nature communications

Article

https://doi.org/10.1038/s41467-023-35794-9

Micro and macroevolution of sea anemone venom phenotype

Received: 3 July 2022

Accepted: 3 January 2023

Published online: 16 January 2023

Check for updates

Edward G. Smith[1,5,7], Joachim M. Surm[2,7], Jason Macrander[1,3], Adi Simhi[2,4], Guy Amir[2,4], Maria Y. Sachkova[2,6], Magda Lewandowska[2], Adam M. Reitzel[1] & Yehu Moran[2]

Venom is a complex trait with substantial inter- and intraspecific variability resulting from strong selective pressures acting on the expression of many toxic proteins. However, understanding the processes underlying toxin expression dynamics that determine the venom phenotype remains unresolved. By interspecific comparisons we reveal that toxin expression in sea anemones evolves rapidly and that in each species different toxin family dictates the venom phenotype by massive gene duplication events. In-depth analysis of the sea anemone, *Nematostella vectensis*, revealed striking variation of the dominant toxin (*Nv1*) diploid copy number across populations (1-24 copies) resulting from independent expansion/contraction events, which generate distinct haplotypes. *Nv1* copy number correlates with expression at both the transcript and protein levels with one population having a near-complete loss of Nv1 production. Finally, we establish the dominant toxin hypothesis which incorporates observations in other venomous lineages that animals have convergently evolved a similar strategy in shaping their venom.

Understanding the molecular processes that drive phenotypic diversity among species, populations, and individuals is essential for unraveling the link between micro and macroevolution. Most traits are polygenic, meaning that their phenotype is influenced by multiple genomic loci[1–5]. However, understanding the heritability of these complex traits is challenging. Gene expression is likely an essential feature in determining the type of effect a gene has on a polygenic trait. This is evident with heritable gene expression dynamics contributing to phenotypic variations within and between species[6,7]. The mechanisms that drive these gene expression dynamics, which include mutations to the cis- and trans-regulatory elements[8,9], epigenetic modifications[7,10,11], and gene duplication[12–14], are subject to selective pressures that can result in adaptive traits in an organism.

Among the mechanisms capable of driving rapid shifts in gene expression dynamics is gene duplication, which can cause an increase in transcript abundance leading to phenotypic variations within and between species. Gene duplications, resulting in copy number variation (CNV), can originate from a combination of replication slippage, unequal crossing over during meiosis, retroposition of gene transcripts, and whole-genome duplications[15,16]. In addition to providing substrate for molecular evolution to act on via diversification, CNV arising from gene duplications can also cause immediate fitness effects resulting from increased gene expression through dosage[17]. Indeed, the potential for immediate phenotypic effects and the high mutation rates of duplicated genes suggest that CNV may be an important mechanism for rapid adaptation to new ecological niches. While CNV is studied mostly in the context of human genetic diseases and recent

[1]University of North Carolina at Charlotte, Department of Biological Sciences, Charlotte, NC, USA. [2]Department of Ecology, Evolution and Behavior, Alexander Silberman Institute of Life Sciences, Faculty of Science, The Hebrew University of Jerusalem, Jerusalem, Israel. [3]Florida Southern College, Biology Department, Lakeland, FL, USA. [4]The Hebrew University of Jerusalem, The School of Computer Science & Engineering, Jerusalem, Israel. [5]Present address: School of Life Sciences, University of Warwick, Coventry, United Kingdom. [6]Present address: Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway. [7]These authors contributed equally: Edward G. Smith, Joachim M. Surm. e-mail: ed.g.smith@warwick.ac.uk; joachim.surm@mail.huji.ac.il; yehu.moran@mail.huji.ac.il

adaptations[18–20], there is an increasing appreciation for the role of individual and population-scale CNV in ecological and evolutionary processes in other species and its impact on complex traits[21–23].

A complex trait hypothesized to be evolving under strong selective pressure is venom due to its essential ecological roles related to predation and defense[24–26]. The venom phenotype is often a complex trait because it relies on the coordinated expression of multiple toxin-coding genes. These toxins combine to produce venom profiles that are highly distinct, varying significantly within and between species[26,27]. Evidence supports that differences in toxin gene expression among species are a major contributor to the rapid evolution of venom phenotypes[28,29]. Toxin gene families have been hypothesized to evolve through a birth and death model of adaptive evolution as these proteins are central to an individual's fitness in mediating the interactions for both nutrition and survival[24,25]. Comparative genomics has revealed evidence supporting a hypothesis that a number of venomous organisms rapidly accumulate gene duplications in their genomes: examples include spiders[30,31], cone snails[32], and scorpions[33], although there are exceptions such as widow spiders[34].

Cnidarians represent an ancient venomous phylum where likely all species rely on toxins for prey capture and defense from predators[35]. Among cnidarians, sea anemone venom is arguably the most well-characterized[36] and past research has shown that toxin gene duplication is an important feature in these organisms[13,37,38]. Various sea anemone toxin families have been structurally and functionally validated or their expression localized to epithelial gland cells and specialized stinging cells called nematocytes[39,40]. These include pore-forming toxins such as Actinoporins[41,42], neurotoxins such as Nematocyte Expressed Protein 3 (NEP3[43,44]), sodium channel modulators (NaTx[45,46]), potassium channel toxins (KTx type 1, 2, 3, and 5 families[36,43,47–49]) and proteases such as NEP6 Astacins[44]. The characterization of these venom components has led to the investigation of their phylogenetic and evolution histories, revealing that these toxin families evolve under purifying selection[37,42,50], with the exception of KTx3 which has been shown to evolve under the influence of diversifying selection[50]. One of the most well-characterized cnidarian toxins is the *Nv1* family from the estuarine sea anemone, *Nematostella vectensis* Stephenson, 1935. Located in the ectodermal gland cells[40], this sodium channel toxin is the major component of the *N. vectensis* venom and has previously been shown to be encoded by at least 11 nearly identical genes that are clustered on one chromosome[13,51,52]. Furthermore, population-specific variants of *Nv1* absent from the reference genome assembly have been identified at specific locations across this species' geographic range along the Atlantic coast of the United States[13] and suggests the potential for location-specific alleles and the presence of unresolved intraspecific variability in the *Nv1* gene family.

Here, we investigate the evolution of venom in sea anemones at both macro- and microevolutionary scales. We employed a combination of comparative transcriptomics and modeling to understand the macroevolution of venom as a complex trait in sea anemones to reveal that toxin expression evolves rapidly among sea anemones with little constraint in their combinations. We find that in sea anemones, a single toxin family dominates their venom phenotype and can dynamically shift even between closely-related species or convergently evolve among distantly-related species. Phylogenomic analysis supports that the dominant toxin family undergoes massive gene duplication events. By investigating different populations of *N. vectensis* using a combination of transcriptomics, long-read genome sequencing, genomic qPCR, and proteomics, we further show that significant expansion and contractions events are driving dynamic shifts in the gene expression of the dominant toxin even at the microscale.

Results

Macroevolution of sea anemone venom phenotype

To investigate the macroevolution of venom as a complex trait, we employed comparative transcriptomics to quantify the gene expression of different toxin components and generate the venom expression phenotype among sea anemone species. Using publicly available transcriptomes, we identified single-copy orthologs to reconstruct the relatedness among sea anemones (Fig. 1A and Supplementary Data 1). In concert, we mapped the expression of multiple toxin families to each de novo assembled transcriptome. This included Actinoporin, NEP3 and NEP6, NaTx, and KTx1, 2, 3, and 5. Transcripts per million (TPM) values generated from the mapping were then used to reconstruct the venom expression phenotype for each species (Fig. 1A, pie graphs at tips). By performing ancestral state reconstruction (ASR) of the venom expression phenotype among sea anemones (Fig. 1A), we revealed that the NaTx toxin family was most likely the dominant toxin in the last common ancestors of sea anemones.

For most sea anemones (17 of 29), a single toxin family contributed to the majority of the venom expression phenotype and accounted for >50% of the total toxin expression (Supplementary Data 2). During diversification of Actinioidea, ASR suggests that KTx3 evolved to become the dominant toxin family. The KTx3 family is the dominant toxin family in 10 of the 17 Actinioidea species, with Actinoporin, KTx1, and KTx2 dominant in four, one and two species, respectively. Outside of Actinioidea, the Edwardsiid *Scolanthus callimorphus* Gosse, 1853 convergently evolved to have KTx3 as the dominant toxin. These shifts in the dominant toxin can be explained by a model of punctuated evolution[53,54]. We tested this by modeling the rates of evolution acting on the expression of toxins. We find evidence that all sea anemone venom components undergo dramatic and unique shifts that is best explained through a mode of rapid pulses (Pulsed) as opposed to Brownian motion (BM), Ornstein–Uhlenbeck (OU), or early burst (EB) models (Fig. 1C).

To understand the constraint acting on the toxin families themselves as well as the combinations of toxins they can form, we performed phylogenetic covariance analysis. Broadly, our analysis shows that sea anemones have minimal constraint acting on the combinations of toxins they employ to capture prey and defend against predators (Fig. 2A and Supplementary Data 3). While our results revealed that the venom expression phenotype of sea anemones has considerable flexibility in the combinations of toxins they express, there was an exception with NEP3 neurotoxin[43], and NEP6 protease families[44], which have a significant correlation in their expression. In concert, these two toxin families have the most pronounced phylogenetic signal in their expression among all toxins (Supplementary Data 4, with a strong signal having values close to 1), providing evidence that the expression for each toxin family is more similar among closely related species.

We then explored the venom expression phenotype of sea anemones by clustering the phylogenetic covariance of toxin expression using principal component analysis (PCA; Fig. 2B). This reconstructed the phylomorphospace of sea anemone venom, revealing that this complex trait has relatively low dimensionality (Supplementary Fig. 1), with two principal components accounting for the majority of variation (62%). While our analysis focused on transcriptomes generated from adults, RNA was generated from different tissue types with the majority coming from multiple tissue types. We therefore tested whether different tissues impacted this our results by using tissue type as a fixed effect in our PCOV analysis and found that this was not significant (Supplementary Data 5). Broadly, the venom expression phenotype clustered together depending on the toxin family with the highest expression, even among distantly-related species found in different superfamilies. While the expression of NEP3 and NEP6 show significant phylogenetic covariance, this had little impact on the broad clustering of the venom expression phenotype among sea anemones.

https://doi.org/10.1038/s41467-023-35794-9

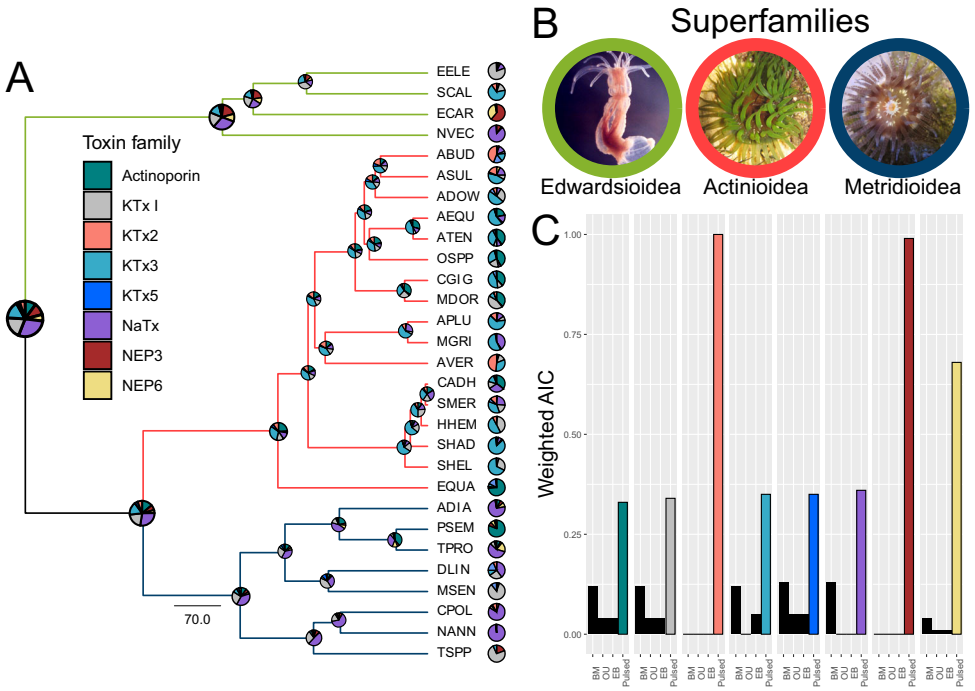

Fig. 1 | Ancestral reconstruction and phylogenetic constraints on the venom expression phenotype in Actiniaria. A Phylogenetic tree of sea anemones with pie charts nodes represents the ancestral reconstruction of known toxins and pie chart at tips represents the venom expression phenotype. All nodes had ultrafast bootstrap support >95% at nodes. **B** Representative images of the three sea anemone superfamilies included in the phylotranscriptomic analyses (Edwardsioidea—*N. vectensis*; Actinioidea—*Aulactinia veratra*; Metridioidea—*Calliactis polypus*). Actinioidea and Metridioidea photos courtesy of Peter Prentis. **C** Models of trait evolution fitted to toxin expression highlights that pulsed evolutionary process best describes sea anemone venom evolution. Model of best fit highlighted in color based on weighted AIC and are colored according to the toxin family key in panel **A**. See Supplementary Data 1 for species code and reference.

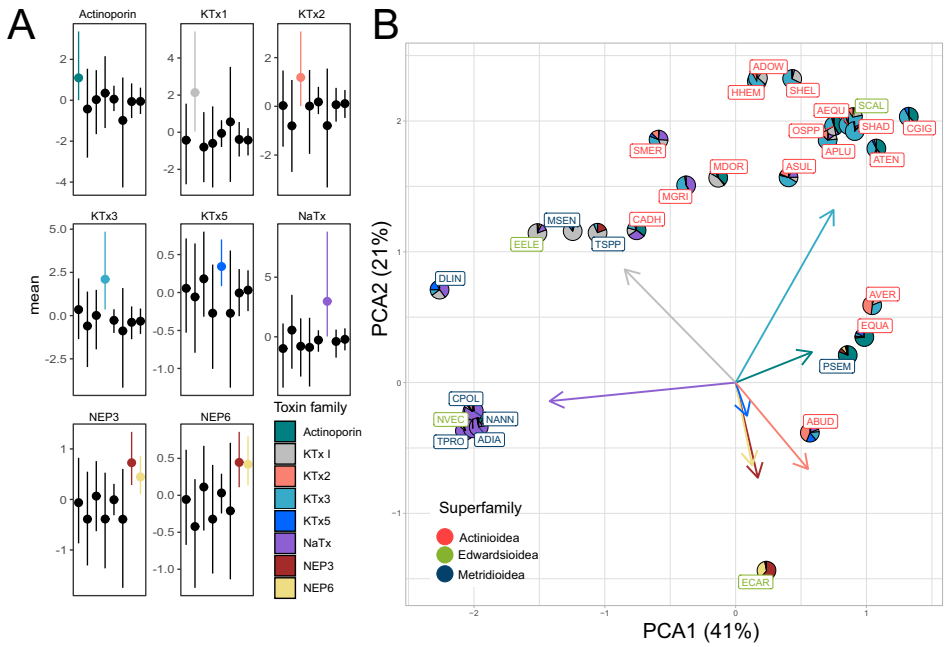

Fig. 2 | Sea anemone venom expression phenotype characterized by a single dominant toxin that evolves through major pulses. A Toxin combinations with significant phylogenetic covariance represented in color, with those not significant in black and lower and upper 95% confidence intervals represented as whiskers. Phylogenetic covariance analysis was run using Markov chain Monte Carlo for a total of 20 million iterations. **B** Sea anemone venom phylomorphospace cluster on the toxin family that contributes to the majority of expression. Loadings for each toxin family represented as arrows.

Furthermore, KTx3 and NaTx are the two toxin families that have the largest loadings, suggesting that they are the major families that define the venom profile employed by sea anemones. These findings suggest that in sea anemones a single dominant toxin family is the major driver in dictating the venom expression phenotype of each species.

Taken altogether, these results highlight that although the venom in sea anemones is comprised of many different toxins, we see that a single toxin family dominates the venom expression phenotype in sea anemones. This is supported by evidence that for most toxins, the phylogenetic signal acting on toxin expression is weak, the venom expression phenotype has low dimensionality, and that little constraint appears to be acting on the combinations of toxins expressed. Furthermore, the evolution of toxin expression appears to be highly dynamic, undergoing a process of rapid pulses. Strikingly, we see convergent shifts in the venom expression phenotype among distantly related species, likely the result of independently adapting the same dominant toxin family.

Genomic architecture of the dominant toxin family

While our comparative transcriptomics and phylogenetic covariance analysis revealed that the sea anemone venom expression phenotype is largely dictated by a single toxin family, the genetic architecture that underlies the dominant toxin family requires investigation at the genomic level. We investigated the sea anemone genomes recently assembled using long-read sequencing for three species, *Actinia equina* (Linnaeus, 1758)[55], *S. callimorphus*, and *N. vectensis*, from two superfamilies (Actinioidea and Edwardsioidea). Remarkably, we find evidence that massive duplication events underly the signal driving a toxin family to become dominant (Fig. 3A), with all three sea anemone species possessing more than 15 copies of each of their respective dominant toxin gene family. Our phylogenetic covariance analysis and comparative transcriptomics revealed that in *S. callimorphus* and *A. equina*, the dominant toxin is KTx3, whereas, in *N. vectensis*, the dominant toxin is NaTx. For each species, the dominant toxin family accounts for the highest number of copies among all other toxin families (Supplementary Data 6). While *A. equina* contains both NaTx and KTx3, the KTx3 toxin family underwent a much greater series of duplication events, with eight members from the NaTx family and 52 members from the KTx3 family.

Next, we aimed to unravel the evolutionary steps that led to the amplification of the dominant toxin family in sea anemones by investigating the genomic location and macrosystemic relationship of chromosomes/scaffolds. To do this we performed phylogenomic analyses and discovered that macrosynteny is broadly shared among the three species (Fig. 3B), which confirms that the macrosyntenic relationship of chromosomes between *N. vectensis* and *S. callimorphus* is consistent with previous analyses[56]. This is particularly evident between *N. vectensis* and *S. callimorphus* whose assemblies utilized long-read sequencing and high-throughput chromosome conformation capture to generate chromosome-level genome assemblies, whereas *A. equina* genome was generated from only long-read sequencing. From our analysis, we find 15 chromosomes are linked between *N. vectensis* and *S. callimorphus*, and that these are linked to 108 scaffolds found in *A. equina* (Fig. 3B and Supplementary Data 7). Our analysis further reveals that while macrosynteny is largely conserved among the three species, synteny among toxin loci for the KTx3 family in *S. callimorphus* and *A. equina*, or the NaTx family in *N. vectensis* and *A. equina*, is absent. This was further confirmed by exploring the genes and genomic sequence up and downstream of each toxin loci. In contrast, the NEP3 and NEP6 gene families can be seen to lie on scaffolds that share macrosynteny among the three genomes (Supplementary Data 8). This supports that the evolution of genes encoding some toxin families are highly dynamic including the NaTx and KTx3 families which have become dominant in *N. vectensis,* and *S. callimorphus* and *A. equina*. Interestingly, while these two toxins are distinct from each other (Fig. 3C), evident from the CLANS clustering, they likely share a common evolutionary history, which is supported by evidence that they share the same cysteine framework (Fig. 3D) and some KTx3 toxins having similar activity to NaTx toxins[57–60]. Because of this likely shared evolutionary history, we also explored whether any synteny was shared between the NaTx and KTx3 families to test a hypothesis for an ancestral NaTX/KTx3, however, no macro or microsynteny was found. This further suggests that these toxin families undergo rapid evolution in their genomic architecture compared to other genes and even other toxin genes.

We further explored the molecular evolution of the dominant toxin family within each species to gain insight into the modes of gene duplication that might be shared among species. In *N. vectensis*, 14 of a total of 18 NaTx copies share 99% sequence similarity at the mRNA level and were hypothesized to evolve through tandem duplication and possibly concerted evolution to result in the *Nv1* cluster[13]. In *A. equina*, four NaTx copies are found on a single cluster, with another four located throughout the genome, yet still they share an average of 87% similarity at the mRNA level. In *S. callimorphus* and *A. equina*, KTx3 copies frequently also cluster together in tandem, however, they also underwent repeated translocation events. They also do not display the same degree of gene homogenization observed for the *Nv1* cluster or NaTx copies in *A. equina*, with *S. callimorphus* and *A. equina* KTx3 copies sharing 73% and 34% similarity at the mRNA level, respectively (Supplementary Data 9). These results support that the amplification of the KTx3 gene family is likely occurring through lineage-specific duplications, and that tandem duplication events play a major role for both NaTx and KTx3 families.

Overall, our comparative transcriptomics and phylogenomic analysis have provided striking insights into the macroevolution of venom in sea anemones. From these analyses, we see that a dominant toxin family dictates the venom expression phenotypes in sea anemones and that this evolves in a highly dynamic process through rapid pulses that are driven by gene duplication events. However, it is unclear how these patterns occur at the population and individual scale and understanding this link would provide important insights into the microevolution of venom in sea anemones.

Population dynamics of the venom phenotype in *N. vectensis*

Previous work has revealed that the *N. vectensis* NaTx cluster of genes are overall highly similar but also that population-specific variants of *Nv1* exist[13]. This led us to explore the population dynamics of the *Nv1* cluster in *N. vectensis* by performing comparative transcriptomics, quantitative genomic copy number PCR, proteomics, and genomics using long-read sequencing.

To explore the microevolution of the venom phenotype in *N. vectensis*, we first needed to understand its population structure across the native geographical range along the Atlantic coast of North America. To do this, highly complete transcriptomes were generated from nine *N. vectensis* populations originating from locations on the Atlantic coast of North America (Fig. 4A). Specifically, all transcriptomes had a BUSCO score >90%, except for Massachusetts (Supplementary Data 10, BUSCO = 72.2%). In all, 2589 single-copy orthologs were identified using OrthoFinder and used to generate a well-supported maximum-likelihood phylogenetic tree (Fig. 4B). Broadly, populations clustered according to geographical location, with populations from North (Massachusetts, Maine, New Hampshire, New Jersey, and Nova Scotia) and South (North Carolina, South Carolina, and Florida) clustering independently together. This phylogenetic analysis also supports that the Maryland population, which serves as the source for the most common *N. vectensis* lab strain[61], clusters more closely with southern populations, consistent with the previous analyses[62]. Differences among populations from close geographical locations are also observed, specifically with South Carolina populations clustering more closely with Florida than North Carolina.

https://doi.org/10.1038/s41467-023-35794-9

A

Species	NaTx Copies	NaTx Clusters	NaTx Scaffolds	KTx3 Copies	KTx3 Clusters	KTx3 Scaffolds
Actinia	8	4	3	52	37	29
Nematostella	18	5	4	0	0	0
Scolanthus	0	0	0	18	10	5

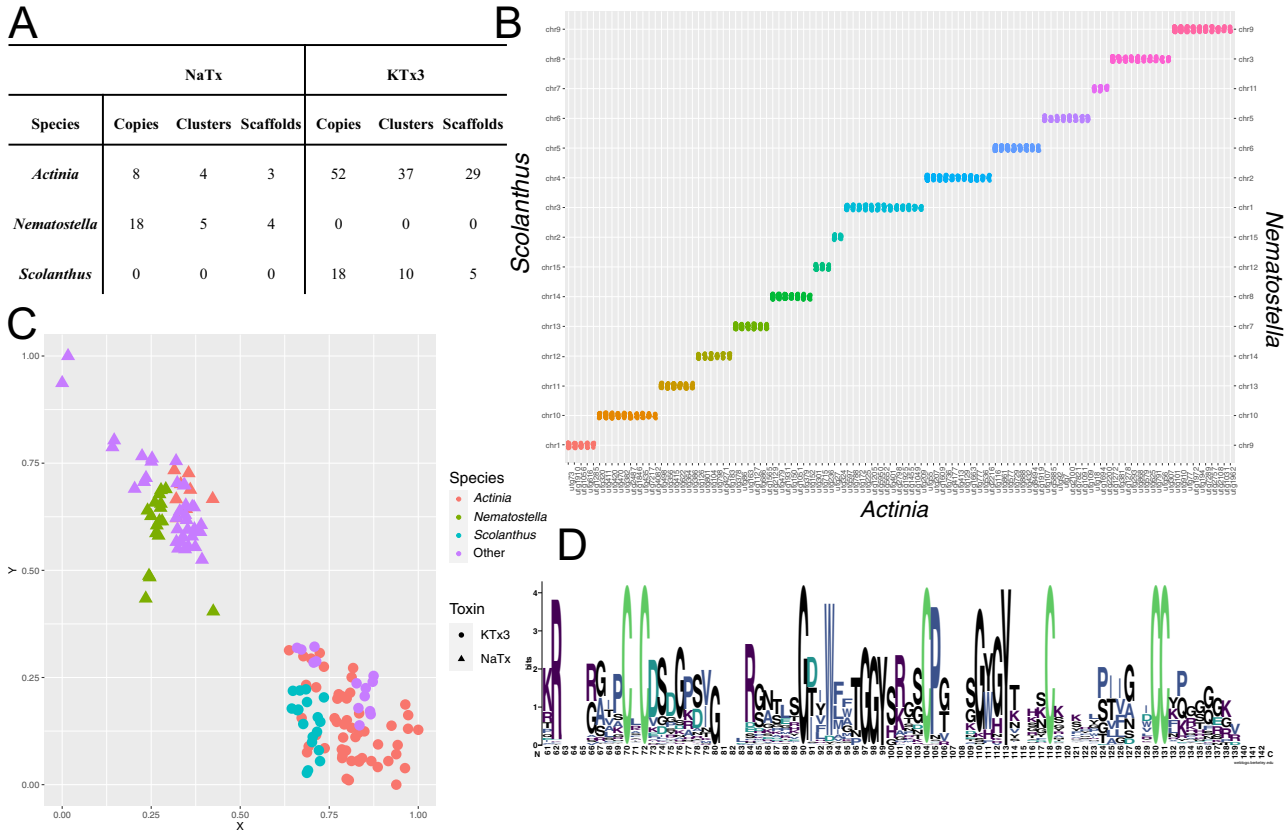

Fig. 3 | Phylogenomic analysis of the dominant toxin family in *A. equina*, *N. vectensis*, and *S. callimorphus*. **A** Table representing the copy number of toxins found across genomic scaffolds assembled. **B** Oxford plot representing the macrosyntenic relationships of chromosomes among *A. equina*, *N. vectensis*, and *S. callimorphus*. **C** Pairwise similarity based clustering of NaTx and KTx3 found in the three genomes as well as other sea anemone species by Cluster Analysis of Sequences (CLANS) software[113]. Sequences from *A. equina*, *N. vectensis*, and *S. callimorphus* represented by different colors. Other includes toxin used from previous work[50]. The KTx3 and NaTx families are represented as circles and triangles, respectively. **D** Shared evolutionary history of NaTx and KTx3 highlighted by a conserved cysteine framework identified in the mature peptide.

After reconstructing the population structure of *N. vectensis*, we then aimed to explore the venom phenotype among populations. We focused on the evolution of the NaTx gene family, which is the dominant toxin family in the model representative *N. vectensis*. Our comparative transcriptomic approach identified allelic variation for the members of the NaTx gene family (Fig. 4A). Specifically, we were able to capture variants of *Nv1* from all populations with more than eight variants captured in all populations, except for Florida samples in which we were only able to capture a single variant. A possible explanation for only a single variant being captured in Florida is this copy is highly conserved and still maintained in high copy numbers. Investigating the expression patterns for *Nv1* among all populations, however, revealed that *Nv1* has massively reduced expression in Florida with TPM for *Nv1* in all populations >500, while Florida had a TPM of five (Supplementary Data 11A). Expression differences of *Nv1* among populations were further validated using nCounter platform, revealing that indeed *Nv1* gene expression is massively reduced in the Florida population (Supplementary Data 11B). This striking result suggests that the *Nv1* cluster in Florida has undergone a massive contraction.

While we were able to get a representation of the sequence diversity of *Nv1* among populations, capturing the copy number variation of *Nv1* is beyond our capacity using comparative transcriptomics. This is especially significant for the *Nv1* family which can contain identical gene copies within the loci. Therefore, we performed individual quantitative PCR estimates of *Nv1* diploid copy number, (Fig. 4C and Supplementary Data 12) for five populations (North Carolina, Maine, Massachusetts, New Hampshire, and Nova Scotia). From this we found that *Nv1* copy number ranges from 8 to 24 genomic copies for different populations on the Atlantic coast of North America. We observed significant differences in the mean copy number across populations (ANOVA, $P < 2e\text{-}16$; Supplementary Data 13), with pairwise post hoc tests revealing significant differences among all population pairs (Tukey HSD, $P < 0.05$; Supplementary Data 14), except the Maine-New Hampshire comparison. The mean population copy number was lowest in Maine and New Hampshire (11 copies) and highest in North Carolina (20 copies).

While genomic and transcriptomic measurements can provide the copy number and expression level of a gene, respectively, the biology of a trait heavily depends on the synthesis level of the protein product of a gene. Moreover, in some cases protein levels are not in direct correlation to RNA levels, and proteomic and transcriptomic dataset might give contrasting pictures[63–65]. Thus, we tested the notion that Florida Nv1 protein levels are massively reduced using a proteomics approach, comparing samples from Florida with North Carolina, the closest population to Florida where we had genomic data. This analysis revealed that Nv1 in Florida is at either negligible or undetectable levels, both when using iBAQ and label-free quantification (LFQ) values. In contrast, Nv1 in North Carolina was measured as the third most abundant protein in the whole proteome (Supplementary Data 15), resulting in Nv1 being the most significantly differentially abundant protein between the two populations (Fig. 4D and Supplementary Data 16). This striking difference cannot be explained by a technical limitation in measuring the Florida samples as overall iBAQ and LFQ values were similar for most proteins in the two populations, and the two proteomes significantly correlated ($R^2 = 0.98$; Supplementary Data 17 and Supplementary Fig. 2). Thus, we see a clear

https://doi.org/10.1038/s41467-023-35794-9

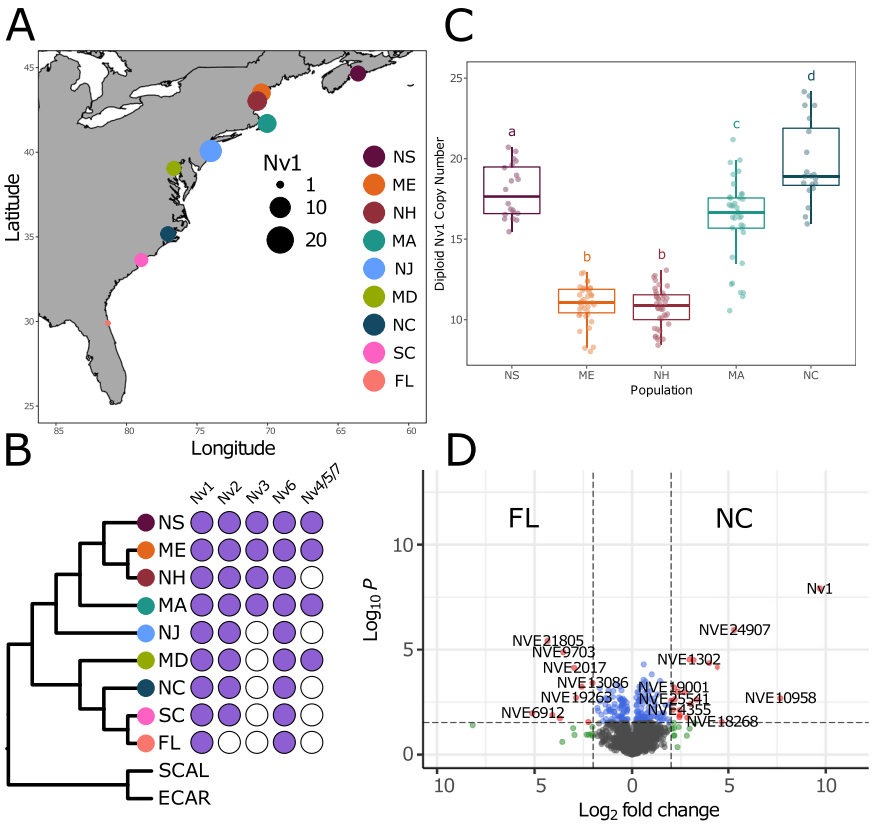

Fig. 4 | Diversity of NaTx paralogs among *N. vectensis* populations. A Map showing the location of the sampled populations across North America. Allele diversity represented by size of dot plot at different locations. Florida (FL), Massachusetts (MA), Maryland (MD), Maine (ME), North Carolina (NC), New Hampshire (NH), New Jersey (NJ), Nova Scotia (NS), and South Carolina (SC). **B** Population structure of *N. vectensis* generated using a maximum-likelihood tree from protein sequences and presence/absence of previously characterized NaTx paralogs in *N. vectensis*. **C** Boxplots of diploid copy number estimated using qPCR from samples collected from the five populations. The median is represented by the bold horizontal line and the upper and lower hinges represent the 75th and 25th percentiles, respectively. The boxplot whiskers extend to the largest and smallest values within 1.5 * IQR (interquartile range) for the upper and lower whiskers, respectively. All individual copy number estimates are shown for each boxplot. ANOVA revealed a significant effect of population on diploid copy number (one-tailed test, $P < 2e\text{-}16$; Supplementary Data 13). The letters above boxplots indicate the results of a Tukey–Kramer post hoc test controlling for the family-wise error rate ($\alpha = 0.05$), with all population comparisons showing significant copy number differences ($P < 0.05$; see Supplementary Data 14 for pairwise comparisons) unless they share the same letter. **D** Volcano plot representing proteins of significantly different abundance, measured as label-free quantification (LFQ) intensity, between FL and NC. Gray dots represent proteins that are not significant, blue dots are proteins with significant P value < 0.01, green dots are proteins with Log2 fold change of >2, red dots are proteins with significant P value and Log2 fold change.

correlation between the reduction in allelic variation found in Florida samples, the reduction in *Nv1* copies, and exceptional reduction of Nv1 at the protein level in the Florida population.

A slight variation in the members of the NaTx gene family were also captured in the transcriptomes generated from different populations. While numbers did vary, all transcriptomes captured at least a single variant of *Nv6*, with some as many as three (Fig. 4B). *Nv2* was captured in all transcriptomes, except for Florida. Notably, *Nv3*, a previously identified variant that contains a 6-bp deletion altering the N-terminus of the mature peptide[13], is located within the *Nv1* locus unlike other more distinct variants (e.g., *Nv4* and *Nv5*) that have translocated outside the *Nv1* locus[66]. Although *Nv3* is not widely identified in our amplicon analyses, this may be influenced by the presence of mutations in the primer binding sites revealed by our genomic analyses. This is likely the case as our comparative transcriptomics was able to recover *Nv3* copies in all North populations (including Massachusetts, Maine, New Hampshire, and Nova Scotia) and absent in all other populations. No copies of *Nv8* were captured, which is also consistent with previous finding that this member of the NaTx gene family is expressed at very low levels. The distribution of *Nv4*, *Nv5*, and *Nv7* is patchy, which may be explained by their expression being restricted to early life stages and maternally deposited in the egg and hence only captured if individuals sampled were females containing egg packages[66].

Here, we provide multiple lines of evidence that confirms that the copy number of *Nv1*, a member of the NaTx family that is the dominant toxin in *N. vectensis*, evolves in a highly dynamic manner among populations, while other toxin families appear to be much more stable. This is most striking in the Florida population that has undergone a dramatic contraction of the *Nv1* cluster, resulting in the almost total loss of *Nv1* at the mRNA and protein level. This highlights that even within the dominant toxin family (NaTx) a hierarchy exists in which specific members (e.g., the *Nv1* cluster) are the major modifiers of the venom phenotype and that their evolution is highly dynamic. As such, understanding the genomic architecture for the expansions and contractions of the *Nv1* cluster among *N. vectensis* populations is critical to identify the mechanisms that underly the evolution of a dominant toxin family and its role in driving variations in the venom phenotype within species.

Genomic arrangement of *Nv1* loci

To further explore the sequence diversity of *Nv1* among populations of *N. vectensis*, we performed amplicon sequencing of 156 *N. vectensis* individuals from five locations (North Carolina, Massachusetts, New Hampshire, Maine, and Nova Scotia; Fig. 5A). This analysis revealed the presence of 30 distinct *Nv1* variants. Of these 30 distinct *Nv1* variants,

https://doi.org/10.1038/s41467-023-35794-9

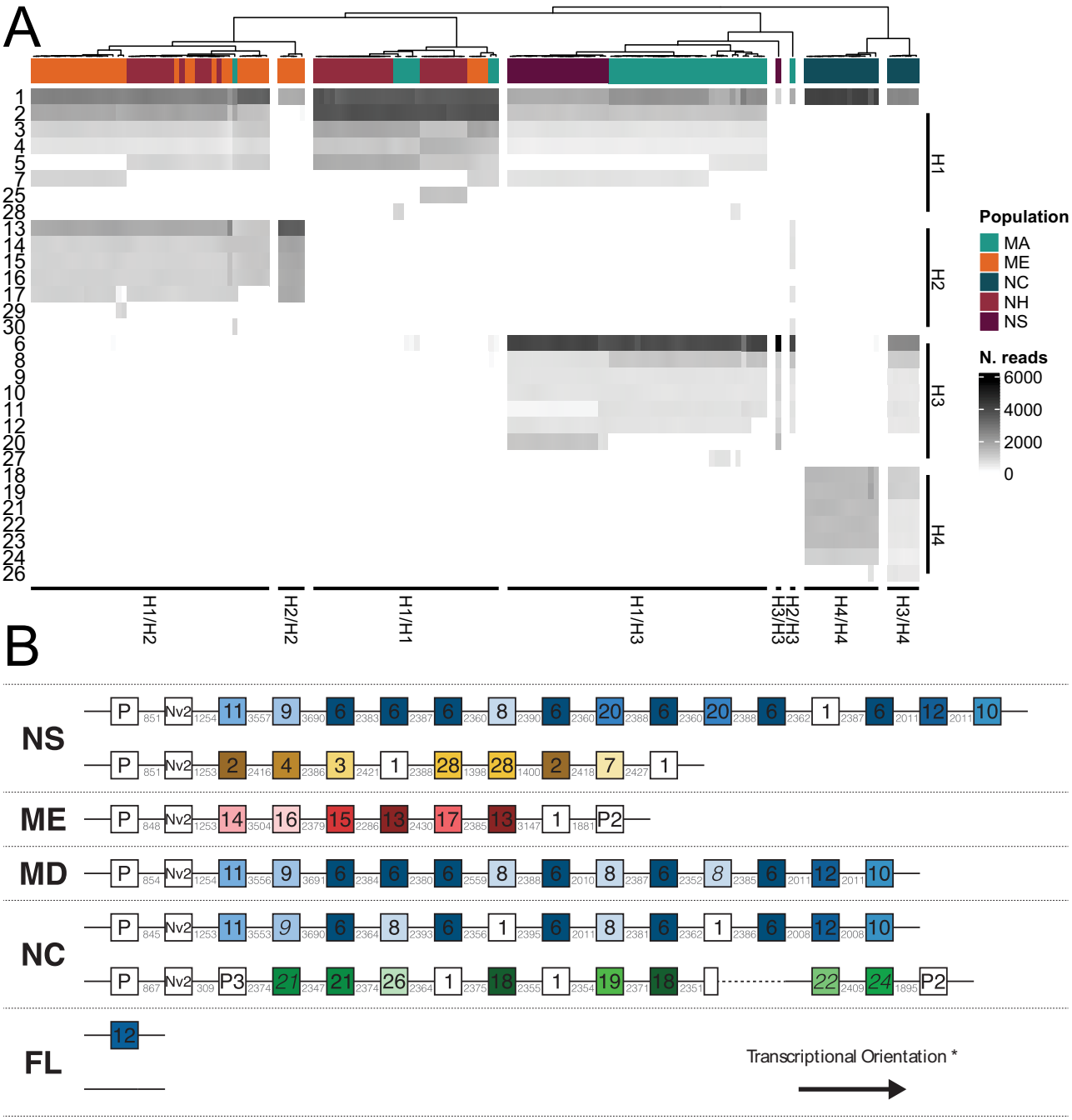

Fig. 5 | Genomic arrangement of *Nv1* loci. A Heatmap of *Nv1* paralog abundance in *N. vectensis* samples. Samples are clustered along the *x* axis according to *Nv1* composition and labeled according to the sampled population. *Nv1* paralogs are grouped on the *y* axis according to the inferred haplotype (H1–4). **B** Genomic arrangement of the *Nv1* locus across individuals Each box denotes an *Nv1* paralog with the internal label corresponding to the numeric identifier of the associated amplicon variant. Paralogs are colored according to the core haplotypes identified in the amplicon analyses (Yellow = H1, Red = H2, Blue = H3, Green = H4, White = shared/undetermined) and corresponding multiple sequence alignment and amino acid variants are shown in Supplementary Figs. 3 and 4. Italicized variant labels correspond to the amplicon variant with the lowest pairwise distance to the genomic variant. Pseudogenes (P1, P2, and P3) and *Nv2* copies are labeled separately. Intergenic distances are shown in gray and dashed line indicates a gap of unknown distance. *All loci are shown in the same transcriptional orientation and all variants within each locus share the same orientation.

27 were also found in the transcriptomes of *N. vectensis* from different populations. At the coding sequence level 11 variants are found broadly across multiple different populations, 12 are restricted to the Northeast, and four are found in both North Carolina and New Jersey. Hierarchical clustering of these amplicon-derived *Nv1* paralogs at the DNA level revealed groups of samples that share *Nv1* locus genotypes, from which four "core" haplotypes can be deduced from homozygous individuals (Fig. 5B). While *Nv1.var1* is shared across all haplotypes and confirmed to be present in all transcriptomes (except for Florida), the remainder of the paralogs are exclusive to a single haplotype. The core haplotypes shared multiple haplotype-specific paralogs although there is evidence of some variability within these haplotypes. For example, heterozygous individuals with an H1 haplotype possess either *Nv1.var5* or *Nv1.var7*.

The distribution of the core haplotypes varies across the range of *N. vectensis*. While H1 is present in 87–100% of individuals from Nova

Scotia, Maine, New Hampshire, and Massachusetts, it is absent from the North Carolina samples (Fig. 5A and Supplementary Fig. 5). Conversely, the H4 haplotype is only present in all samples from North Carolina but absent from all other populations.

Given the significance of *Nv1* cluster to the venom phenotype in *N. vectensis* and its dynamic evolution among populations, we performed long-read sequencing and assembly of *Nv1* locus haplotypes for four individuals from different populations (Florida, North Carolina, Maine, and Nova Scotia). Our analysis yielded seven distinct haplotypes, in addition to the haplotype of the recently assembled *N. vectensis* reference genome[56]. Of the seven haplotypes assembled in this study, six were assembled completely and spanned by single reads (Supplementary Data 18). The number of *Nv1* copies per haplotype (including pseudogenes) ranged from 0 to 17 copies. These copy number estimates are higher than the amplification-based analyses due to the presence of mutations in the primer binding sites of some variants (e.g., *Nv1.var28*). The Floridian haplotype without a single *Nv1* copy results from a 30 kb deletion relative to the single-copy haplotype (Supplementary Fig. 6). With the exception of the Florida haplotypes, all haplotypes share a pseudogene and a copy of one paralog (*Nv2*). The Maine and North Carolina H4 haplotypes also share a pseudogene at the opposite end of the locus.

The composition of the assembled haplotypes corroborates the inferred core haplotypes identified by the amplicon analyses. Of the eight haplotypes (including the ref. [56]) four belong to the H3 haplotype, yet, show extensive variation in the organization of *Nv1* paralogs. Considering the arrangement of paralogs and the associated intergenic spacing, expansion of Nova Scotia H3 (blue; Fig. 5B) occurred through serial duplication of single *Nv1.var6* copies, and of paired *Nv1.var6-Nv1var.20* copies (*Nv1.var1* and *Nv1.var20* differ by 1 bp intronic indel/mutation). In contrast, North Carolina H3 (blue; Fig. 5B) appears to have undergone a duplication of a *Nv1.var8-Nv1.var6-Nv1.var1-Nv1.var6* quadruplet.

There is evidence that transposable elements (TEs) have impacted the *Nv1* locus as there is a large insertion into the locus in FL262 with Mutator-like elements (MULEs) at either end (positions 60,895 and 76,278), suggestive of a pack-MULE (Supplementary Data 19). The insertion is found in multiple genomic loci and is ~15 kb in length, which is at the extreme end of pack-MULE size distribution seen in plants[67]. Nevertheless, while the TEs can explain the insertion, they do not appear to explain the deletion of the rest of the *Nv1* locus. One plausible explanation for the deletions associated with the Florida haplotypes is the presence of non-b DNA structures. The deletion breakpoints for zero-copy haplotype occurs within 356 bp from a breakpoint associated with the truncated *Nv1* found in the Maine and North Carolina haplotypes, suggesting this region may be predisposed to deletions (Supplementary Fig. 7). The breakpoints in this region occur at or adjacent to features known to cause non-canonical (non-b) DNA structures: inverted repeats and poly(G). These structures induce genomic instability and have been associated with large deletions in humans and yeast[68].

Discussion

Our findings provide a striking example of how gene duplication can impact both micro and macroevolutionary patterns in shifting the venom expression phenotype within and between species of sea anemones. In contrast to the signature of evolutionary constraint acting on toxin genes at the sequence level[37,42,50], here we demonstrate that toxin gene expression evolves rapidly and dynamically, suggesting strong selective forces are acting on toxin gene expression. Phylogenomic analysis supports that gene duplication is likely the underling mechanism that accounts for these adaptive shifts in gene expression. We find similar patterns at the populations scale and show that the dominant toxin in the model sea anemone, *N. vectensis*, exhibits extreme copy number variation between populations and even individual chromosomes.

Here, we investigated venom evolution in sea anemones across both ecological and evolutionary timescales. Broadly, our results reveal that the expression of a single toxin family dominates the venom expression phenotype of sea anemones and that they can shift between, and even within a species in a highly dynamic manner. Modeling the expression of toxin families confirms that their evolution is best explained by rapid pulses of evolution. Convergent shifts in the dominant toxin are observed to occur, with the venom expression phenotype of species found across superfamilies clustering together. Convergent evolution is often a hallmark of adaptive evolution, thus indicating that the dominant toxin is having an adaptative role required for ecological specialization.

Our findings also show a similar mechanism to venomous snakes[28,53]. In work by Barua & Mikheyev[28], the snake venom phenotype is largely dictated by a single dominant toxin, which explains its low dimensionality and lack of phylogenetic constraint acting on the venom combinations. By comparing our result with those found in snakes we see that selection driving toxin families to become dominant, rather than intrinsic constraints, likely plays the major role in shaping the venom phenotype for both sea anemones and snakes. A similar pattern was reported in cone snails, in which a single toxin superfamily often accounts for >50% of the total conotoxin expression[69]. These dominant toxin superfamilies convergently evolve in a highly dynamic manner, where closely-related species have different dominant toxins[69]. From these results, we suggest that given venom is a polygenic trait in many other venomous animals, a single dominant toxin family is the major dictator of the venom phenotype and the shift in the dominant toxin is likely driven by selection to meet the ecological requirements of these animals.

Our findings provide evidence that a single toxin family dictates the phenotype of venom in sea anemones, while other venom components likely have a more indirect effect. A potential constraint of this phylotranscriptomic approach is the assumption that toxin transcript abundances accurately represent the venom phenotype. The correlation between transcript abundances and protein/peptide expression has been the subject of debate within the venom field[70–73], and more widely (e.g., refs. [74,75]). Nevertheless, we consider our transcriptomic approach robust for the following three reasons: First, a previous quantitative interspecies study did not find evidence of protein-level buffering in venoms that could complicate interspecific comparisons[73]. Second, to avoid known issues with false positives in transcriptomic analyses, we applied stringent filters to restrict our analyses to bona fide sea anemone toxins. Lastly, our work with *N. vectensis* (ref. [43]; this study) has shown strong congruence between toxin transcript and protein/peptide abundance.

From the transcriptomic observations, we see similarities with the omnigenic model which is a framework to understand the polygenic architecture of complex traits by categorizing groups of complex trait genes as either core or peripheral genes. Proposed by Boyle et al.[1], the value of a given trait is largely determined by the expression level of a few core genes in the relevant tissue, while genes co-expressed likely have a more indirect effect on the phenotype. We see a correlation between the omnigenic model and the venom expression phenotype in which a single dominant toxin family act as core genes that directly affect the venom expression phenotype. Other toxin genes, however, act more like peripheral genes, affecting the venom expression phenotype in a more indirect manner and could possibly be acting synergistically with the dominant toxin. This has previously been shown in various spitting snakes where phospholipase A2 (PLA2) potentiates the dominant toxin, cytotoxic three-finger toxins which accounts for the majority of the protein abundance in their venom profile[76]. Therefore, while all toxin components of the venom may contribute to the heritable variance of the complex trait, the core genes are the major

modifiers of the venom phenotype. It should be noted, however, that the omnigenic model does not perfectly fit venom as a complex trait. Venom is a relatively unique trait in the sense that the toxic cocktail used by the organism is a terminal component of venom production and that toxins are unlikely to impact this complex pathway through feedback loops, while the omnigenic model was conceptualized to understand how networks of genes impact a complex trait. To apply this to the venom phenotype, it would require exploration into the network involved in venom production. To unravel this in sea anemones, comparative transcriptomics of nematocytes and gland cells would be needed. However, applying the omnigenic model to understand the phenotype of the venom cocktail itself still gives us insights into understanding the complex trait by categorizing different toxin families into groups such as core and peripheral genes.

Recent works are unraveling the impact of gene expression on the fitness of an organism. For example, variation at the nucleotide level driving changes in gene expression was shown to be the major modifier of the fitness landscape of protein-coding genes in an experimental setup using the model yeast *Saccharomyces cerevisiae*[77]. These findings indicate that there is greater constraint acting on the sequence of highly expressed genes and highlights that gene expression levels and sequence evolution are interrelated[77]. A recent ground-breaking study by Monroe et al.[78] provides insight into the mechanisms responsible for the evidence that highly expressed genes are under pronounced signatures of constraint. The authors find that differences in genes that are essential have a reduction in the mutation rate by 37% in *Arabidopsis thaliana* and that this reduced mutation rate for essential genes is associated with epigenomic features, such as H3K4me1[78]. In the context of venom evolution, we suspect that the distinction between a toxin family being categorized as either a core gene or peripheral gene may have important implications in the selection pressures acting at the sequence level. This is supported by previous work in different populations of eastern diamondback rattlesnakes (*Crotalus adamanteus*) that revealed that toxin gene expression dynamics, not positive selection at the nucleotide level, was the mechanism for these animals to overcome the resistance of population-specific prey, highlighting the ecological impact and selective pressure acting on toxin gene expression levels[79]. Dominant toxins have been proposed to be essential for broad ecological functions (such as general prey capture), while the peripheral toxins may have more prey-specific functions and are characterized by having greater divergence both at the expression and amino acid sequence levels[80–82].

Taken together with our findings, we report that the dominant toxin dictates the venom phenotype of sea anemones and hypothesize that this phenomenon might be shared across sea anemones, snakes and cone snails as well as other venomous groups, suggesting this is a trend that has evolved convergently among distantly related lineages. We argue that gene duplication is the mechanism that underlies this process.

Gene duplication represents an important mechanism for generating phenotypic variation over ecological and evolutionary timescales through the alteration of gene expression and diversification of variants[83–85]. We find that gene duplication plays a role in shaping the venom phenotype in sea anemones across both micro and macroevolution. The maintenance of clusters of duplicated genes is hypothesized to occur due to conserved regulation of expression. For toxin genes, highly duplicated toxins retained in a cluster could result in increased production of toxin protein due to the transcription of many copies of highly similar or identical genes[12]. In the case of Nv1, this is well-supported by measurements at the transcriptomic and proteomic levels in our current study and previously published works[43]. However, the transcription of *Nv1* varies significantly during the life cycle[43] and in response to a variety of environmental variation such as temperature and salinity[86] and light periodicity[87]. Environmentally elicited expression of *Nv1* differs based on the geographic origin and this transcriptional variation correlates with CNV, suggesting that gene dosage is the potential mechanism for local adaptation[86] (Supplementary Fig. 8). These results are consistent with snake myotoxins where it has been proposed that selection acts to increase expression as opposed to providing diversity through the permanent heterozygote or multiallelic diversifying selection models[79]. However, these myotoxin analyses excluded sequence variation in the exon responsible for the signal peptide as it is cleaved from the mature toxin. While we also observed low diversity of the mature toxin, consistent with ref. [13], our analyses across multiple populations show non-synonymous variation in the signal and propart peptide sequences. While the functional role of sequence variation in these regions in venom genes has not yet been explored, the amino acid composition and arrangement in signal and propart peptides has been shown to alter translocation, translation and cleavage efficiency[88]. As such, variation in this region of the gene could presumably alter the post-translational regulation of *Nv1*.

The Florida haplotypes raise important questions regarding their origin and the ecology of these populations. The presence of a haplotype without the *Nv1* locus suggests that *Nv1*-less homozygotes may be present in wild *N. vectensis* populations. The *Nv1*-less haplotype could reflect a phenomenon similar to the A-B dichotomy observed in snakes, where two distinct types of venoms exist in a largely mutually exclusive manner[89]. Under this scenario, Florida individuals may have compensated for low *Nv1* copies through the expansion of other toxin genes. However, we find no evidence of compensatory gene family expansion in 11 other known *N. vectensis* toxin genes (Supplementary Fig. 9). Alternatively, the low copy numbers associated with the Florida individual could result from the fitness costs associated with high gene expression. Venom production has a significant metabolic cost in *N. vectensis*[86] and *Nv1* is expressed by almost two orders of magnitude higher compared to the other toxins[43]. Thus, reduced venom capacity in a population at the upper thermal limit of the species range could potentially reflect the metabolic strain of venom production. However, summer temperatures in the South Carolina habitat are relatively similar to the ones in the Florida habitat. Instead, we suggest that such a massive reduction in toxin production as observed here should be associated with at least some differences in prey and/or predator composition and abundance between the Florida and South Carolina habitats as loss of defense or ability to predate with venom can be highly deleterious.

The exclusivity of paralogs to particular haplotypes suggests that recombination between contemporary haplotypes does not occur or is rare enough that it is beyond our limits of detection with these samples. The observed lack of recombination does not appear to result from the absence of heterozygous individuals as they are present in all populations, although, it is important to note that recombination between contemporary haplotypes could occur but its prevalence may be impacted by other factors such as selection. Nevertheless, this lack of evidence for recombination between core haplotypes helps provide insight into the mechanisms governing expansion and contraction within haplotypes. We observe substantial variation in the copy number, composition, and organization of paralogs within haplotypes including tandem duplications of singlet, duplet, and quadruplet *Nv1* paralogs (Supplementary Fig. 10). Furthermore, the expansion of different paralogs indicates that multiple independent expansion events have likely occurred at the same locus. The expansion and contraction of *Nv1* paralogs within core haplotypes in *N. vectensis* could be driven by non-allelic homologous recombination (NAHR) or replication slippage. NAHR is commonly associated with CNVs, including toxin genes, and within a single *Nv1* haplotype, there is a sufficient substrate for NAHR with regions of high sequence homology extending over >300 bp. In snakes, transposable elements have been proposed as the NAHR substrate[90,91]; however, this does not appear to be the case for the *Nv1* locus as TEs are largely absent from within the *Nv1* locus. If

NAHR is the mechanism driving the expansion and contraction of the *Nv1* haplotypes, it is unclear why it would not occur across haplotypes because sufficient NAHR substrate is evident between haplotypes despite the presence of haplotype-specific paralogs. An alternative hypothesis would be that expansions and contractions at the locus are a result of backward replication slippage[92]. We suggest that this mechanism is more likely responsible for the tandem duplications at the *Nv1* locus as there is the sufficient substrate, the duplication sizes are consistent with past observations of replication slippage, and importantly, would maintain the strong haplotype structure.

The presence of *Nv1.var12* in the single-copy Florida haplotype indicates that it is most closely related to the core haplotype H3. However, considering that this copy is not at the end of the locus, it suggests that even the single-copy Florida haplotype is at least two mutational steps from its closest relative and warrants further exploration for intermediate haplotypes in populations in the southeastern United States (e.g., Georgia). Analysis of the genomic context of the *Nv1* locus in the Florida haplotypes suggests that NAHR is unlikely to be the cause of these extreme contractions (Supplementary Fig. 6) and indicates that other processes are involved in the evolution of the *Nv1* locus.

The homogeneity of *Nv1* genes in the gene cluster has previously been hypothesized to maintain sequence similarity of duplicated genes through concerted evolution[13]. Later analyses of *Nv1*-like paralogs that translocated out of the cluster, which accrued proportionally more sequence divergence, further supported a hypothesis for concerted evolution of the *Nv1* cluster[66]. Toxin genes in other cnidarians also showed patterns of highly similar genes resulting from lineage-specific duplication events[13,38], suggesting concerted evolution may be common in the expansion of toxin families. Here, evidence for concerted evolution at the *Nv1* locus is confounded by our analysis of the spatial organization of *Nv1* genes in the cluster. First, although the reference haplotypes for the current and past genome assemblies contain a numerically overrepresented sequence, this is not a feature of all of the *Nv1* haplotypes. Second, the tandem arrangement of groups of paralogs (doublets, quadruplets) with consistent intergenic spacing might suggest that some of the similarities in loci is due to more recent duplications that retain the evolutionary history of the ancestral loci prior to duplications rather than homogenization of the array.

An alternative or additional hypothesis to concerted evolution for this locus is the birth–death model that has been proposed for other venom genes including *Nv1* paralogs that have escaped the *Nv1* locus[66]. Here, new gene copies arise through repeated duplications with some copies retained in the genome, while others become non-functional through mutation or are deleted[93]. Our analyses of the composition and spatial organization of *Nv1* genes demonstrate repeated duplications of genes and pairs of genes, providing support for the birth process. Furthermore, we also observed pseudogenes highlighting that not all genes are retained after duplication. Nevertheless, it is important to note that their number is very small compared to the seemingly functional copies. Due to the absence of an ancestral sequence, it is not possible to conclusively determine the extent of gene losses versus gene gains, however, the single-copy and *Nv1*-less haplotypes in Florida could represent a rapid gene death process. This would be consistent with other venom gene families where large deletions of genes have been observed[91].

Overall, our observations across macro- and microevolutionary timescales demonstrate that a single toxin family dictates the complex venom phenotype among sea anemones. Gene duplication underlies which toxin family becomes dominant through a process of increasing gene expression and this process is highly dynamic resulting in the rapid evolution of the venom phenotype across different species. High gene turnover rates of the dominant toxin family are found even within species, further signaling that strong selective forces are acting on toxin gene expression. Finally, as we see a similar trend is found in other venomous species, we hypothesize that gene duplication-driven dominance by a single toxin family is a fundamental process shaping the venom phenotype.

Methods

Phylotranscriptomics

We analyzed transcriptomes from 29 sea anemone species, spanning three of the five Actiniarian superfamilies (Actinioidea, Edwardsioidea, and Metrioidea). These transcriptomes that were sampled from either multiple tissues or tentacles were downloaded from NCBI SRA using FASTQ-DUMP in the SRA toolkit. Raw reads retrieved were assessed for quality and trimmed using Trimmomatic[94]. Trinity was used to assemble transcriptomes de novo from the filtered raw reads[95]. BUSCO (v4) was used to validate the quality and completeness of the transcriptomes[96]. Transcripts corresponding to toxins were identified using previously established methods[37], and then manually curated. Briefly, predicated open-reading frames encoding proteins for transcripts from each transcriptome was identified using ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/) and BLASTp (*E*-value 1e-01) performed against the swiss-prot database. Top hits against sea anemones toxins characterized in the Tox-Prot database[97] were retained and used to determine the presence of a signal peptide using SignalP (v5.0[98]). These sequences were then characterized into toxin families and aligned using[99] to retain only those with conserved cysteine frameworks are essential residues. Toxin families used in this analysis included only those that have been functionally characterized as toxins in multiple sea anemone species or shown to be localized to venom producing cells using multiple experimental approaches.

Toxin expression data were generated using software leveraged in the Trinity package (v > 2.2[100]). This included individual reads being mapped back to reference de novo transcriptome assemblies independently for each species using Bowtie2[101], and abundance estimated using RSEM[102]. Normalized abundance estimates of the transcript were calculated and corrected for their length to generate TPM values. Finally, we calculated the cumulative TPM values for each toxin family and the venom phenotype was generated as the percentage that each family contributes.

Transcripts with TPM values greater than zero were retained and their predicated open-reading frame was detected using ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/). Open-reading frames encoding proteins >100 amino acids in length were retained and redundant sequences with >88% similarity were removed to produce a predicted proteome for the 29 transcriptomes using CD-HIT[103].

Single-copy orthologs were identified using DIAMOND within the Orthofinder package[104]. This identified 138 single-copy orthologs that were individually aligned using MAFFT[99] and nucleotide alignment was generated using Pal2Nal using the coding sequence[105]. Aligned orthologs were concatenated and imported into IQ-TREE to determine the best-fit model of evolution[106]. The JTT model with gamma rate heterogeneity, invariable sites, and empirical codon frequencies were selected, and a maximum-likelihood tree was generated using 1000 ultrafast bootstrap iterations. An ultrametric tree was generated using by calibrating the maximum-likelihood tree Chronos function within the R package Ape using minimum and maximum age of root set to 424 and 608 million years ago[107,108]. Different calibration models were tested, including correlated, discrete, and relaxed models, with the discrete model determined to be the best fit.

Phylogenetic covariance analysis

PCA was performed as per ref. [28] using the R package MCMCglmm[109] with a multivariate model being used and toxin families as the

response variable. 20 million iterations were used, which included burnin and thinning values of 1 million and 1500, respectively. The phylogenetic signal was determined as previously described[28,110]. Principal component analysis was used by obtaining the phylogenetic covariances generated from the MCMCglmm analysis. Given sea anemones have a decentralized venom system, we tested whether the tissue type used to generate the raw reads significantly impacted the phylogenetic effect[25,26].

Modeling the ancestral states and modes of evolution acting on sea anemone venom

The R packages SURFACE and pulsR were used to test the models of evolution[111]. Evidence of phenotypic convergence was tested using SURFACE. The pulsR package was used to test the evolution of venom expression phenotype as either through a model incremental evolution or through pulsed evolution as modeled using the Lévy process[54]. The ancestral venom expression phenotype was reconstructed using *fastAnc* in the Phytools package[112].

Macro and microsynteny

Homologous chromosomes were found among the three genomes to determine macrosynteny. This was achieved by identifying 3767 single-copy orthologs using proteins annotated from all three genomes. The genomes of *N. vectensis*, *S. callimorphus* and *A. equina* were all investigated for the presence of NaTx and KTx3 toxins. Toxins from these genomes were identified using transcripts previously assembled using Trinity and mapped to the genome using Splign online software (https://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi). The chromosomal locations for the single-copy orthologs were compared to generate a broad macrosyntenic map of chromosomes among sea anemone genomes. The microsynteny neighboring NaTx/KTx3 loci was investigated by using BLAT from 30Kb upstream and downstream of the loci as well as comparing the 3 protein-coding genes upstream and downstream from NaTx/KTx3 loci. The NaTx/KTx3 copies identified from the genomes of the three species were clustered with publicly available copies previously used for evolutionary analyses[50], using CLANS software[113] with default settings and 300,000 rounds.

Population transcriptomics

Animal collection. Adult *N. vectensis* were collected from estuaries along the Atlantic coast of the United States and Canada. We collected 20 individuals from five locations (Crescent Beach, Nova Scotia; Saco, Maine; Wallis Sand, New Hampshire; Sippewissett, Massachusetts; Ft. Fisher, North Carolina) in March 2016, and an additional 10 individuals/month from three of these locations (Saco, Maine; Wallis Sands, New Hampshire; Sippewissett, Massachusetts) in June and September 2016. Individuals were stored in RNAlater and stored at −20 °C prior to nucleic acid extraction for qPCR and amplicon analyses. At each collection, additional individuals were transported to UNC Charlotte and cultured in the laboratory under standard laboratory conditions (15 parts per thousand artificial seawater, room temperature, fed freshly hatched Artemia 2-3 times per week). In addition, eight adult *N. vectensis* collected near St. Augustine, Florida were kindly provided by Lukas Schäre (University of Basel). From these laboratory populations, we selected four individual anemones to grow clonal lines for long-read sequencing; single individuals from Nova Scotia, Maine, North Carolina, and Florida were grown and bisected to generate the lines.

To investigate the population-level comparison of venom among *N. vectensis*, transcriptomics was performed. Multiple individuals from nine locations in North America were collected. This included the same locations as mentioned above in Florida, Massachusetts, Maine, North Carolina, New Hampshire, Nova Scotia and as well as New Jersey (Brigantine), Maryland (Rhode River), and South Carolina (Georgetown). Individuals from these locations were brought back to the lab and allowed to acclimatize for 2 weeks.

Total RNA was extracted from pools of three whole specimens per site from nine different locations using RNeasy Mini Kit (Qiagen, USA). Quality and integrity of extracted RNA was assessed using Bioanalyzer 2100 (Agilent, USA) using an RNA nano chip (RIN > 8). Sequencing libraries were prepared using the Kapa Stranded mRNA-seq kit (Roche, Switzerland) and sequenced on an Illumina HiSeq 4000 using 150 bp paired-end chemistry performed at Duke Center for Genomic and Computational Biology (Durham, NC, USA). Raw reads from each population were cleaned using Trimmomatic[94] to retain only high-quality reads and to remove non-biological sequence and assembled into nine transcriptomes using the Trinity 2.6.6[95]. To assess the completeness of de novo transcriptomes, BUSCO (v3.0) was performed on each transcriptome to assess the completeness of each assembly, by determining the percentage of full-length sequences in each transcriptome corresponding to a conserved set of metazoan orthologs[114].

Comparative transcriptomics were performed to reconstruct the phylogenetic relatedness of *N. vectensis* populations across North America. For each transcriptome, open-reading frames were identified using ORFfinder and translated using Transeq. Redundant sequences with >88% sequence similarity were removed using CD-HIT[103]. Protein sequences >100 amino acids in length were used to identify single-copy orthologs using OrthoFinder[115] and leveraged using DIAMOND[116]. In addition, we added *S. callimorphus* and *Edwardsiella carnea* as outgroups to the *N. vectensis* populations. This resulted in 2589 single-copy orthologs shared among the 11 transcriptomes. Protein sequences for each single-copy ortholog were individually aligned using MAFFT[99]. Protein alignments were then concatenated and imported into IQ-TREE, and the best-fit model of evolution selected using ModelFinder[106], and posterior mean site frequency models were used to reduce long-branch attraction artefacts[117]. A maximum-likelihood phylogenetic tree was generated using 1000 ultrafast bootstrap iterations.

Variations of toxins from the NaTx gene family in *N. vectensis* (*Nv1*, *Nv2*, *Nv3*, *Nv4*, *Nv5*, *Nv6*, and *Nv7*) were investigated among the different populations using multiple approaches. Initially, BLASTp was performed to identify toxins using ORFs from the transcriptomes against a custom database consisting of all known sequences from the *Nv1* gene family. Sequences with a significant hit (*E*-value 1e-05) were then manually curated to determine the presence of a signal peptide and conserved cysteine framework. In addition, *Nv1* copies have been previously reported to be massively duplicated (with at least ten copies previously reported) and highly homogenous in the genome of *N. vectensis*[13]. For these reasons, additional approaches were required to capture these limited variations of *Nv1* copies among the populations. To achieve this, cleaned raw reads were mapped using Bowtie2 plugin in Trinity using default settings[95,101] to the *N. vectensis* gene models with *Nv1* reduced to a single copy[86]. Paired-end reads mapping to *Nv1* were then extracted and aligned to the *Nv1* gene model using MAFFT[99], and a new consensus *Nv1* sequence generated for each mapped paired-end read using cons in EMBOSS. Identical *Nv1* sequences were then clustered using CD-HIT-EST[103] and only the top most abundant sequences that accounted for 70% of the total number of sequences or had a minimum of 10 identical copies were retained. Florida sample was an exception in which only the most abundant sequence was retained as it had four identical copies. The open-reading frame was identified and redundant coding sequences with removed to give a representation of allelic variation in *Nv1* in different populations. To obtain a allelic variation of *Nv3*, mapped paired-end reads that had a *Nv3* signature (AAACGCGGCTTTGCT, which encodes for KRGFA, as opposed to *Nv1* AAACGCGGCATTCCT which encodes for KRGIP) were extracted and aligned to *Nv3* coding sequence using MAFFT[99]. The most abundant consensus sequences that accounted for 70% of the total number of sequences or had a minimum of 10 identical copies were retained, and redundant coding sequences removed.

Population proteomics

Semi-quantitative MS/MS analysis was performed using adults (four replicates, each made of three individuals) from both North Carolina and Florida. Samples were snap frozen and lysed using in 8 M urea and 400 mM ammonium bicarbonate solution. Lysed samples were centrifuged (22,000 × *g*, 20 min, 4 °C) and supernatant collected. Protein concentrations were measured with BCA Protein Assay Kit (Thermo Fisher Scientific).

Sample preparation for MS analysis. Ten micrograms of protein were dissolved in 100 µl of 8 M urea, 10 mM DTT, 25 mM Tris-HCl pH 8.0 for 30 min at 22 °C. Iodoacetamide (55 mM) was added and followed by incubation for 30 min (22 °C, in the dark). The samples were diluted with 8 volumes of 25 mM Tris-HCl pH 8.0 followed by the addition of sequencing-grade modified Trypsin (Promega Corp., Madison, WI) (0.4 µg/ sample) and incubation overnight at 37 °C. The peptides were acidified by the addition of 0.4% formic acid and transferred to C18 home-made stage tips for desalting. The peptide concentration was determined by absorbance at 280 nm and 0.3 µg of peptides were injected into the mass spectrometer.

nanoLC-MS/MS analysis. nanoLC-MS/MS analysis was performed as previously described in ref. [118] with the exception that peptides dissolved in 0.1% formic acid were separated without a trap column over an 80 min acetonitrile gradient run at a flow rate of 0.3 µl/min on a reverse phase 25-cm-long C18 column (75 µm ID, 2 µm, 100 Å, Thermo PepMapRSLC). The instrument settings were as described in ref. [119].

MS data analysis. Mass spectra data were processed using the MaxQuant computational platform, version 2.0.3.0. Peak lists were searched against an NVE FASTA sequence database (https://figshare.com/articles/Nematostella_vectensis_transcriptome_and_gene_models_v2_0/807696). The search included cysteine carbamidomethylation as a fixed modification, N-terminal acetylation, and oxidation of methionine as variable modifications and allowed up to two miscleavages. The "match-between-runs" option was used. Peptides with a length of at least seven amino acids were considered and the required FDR was set to 1% at the peptide and protein level. Relative protein quantification in MaxQuant was performed using the LFQ algorithm[120]. MaxLFQ allows accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction.

Statistical analysis ($n = 4$) was performed using the Perseus statistical package, Version 1.6.2.2[121]. Only those proteins for which at least three valid LFQ values were obtained in at least one sample group were accepted and log2 transformed. Statistical analysis by Student's *t* test and permutation-based FDR (*P* value <0.05). After the application of this filter, a random value was substituted for proteins for which LFQ could not be determined ("Imputation" function of Perseus). The imputed values were in the range of 10% of the median value of all the proteins in the sample and allowed the calculation of *P* values. To test if proteomes were comparable, we performed linear regression between the Florida and North Carolina samples. Proteins with non-zero LFQ values in at least one sample for each population were used and transformed per million.

Population genomics

Quantification of Nv1 copy number. We used quantitative PCR to determine the number of *Nv1* copies in individuals collected from each location using hydrolysis probe-based quantitative PCR. DNA for anemones from each location was isolated with the AllPrep DNA/RNA kit (Qiagen) using the manufacturer's protocol. Primers and hydrolysis probes were designed for *Nv1* and *Catalase* using Primer3[122]. The hydrolysis probes contained distinct fluorophores for each gene in addition to 3′ and internal quenchers (*Nv1* = 5′ Cy5/TAO/3′ IBRQ; Cat = 5′ 6-FAM/ZEN/3′ IBFQ). Amplifications were performed in an Applied Biosystems 7500 Fast Real-Time PCR System using the Luna Universal Probe Mix (NEB). We evaluated the performance of the qPCR primers and probes using a DNA concentration gradient spanning 0.1–100.0 ng/reaction in single gene reactions as well as in a multiplex reaction. The efficiencies of both gene assays were within the recommended range (90–110%), comparable across the concentration gradient, and consistent between the single and multiplex reactions. As such, we amplified *Catalase* and *Nv1* in triplicate multiplex reactions for each sample, and each 96-well plate contained samples from all populations. In addition, each plate contained triplicate reactions of a reference sample of known copy number from Florida (diploid copy number = 1; derived from genome assembly), a no template control (NTC), and two samples to monitor variability between plates. The diploid copy number was estimated using the ΔΔCt approach with Catalase as the single-copy control gene and the Floridian sample of known copy number as our reference sample. There was no amplification observed in any NTCs.

The Cq values were determined automatically in the Applied Biosystems software. We filtered individuals from the qPCR results where the Cq values for either gene was outside of the range used for efficiency estimation (two individuals), and filtered individual reactions where the Cq values deviated by more than 0.2 Cq between any of the triplicate reactions for either gene (6/549 reactions). As we used multiplex reactions, mean ΔCt was calculated as the mean of ΔCt across individual reactions. We performed a two-way ANOVA (diploid copy number ~ population * plate) in cab package R using Type II SS (to account for the unbalanced design) to test for the effect of population on diploid *Nv1* copy number while accounting for any potential batch effects. The ANOVA tests revealed no significant effect of plate or population-by-plate on our copy number estimates. Tukey HSD post hoc tests (alpha = 0.05) were performed using the agricolae package[123].

Amplification and sequencing of *Nv1*. We designed primers to amplify *Nv1* loci from genomic DNA for sequence analysis with the Illumina MiSeq. Primers were designed to amplify the full coding sequence for *Nv1* and minimized mismatches with SNPs identified between known *Nv1* variants. Primers contained the adapter overhang for Nextera Indexing. PCRs were performed with HiFi HotStart ReadyMix (Kapa Biosciences) using the following conditions: 95 °C−3 min; 8 x (95 °C−30s, 55 °C−30 s, 72 °C−30 s), 72 °C−5 min. PCR products were purified with Ampure XP beads (Beckman Colter). Successful amplification of the anticipated product size was verified by gel electrophoresis. Amplicons from each sample were quantified by Qubit (Thermo Fisher Scientific) for normalization. Equal concentrations of each sample were pooled with 5% PhiX for sequencing using a MiSeq v3 reagent kit (600 cycles). We used mothur v1.44.3[124] to join overlapping reads to make contigs that were subsequently filtered to remove amplicons outside of *Nv1* size expectations (300–500 bp) and with ambiguous bases. Cutadapt v2.6[125] was subsequently used to remove primer sequences. We randomly subsampled the FASTA files to a depth of 14,800 reads, with four samples removed from future analyses due to insufficient reads. In order to distinguish biological sequence variation from methodological artifacts (e.g., PCR and sequencing errors), we identified a list of variants based on a minimum sample read abundance of 100 and presence in more than one individual. A heatmap of relative abundance of variants across samples was generated using the ComplexHeatmap package[126] in R. Hierarchical clustering of samples was performed using Spearman rank correlation as the distance measure.

PacBio and nanopore sequencing. High-quality DNA was extracted from individual clone lines from four geographic locations (Nova Scotia, Maine, North Carolina, and Florida) using a previously described extraction protocol[127]. This protocol was adapted for HMW DNA by the addition of tissue grinding in liquid nitrogen, decreasing the incubation time and temperature to one hour at 42 °C, increasing the elution time to 24 h, and the use of wide-bore tips throughout the protocol.

For Nova Scotia and Florida anemones we sequenced DNA from single genotypes with PacBio technology. DNA was shipped to Brigham Young University (Provo, UT, USA) for quality check with pulsefield capillary electrophoresis followed by CLR library construction and sequencing (PacBio Sequel II). The unique molecular yields were 38 Gb and 123 Gb, with the longest subread N50s of 35 kb and 28 kb, respectively. PacBio reads were assembled into contigs using Canu v2.0[128], configured to assemble both haplotypes at each locus separately. Two rounds of polishing were applied to each assembly by aligning raw PacBio data using pbmm2 (v1.3.0) and using the multimolecule consensus setting of the Arrow algorithm implemented in gcpp (v.1.9.0)[129]. Transposable elements annotations for the PacBio assemblies, in addition to the Maryland reference, were generated by EDTA v1.9.6[130] using a combined fasta file containing all three assemblies.

For Maine and North Carolina anemones, short DNA fragments were removed using the short read eliminator kit (Circulomics). Libraries were prepared for Nanopore sequencing using the ligation sequencing kit (LSK109) and sequenced on a single MinION flow cell (R9.4.1; Oxford Nanopore Technologies). The Nanopore long reads were basecalled using guppy (v4.5.2), assembled into contigs using Canu v2.1[128], and *Nv1* contigs polished using Racon v1.4.21[131]. For the Maine sample, only one *Nv1* haplotype was assembled. Evaluation of the intergenic spacing of *Nv1* copies in raw reads based on BLASTn searches was consistent across all reads suggestive of a homozygous individual. In contrast, evaluation of the North Carolina raw reads showed reads could be split into two separate groups based on disparate intergenic spacings. These two sets of reads were assembled separately.

For this study, we have only focused our analysis on the contigs corresponding to the *Nv1* cluster. These contigs and their respective *Nv1* copy number and localization were identified using BLASTn searches against the assemblies. Pseudogenes were identified as copies as *Nv1* copies with premature stop codons and truncated mature peptide sequences. An analysis of the remaining portions of the genome for each clone line will be reported in a future publication.

Expression of *Nv1* for individuals originally collected from Florida was quantified with nCounter technology. This approach was identical to methods reported for quantification of *Nv1* for *N. vectensis* from other geographic locations reported in Sachova et al. [86]. Briefly, individuals were acclimatized at 20 °C for 24 h in the dark in 15‰ artificial seawater (ASW). Individuals were subsequently exposed to one of three temperatures in the dark: 20 °C (control), 28 °C, and 36 °C for 24 h. Animals were placed into tubes and frozen to obtain three replicates for each condition, two animals/replicate. Extracted RNA was shipped for analysis using the nCounter platform (NanoString Technologies, USA; performed by MOgene, USA) for expression of *Nv1* using the same custom probe previously reported.

Reporting summary

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

Data availability

Raw sequencing data for *N. vectensis* populations have been submitted to the NCBI SRA database for transcriptomics (BioProject: PRJNA831625), amplicon sequencing (BioProject: PRJNA836916) and genomics (BioProject: PRJNA844989). Proteomics from North Carolina and Florida populations has been submitted to the proteome exchange (PXD034383). Sequences used in this study have also been uploaded as FASTA files to figshare (https://doi.org/10.6084/m9.figshare.20115719.v1). Accession numbers for data used in this project for the phylotranscriptomics can be found in Supplementary Data 1. Source data are provided with this paper.

References

1. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
2. Lynch, M. & Walsh, B. Genetics and analysis of quantitative traits. *Q. Rev. Biol.* **74**, 225–225 (1999).
3. Mathieson, I. The omnigenic model and polygenic prediction of complex traits. *Am. J. Hum. Genet.* **108**, 1558–1563 (2021).
4. Sella, G. & Barton, N. H. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* **20**, 461–493 (2019).
5. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
6. Hill, M. S., Vande Zande, P. & Wittkopp, P. J. Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* 1–13 https://doi.org/10.1038/s41576-020-00304-w (2020).
7. Zheng, W., Gianoulis, T. A., Karczewski, K. J., Zhao, H. & Snyder, M. Regulatory variation within and between species. *Annu. Rev. Genomics Hum. Genet.* **12**, 327–346 (2011).
8. Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. & Teichmann, S. A. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* **14**, 283–291 (2004).
9. Yu, H. & Gerstein, M. Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl Acad. Sci. USA* **103**, 14724–14731 (2006).
10. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
11. Raveh-Sadka, T. et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.* **44**, 743–750 (2012).
12. Giorgianni, M. W. et al. The origin and diversification of a novel protein family in venomous snakes. *Proc. Natl Acad. Sci. USA* **117**, 10911–10920 (2020).
13. Moran, Y. et al. Concerted evolution of sea anemone neurotoxin genes is revealed through analysis of the *Nematostella vectensis* genome. *Mol. Biol. Evol.* **25**, 737–747 (2008).
14. Robinson, D., Place, M., Hose, J., Jochem, A. & Gasch, A. P. Natural variation in the consequences of gene overexpression and its implications for evolutionary trajectories. *eLife* **10**, e70564 (2021).
15. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).
16. Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10**, 451–481 (2009).
17. Brown, C. J., Todd, K. M. & Rosenzweig, R. F. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol. Biol. Evol.* **15**, 931–942 (1998).
18. Perry, G. H. et al. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).
19. Pös, O. et al. Copy number variation: methods and clinical applications. *Appl. Sci.* **11**, 819 (2021).
20. Sudmant, P. H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).

21. Lighten, J., van Oosterhout, C., Paterson, I. G., McMullan, M. & Bentzen, P. Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Mol. Ecol. Resour.* **14**, 753–767 (2014).
22. Pajic, P. et al. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *eLife* **8**, e44628 (2019).
23. Weetman, D. et al. Contemporary evolution of resistance at the major insecticide target site gene Ace-1 by mutation and copy number variation in the malaria mosquito *Anopheles gambiae*. *Mol. Ecol.* **24**, 2656–2672 (2015).
24. Casewell, N. R., Wüster, W., Vonk, F. J., Harrison, R. A. & Fry, B. G. Complex cocktails: the evolutionary novelty of venoms. *Trends Ecol. Evol.* **28**, 219–229 (2013).
25. 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).
26. Surm, J. M. & Moran, Y. Insights into how development and life-history dynamics shape the evolution of venom. *EvoDevo* **12**, 1 (2021).
27. Casewell, N. R., Jackson, T. N. W., Laustsen, A. H. & Sunagar, K. Causes and consequences of snake venom variation. *Trends Pharmacol. Sci.* **41**, 570–581 (2020).
28. Barua, A. & Mikheyev, A. S. Many options, few solutions: over 60 my snakes converged on a few optimal venom formulations. *Mol. Biol. Evol.* **36**, 1964–1974 (2019).
29. Mason, A. J. et al. Trait differentiation and modular toxin expression in palm-pitvipers. *BMC Genomics* **21**, 147 (2020).
30. Pineda, S. S. et al. Structural venomics reveals evolution of a complex venom by duplication and diversification of an ancient peptide-encoding gene. *Proc. Natl Acad. Sci. USA* **117**, 11399–11408 (2020).
31. Sanggaard, K. W. et al. Spider genomes provide insight into composition and evolution of venom and silk. *Nat. Commun.* **5**, 3765 (2014).
32. Chang, D. & Duda, T. F. Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Mol. Biol. Evol.* **29**, 2019–2029 (2012).
33. Cao, Z. et al. The genome of *Mesobuthus martensii* reveals a unique adaptation model of arthropods. *Nat. Commun.* **4**, 2602 (2013).
34. Haney, R. A. et al. Effects of gene duplication, positive selection, and shifts in gene expression on the evolution of the venom gland transcriptome in widow spiders. *Genome Biol. Evol.* **8**, 228–242 (2016).
35. Daly, M. et al. The phylum Cnidaria: a review of phylogenetic patterns and diversity 300 years after Linnaeus*. *Zootaxa* **1668**, 127–182 (2007).
36. Prentis, P. J., Pavasovic, A. & Norton, R. S. Sea anemones: quiet achievers in the field of peptide toxins. *Toxins* **10**, 36 (2018).
37. Surm, J. M. et al. A process of convergent amplification and tissue-specific expression dominates the evolution of toxin and toxin-like genes in sea anemones. *Mol. Ecol.* **28**, 2272–2289 (2019).
38. Surm, J. M., Stewart, Z. K., Papanicolaou, A., Pavasovic, A. & Prentis, P. J. The draft genome of *Actinia tenebrosa* reveals insights into toxin evolution. *Ecol. evolution* **9**, 11314–11328 (2019).
39. Beckmann, A. & Özbek, S. The nematocyst: a molecular map of the cnidarian stinging organelle. *Int. J. Dev. Biol.* **56**, 577–582 (2012).
40. Moran, Y. et al. Neurotoxin localization to ectodermal gland cells uncovers an alternative mechanism of venom delivery in sea anemones. *Proc. Biol. Sci.* **279**, 1351–1358 (2012).
41. Anderluh, G. & Maček, P. Cytolytic peptide and protein toxins from sea anemones (Anthozoa: Actiniaria). *Toxicon* **40**, 111–124 (2002).
42. Macrander, J. & Daly, M. Evolution of the cytolytic pore-forming proteins (Actinoporins) in sea anemones. *Toxins* **8**, 368 (2016).
43. Columbus-Shenkar, Y. Y. et al. Dynamics of venom composition across a complex life cycle. *eLife* **7**, e35014 (2018).
44. Moran, Y. et al. Analysis of soluble protein contents from the nematocysts of a model sea anemone sheds light on venom evolution. *Mar. Biotechnol.* **15**, 329–339 (2012).
45. Moran, Y., Gordon, D. & Gurevitz, M. Sea anemone toxins affecting voltage-gated sodium channels—molecular and evolutionary features. *Toxicon* **54**, 1089–1101 (2009).
46. Wanke, E., Zaharenko, A. J., Redaelli, E. & Schiavon, E. Actions of sea anemone type 1 neurotoxins on voltage-gated sodium channel isoforms. *Toxicon* **54**, 1102–1111 (2009).
47. Castañeda, O. & Harvey, A. L. Discovery and characterization of cnidarian peptide toxins that affect neuronal potassium ion channels. *Toxicon* **54**, 1119–1124 (2009).
48. Orts, D. J. B. et al. Biochemical and electrophysiological characterization of two sea anemone type 1 potassium toxins from a geographically distant population of *Bunodosoma caissarum*. *Mar. Drugs* **11**, 655–679 (2013).
49. Tudor, J. E., Pallaghy, P. K., Pennington, M. W. & Norton, R. S. Solution structure of ShK toxin, a novel potassium channel inhibitor from a sea anemone. *Nat. Struct. Mol. Biol.* **3**, 317–320 (1996).
50. Jouiaei, M. et al. Evolution of an ancient venom: recognition of a novel family of cnidarian toxins and the common evolutionary origin of sodium and potassium neurotoxins in sea anemone. *Mol. Biol. Evol.* **32**, 1598–1610 (2015).
51. Moran, Y. et al. Intron retention as a posttranscriptional regulatory mechanism of neurotoxin expression at early life stages of the Starlet Anemone *Nematostella vectensis*. *J. Mol. Biol.* **380**, 437–443 (2008).
52. Moran, Y. & Gurevitz, M. When positive selection of neurotoxin genes is missing. *FEBS J.* **273**, 3886–3892 (2006).
53. Barua, A. & Mikheyev, A. S. Toxin expression in snake venom evolves rapidly with constant shifts in evolutionary rates. *Proc. R. Soc. B: Biol. Sci.* **287**, 20200613 (2020).
54. Landis, M. J. & Schraiber, J. G. Pulsed evolution shaped modern vertebrate body sizes. *Proc. Natl Acad. Sci. USA* **114**, 13224–13229 (2017).
55. Wilding, C. S. et al. The genome of the sea anemone *Actinia equina* (L.): meiotic toolkit genes and the question of sexual reproduction. *Mar. Genomics* **53**, 100753 (2020).
56. Zimmermann, B. et al. Sea anemone genomes reveal ancestral metazoan chromosomal macrosynteny. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.30.359448 (2020).
57. Diochot, S., Schweitz, H., Béress, L. & Lazdunski, M. Sea anemone peptides with a specific blocking activity against the fast inactivating potassium channel Kv3.4. *J. Biol. Chem.* **273**, 6744–6749 (1998).
58. Peigneur, S. et al. A natural point mutation changes both target selectivity and mechanism of action of sea anemone toxins. *FASEB J.* **26**, 5141–5151 (2012).
59. van Vlijmen, H. W. T., Gupta, A., Narasimhan, L. S. & Singh, J. A novel database of disulfide patterns and its application to the discovery of distantly related homologs. *J. Mol. Biol.* **335**, 1083–1092 (2004).
60. Zaharenko, A. J. et al. Proteomics of the neurotoxic fraction from the sea anemone *Bunodosoma cangicum* venom: novel peptides

belonging to new classes of toxins. *Comp. Biochem. Physiol. Part D: Genomics Proteom.* **3**, 219–225 (2008).
61. Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86–94 (2007).
62. Reitzel, A. M., Herrera, S., Layden, M. J., Martindale, M. Q. & Shank, T. M. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Mol. Ecol.* **22**, 2953–2970 (2013).
63. Ashwood, L. M. et al. Venoms for all occasions: the functional toxin profiles of different anatomical regions in sea anemones are related to their ecological function. *Mol. Ecol.* **31**, 866–883 (2022).
64. Bathke, J., Konzer, A., Remes, B., McIntosh, M. & Klug, G. Comparative analyses of the variation of the transcriptome and proteome of *Rhodobacter sphaeroides* throughout growth. *BMC Genomics* **20**, 358 (2019).
65. Takemon, Y. et al. Proteomic and transcriptomic profiling reveal different aspects of aging in the kidney. *eLife* **10**, e62585 (2021).
66. Sachkova, M. Y. et al. The birth and death of toxins with distinct functions: a case study in the sea anemone Nematostella. *Mol. Biol. Evol.* **36**, 2001–2012 (2019).
67. Hanada, K. et al. The functional role of pack-MULEs in rice inferred from purifying selection and expression profile. *Plant Cell* **21**, 25–38 (2009).
68. Bacolla, A. et al. Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc. Natl Acad. Sci. USA* **101**, 14162–14167 (2004).
69. Phuong, M. A., Mahardika, G. N. & Alfaro, M. E. Dietary breadth is positively correlated with venom complexity in cone snails. *BMC Genomics* **17**, 401 (2016).
70. Casewell, N. R. et al. Medically important differences in snake venom composition are dictated by distinct postgenomic mechanisms. *Proc. Natl Acad. Sci. USA* **111**, 9205–9210 (2014).
71. Jenner, R. A., von Reumont, B. M., Campbell, L. I. & Undheim, E. A. Parallel evolution of complex centipede venoms revealed by comparative proteotranscriptomic analyses. *Mol. Biol. Evol.* **36**, 2748–2763 (2019).
72. Madio, B., Undheim, E. A. & King, G. F. Revisiting venom of the sea anemone *Stichodactyla haddoni*: Omics techniques reveal the complete toxin arsenal of a well-studied sea anemone genus. *J. Proteom.* **166**, 83–92 (2017).
73. Rokyta, D. R., Margres, M. J. & Calvin, K. Post-transcriptional mechanisms contribute little to phenotypic variation in snake venoms. *G3: Genes, Genomes, Genet.* **5**, 2375–2382 (2015).
74. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**, e270 (2014).
75. Wang, D. et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, e8503 (2019).
76. Kazandjian, T. D. et al. Convergent evolution of pain-inducing defensive venom components in spitting cobras. *Science* **371**, 386–390 (2021).
77. Wu, Z. et al. Expression level is a major modifier of the fitness landscape of a protein coding gene. *Nat. Ecol. Evol.* **6**, 103–115 (2022).
78. Monroe, J. G. et al. Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* **602**, 101–105 (2022).
79. Margres, M. J., Bigelow, A. T., Lemmon, E. M., Lemmon, A. R. & Rokyta, D. R. Selection to increase expression, not sequence diversity, precedes gene family origin and expansion in rattlesnake venom. *Genetics* **206**, 1569–1580 (2017).
80. Gibbs, H. L., Sanz, L. & Calvete, J. J. Snake population venomics: proteomics-based analyses of individual variation reveals significant gene regulation effects on venom protein expression in Sistrurus Rattlesnakes. *J. Mol. Evol.* **68**, 113–125 (2009).
81. Margres, M. J. et al. Expression differentiation is constrained to low-expression proteins over ecological timescales. *Genetics* **202**, 273–283 (2016).
82. Rautsaw, R. M. et al. Intraspecific sequence and gene expression variation contribute little to venom diversity in sidewinder rattlesnakes (*Crotalus cerastes*). *Proc. R. Soc. B: Biol. Sci.* **286**, 20190810 (2019).
83. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010).
84. Kondrashov, F. A. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. B: Biol. Sci.* **279**, 5048–5057 (2012).
85. Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. & Ravikesavan, R. Gene duplication as a major force in evolution. *J. Genet.* **92**, 155–161 (2013).
86. Sachkova, M. Y. et al. Some like it hot: population-specific adaptations in venom production to abiotic stressors in a widely distributed cnidarian. *BMC Biol.* **18**, 121 (2020).
87. Leach, W. B. & Reitzel, A. M. Transcriptional remodelling upon light removal in a model cnidarian: losses and gains in gene expression. *Mol. Ecol.* **28**, 3413–3426 (2019).
88. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: structure, roles, and applications. *Eur. J. Cell Biol.* **97**, 422–441 (2018).
89. Glenn, J. L., Straight, R. C. & Wolt, T. B. Regional variation in the presence of canebrake toxin in *Crotalus horridus* venom. *Comp. Biochem. Physiol. Part C: Pharmacol., Toxicol. Endocrinol.* **107**, 337–346 (1994).
90. Dowell, N. L. et al. The deep origin and recent loss of venom toxin genes in rattlesnakes. *Curr. Biol.* **26**, 2434–2445 (2016).
91. Margres, M. J. et al. The Tiger Rattlesnake genome reveals a complex genotype underlying a simple venom phenotype. *Proc. Natl Acad. Sci. USA* **118**, e2014634118 (2021).
92. Chen, J.-M., Chuzhanova, N., Stenson, P. D., Férec, C. & Cooper, D. N. Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. *Hum. Mutat.* **25**, 207–221 (2005).
93. Nei, M., Gu, X. & Sitnikova, T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl Acad. Sci. USA* **94**, 7799–7806 (1997).
94. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
95. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
96. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evolution* **38**, 4647–4654 (2021).
97. Jungo, F. & Bairoch, A. Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein knowledgebase. *Toxicon* **45**, 293–301 (2005).
98. Almagro Armenteros, J. J. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).

99. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
100. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotech.* **29**, 644–652 (2011).
101. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
102. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
103. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
104. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
105. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucl. Acids Res.* **34**, W609–W612 (2006).
106. 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).
107. McFadden, C. S. et al. Phylogenomics, origin, and diversification of Anthozoans (Phylum Cnidaria). *Syst. Biol.* **70**, 635–647 (2021).
108. Quattrini, A. M. et al. Palaeoclimate ocean conditions shaped the evolution of corals and their skeletons through deep time. *Nat. Ecol. Evol.* **4**, 1531–1538 (2020).
109. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**, 1–22 (2010).
110. de Villemereuil, P. & Nakagawa, S. General quantitative genetic methods for comparative biology. in *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice* (ed. Garamszegi, L. Z.) 287–303 (Springer, 2014).
111. Ingram, T. & Mahler, D. L. SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. *Methods Ecol. Evol.* **4**, 416–425 (2013).
112. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
113. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).
114. Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
115. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 1–14 (2015).
116. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
117. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **67**, 216–235 (2018).
118. Carmon, I. et al. HU308 Mitigates Osteoarthritis by Stimulating Sox9-Related Networks of Carbohydrate Metabolism. *Journal of Bone and Mineral Research* **38**, 154–170 (2023).
119. Scheltema, R. A. et al. The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol. Cell Proteom.* **13**, 3698–3708 (2014).
120. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteom.* **13**, 2513–2526 (2014).
121. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
122. Untergasser, A. et al. Primer3-new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
123. De Mendiburu, F. & Simon, R. Agricolae—ten years of an open source statistical tool for experiments in breeding, agriculture and biology. https://doi.org/10.7287/peerj.preprints.1404v1 (2015).
124. Schloss, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ. Microbiol.* **75**, 7537–7541 (2009).
125. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
126. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
127. Smith, E. G., Ketchum, R. N. & Burt, J. A. Host specificity of Symbiodinium variants revealed by an ITS2 metahaplotype approach. *ISME J.* **11**, 1500–1503 (2017).
128. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
129. Holt, R. A. et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149 (2002).
130. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
131. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

Acknowledgements

The authors are grateful to Dr. William Breuer (Mass Spectrometry Core Facility, The Hebrew University of Jerusalem) for his help with proteomics and Dr. Peter Prentis (Queensland University of Technology) for providing sea anemone photos. This work was supported by the Lady Davis fellowship to J.M.S., National Science Foundation fellowship 1924498 to E.G.S., Israel Science Foundation grant 636/21 to Y.M., incentive funding from the CIPHER Center at UNC Charlotte to A.M.R., and Binational Science Foundation program with the National Science Foundation grants 1536530 and 2020669 to Y.M. and A.M.R.

Author contributions

Conceptualization: E.G.S., J.M.S., J.M., A.M.R., and Y.M.; computational analysis: E.G.S., J.M.S., J.M., A.S., and G.A.; Experimental analysis: E.G.S., J.M.S., J.M., M.Y.S., and M.L.; writing—original draft: E.G.S., J.M.S., A.M.R., and Y.M.; writing—review and editing all authors; supervision: J.M.S., A.M.R., and Y.M.; funding acquisition: E.G.S., J.M.S., A.M.R., and Y.M.

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/s41467-023-35794-9.

Correspondence and requests for materials should be addressed to Edward G. Smith, Joachim M. Surm or Yehu Moran.

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 permissions 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) 2023