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. 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 β-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, β-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 streamlined of venom arsenal occurs alongside the recruitment and diversification of new components.

The β-PFTxs produced by bacteria are virulence factors that contribute to pathogenicity by the lysing of host cells. Interestingly, although they are not expressed in their tentacle venom, cnidarian β-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. There is no experimental data for the role of β-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. The great diversity of β-PFTx transcripts expressed in centipede venom proteomes, and the abundance of their expression, suggest that β-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. Homologous sequences are also present in transcriptomes of other centipedes from both lithobiomorph families (Lithobiidae: L. forficatus, E. caverniculus; 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. forficatus* found no evidence for bacterial symbionts in the venom producing and secreting tissues.

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. The most intensely studied bacterial PCPDPs are pore-forming insecticidal endotoxins.
GM crops36–39. They are produced by Bacillus species in the B. cereus group40,41, especially B. thuringiensis, the entomopathogenic bacterium from which they were first described, and which feeds on the insects killed by the toxin42. Cry toxins consist of three conserved domains: an N-terminal domain of α-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 β-sheets that are involved in receptor interactions, and which may confer host-specific toxicity37,43,44. 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 group37,41.

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 centipepe and other eukaryotic PCPDF-like sequences likewise only contain this N-terminal domain. A hint of how centipedes may have repurposed an insectical bacterial toxin into a venom protein is suggested by the most closely related bacterial sequences. All bacterial sequences that group together with the centipepe sequences in the clad at the top of Fig. 2 also only contain the pore-forming N-terminal domain, and like the centipepe sequences include a signal peptide region. This suggests that the bacterial proteins are exotoxins that are secreted from cells, like the centipepe PCPDF-like proteins. Unlike the centipepe sequences, the bacterial sequences in this clad 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 centipepe PCPDF-like sequences may derive from such putative bacterial exotoxins, followed by loss of these target-binding domains. Alternatively, the centipepe 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 centipepe 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 centipepe sequence because BLAST searches of the C-terminal region of the PCPDF-like sequences against centipepe 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 venom of two distantly related centipepe species, Therapinodora longicornis (order Scutigeromorpha), and Lithobius forficatus (order Lithobiomorpha)22,23. 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 unrelated45,46. Interestingly, the nesting of four fungal branches and a sequence derived from the black garden ant Lasius niger within the paraplytic 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 sibionts, which, if true, would be the second example of an animal venom component being produced by bacteria47. However, the balance of evidence suggests that centiPADs are bona fide centipepe gene family. CentiPAD is a prominent component of the venom proteome of T. longicornis22, 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 clad. Compared to related PAD sequences derived from the gammaproteobacterial genera Pseudomonas, Cecedea, 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 centipepe sequences.

The Lithobius centiPAD sequences group together in a strongly supported clad without interleaving bacterial sequences. This clad groups sequences from specimens collected in the UK, continental Europe, and North America33,48,49. This strongly suggests that they are bona fide centipepe sequences, a conclusion in line with the lack of evidence for microorganisms in the venom system of L. forficatus35. The European sequences (represented by UK sequences; an identical German sequence was excluded) form a sister clad to the American sequences. Because the latter were not determined to species by the original collectors48, it is unclear if they are L. forficatus, which was imported from Europe to North America some time before the end of the 19th century50.

CentiPAD is absent from the transcriptomes of other lithobiomorph species: Eupolybothrus cavernicola, Paralamyctes validus, and Anopsobius griberti51,52. 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 centipepe species analysed in our previous study23 (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 PCPDP-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 colored 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 Phylomic (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 scutigeromorph 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. Collombolan image was sourced from Phylomic (www.phylomic.org; credit for the Collombola 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 scutigeromorphs (Scutigera weberi, Sphenodonoma gaulingii, and Scutigera coleoptrata) 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 scutigeromorph families, while S. coleoptrata and T. longicornis belong to the family Scutigeridae. 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.35

Bacterial PAD converts peptidylarginine into citrulline residues, and the effects of this process have been most intensively 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.45,46,54,55 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.55,56 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 preserved 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.60

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 transcriptomes23,33,34,57,58. 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
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 β-PFTXs and PCPDF-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 venom 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 paraglogue groups of genes with introns in the genome of S. maritima, with the clade with 74% bootstrap support representing paraglogue groups 1–3 and the collapsed clade of eleven S. maritima sequences representing paraglogue 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 ascomycete 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-infesting 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 paragologue 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 Orchelisella 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. 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.

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 Craterostigmus tasmanianus. Unchar16 maps to two protein-coding paragologue 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, 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. 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. 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, 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 stockpiling venom arsenals from five or six to at least thirteen. We show that 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, 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. 16–21). 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. 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 β-PFTx, into the ancestral centipede venom may have conferred an immediate functional benefit, for example in prey immobilization. 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. The selective benefit of the horizontal transfer of β-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. The recruitment of...
β-PFTx into the ancestral centipede 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 known example of HGT contributing to the evolutionary origin of horizontally transferred genes. 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 a new host. A recent example is the use of the β-PFTx into the ancestral centipede venom.

**Methods**

**Initial identification of HGT candidates.** We used the transcriptomic and proteomic data from Undheim et al. and Jenner et al. 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, or few or no metazoan hits (β-PFTx, centiPAD, CHILOTX01, DUPF4327, GEOTX02, LTHTX01, LTHTX03, PCPDP- like, SCTOX01, SCTOX02, SLPTX02, SLPTX04, SLPTX06, SLPTX10, 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 cutoff of e-3. Significant hits against non-metazoan sequences were found for β-PFTx, centiPAD, DUPF4327, GEOTX02, PCPDP-like, SCTOX01, SCTOX02, 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 cutoff of e-3. Significant hits against non-metazoan sequences were found for β-PFTx, centiPAD, DUPF4327, GEOTX02, PCPDP-like, SCTOX01, SCTOX02, 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 cutoff of e-3. Significant hits against non-metazoan sequences were found for β-PFTx, centiPAD, DUPF4327, GEOTX02, PCPDP-like, SCTOX01, SCTOX02, 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 cutoff of e-3. Significant hits against non-metazoan sequences were found for β-PFTx, centiPAD, DUPF4327, GEOTX02, PCPDP-like, SCTOX01, SCTOX02, 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 cutoff of e-3. Significant hits against non-metazoan sequences were found for β-PFTx, centiPAD, DUPF4327, GEOTX02, PCPDP-like, SCTOX01, SCTOX02, 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 cutoff of e-3. Significant hits against non-metazoan sequences were found for β-PFTx, centiPAD, DUPF4327, GEOTX02, PCPDP-like, SCTOX01, SCTOX02, 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 cutoff of e-3. Significant hits against non-metazoan sequences were found for β-PFTx, centiPAD, DUPF4327, GEOTX02, PCPDP-like, SCTOX01, SCTOX02, SLPTX02, SLPTX04, SLPTX06, SLPTX30, unchar05, and unchar16.
Tree topology tests. Likelihood ratio tests of constrained tree topologies were performed in IQ-TREE 2.0.7, which implements several different tests. Each test compared 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) 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 trans- recordipteran insects. GigaScience 8, 1–13 (2019). 10. Walker, A. A. et al. Melt with this kiss: paralyzing and liquefying venom of the assassin bug Pristhesancus plagipes (Hemiptera: Reduviidae). Cell. Proteome. 16, 552–566 (2017). 11. Ozben, R. et al. Proteo-transcriptomic characterization of the venom from the endoparasitoid wasp Pimpina taurionellae with aspects on its biology and evolution. Toxins 11, 721 (2019). 12. 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. 18, 3866–3876 (2018). 13. Undheim, E. A. B. et al. Clawing through evolution: venom diversification and convergence in the ancient lineage Chilopoda (centipedes). Mol. Biol. Evol. 31, 2124–2148 (2014). 14. 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). 15. Shear, W. A. & Edgecombe, G. D. The geological record and phylogeny of the Myriapoda. Arthropod Struct. Dev. 39, 174–190 (2010). 16. Ravenhall, M., Škunca, N., Lassalle, F. & Dessimoz, C. Inferring horizontal gene transfer. PLoS Comput. Biol. 11, e1004095 (2015). 17. 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). 18. 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). 19. Gribet, G. & Edgecombe, G. D. The Invertebrate Tree of Life (Princeton University Press 2020). 20. Podobnik, M., Kiovec, M. & Anderl, G., Molecular mechanism of pore formation by aozylan-like proteins. Phil. Trans. R. Soc. B 372, 20160209 (2017). 21. 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). 22. Sher, D. et al. Hyaluronidase, a new category of β-pore-forming toxins in Caenida. J. Biol. Chem. 280, 22847–22855 (2005). 23. Undheim, E. A. B., Fry, B. G. & King, G. F. Centipede venom: recent discoveries and current state of knowledge. Toxins 7, 679–704 (2015). 24. Ellsworth, S. A. et al. Convergent recruitment of adamalysin-like metalloproteases in the venom of the red bark centipede (Scolopocryptops sexspinus). Toxicon 168, 1–15 (2019). 25. Smith, J. I. & Undheim, E. A. B. True lies: using proteomics to assess the accuracy of transcriptome-based venomics in centipedes uncovers false positives and reveals startling intraspecific variation in Scolopendrina subtrispines. Toxins 10, 96 (2018). 26. Rostgenberg, J. & Hilken, G. Fine structure organization of the poison gland of Lithobius forfacatus (Chilopoda, Lithobiomorpha). Norw. J. Entomol. 53, 119–127 (2006).
36. Osman, G. E. H. et al. Bioinsecticidal Bacillus thuringiensis: a comprehensive review. *Egypt J. Biol. Pest Co.* 25, 271–288 (2015).

37. Palma, L., Muñoz, D., Berry, C., Muñoz, J. & Caballero, P. Bacillus thuringiensis toxins: an overview of their biocidal activity. *Toxins* 6, 3296–3324 (2014).

38. Schnepp, 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. Biotechnol.* 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. *Biotech. 9*, 13 (2019).

42. Costa Argóñ-Filho, R. & Lopes Loguerio, 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. Guiry, 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 (2016).

47. Yoshida, N. et al. Chaperonin turned insect toxin. *Proc. Natl. Acad. Sci. U. S. A. 102*, 10499–10504 (2005).

48. Brewer, M. S. & Bond, J. E. Ordinal-level phylogenomics of the arthropod backbone reveals an ancient class of CS genes an ancient class of CS genes an ancient class of CS genes an ancient class of CS genes an ancient class of CS genes an ancient class of CS genes an ancient class of CS genes an ancient class of CS genes an ancient class of CS genes. *Science 333*, 119 (2010).

49. Rehm, P., Meusemann, K., Borner, J., Misof, B. & Burmester, T. Phylogenetic mapping of Paralamyctes validus and *Paralamyctes rufus*: evidence for the position of Myriapoda revealed by 454 transcriptome sequencing. *Syst. Biol.* 65, 871–889 (2016).

50. Quevillon, E. et al. InterProScan: protein domains identifier. *Nucleic Acids Res. 33*, W116–W120 (2005).

51. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Minh, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods 14*, 587–589 (2017).

52. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE 2: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Ann. Med. Biol. Exp. 32*, 268–274 (2015).

53. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary analysis. *http://www.mesquiteproject.org* (2019).
