The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins

Kushal Suryamohan1,2, Sajesh P. Krishnankutty3,4, Joseph Guillory5, Matthew Jevit6, Markus S. Schröder1, Meng Wu1, Boney Kurianose3, Oomen K. Mathew3, Rajadurai C. Perumal6, Ivan Koludarov6, Leonard D. Goldstein1,2, Kate Senger1, Mandumpala Davis Dixon3, Dinesh Velayutham3, Derek Vargas1,2, Subhra Chaudhuri1, Megha Muraleedharan3, Ridhi Goel3, Ying-Jiun J. Chen1, Aakrosh Ratan8, Peter Liu9, Brendan Faherty9, Guillermo de la Rosa10, Hiroki Shibata11, Miriam Baca12, Meredith Sagolla12, James Ziai12, Gus A. Wright13, Domagoj Vucic14, Sangeetha Mohan15, Aju Antony15, Jeremy Stinson1, Donald S. Kirkpatrick9, Rami N. Hannoush14, Steffen Durinck12, Zora Modrusan1, Eric W. Stawiski1,2, Kristen Wiley16, Terje Raudsepp5, R. Manjunatha Kini17, Arun Zachariah18 and Somasekar Seshagiri1,4*

Snakebite envenoming is a serious and neglected tropical disease that kills ~100,000 people annually. High-quality, genome-enabled comprehensive characterization of toxin genes will facilitate development of effective humanized recombinant antivenom. We report a de novo near-chromosomal genome assembly of *Naja naja*, the Indian cobra, a highly venomous, medically important snake. Our assembly has a scaffold N50 of 223.35 Mb, with 19 scaffolds containing 95% of the genome. Of the 23,248 predicted protein-coding genes, 12,346 venom-gland-expressed genes constitute the ‘venom-ome’ and this included 139 genes from 33 toxin families. Among the 139 toxin genes were 19 ‘venom-ome-specific toxins’ (VSTs) that showed venom-gland-specific expression, and these probably encode the minimal core venom effector proteins. Synthetic venom reconstituted through recombinant VST expression will aid in the rapid development of safe and effective synthetic antivenom. Additionally, our genome could serve as a reference for snake genomes, support evolutionary studies and enable venom-driven drug discovery.

*1Molecular Biology Department, Genentech, Inc., South San Francisco, CA, USA. 2MedGenome Inc., Foster City, CA, USA. 3AgriGenome Labs Private Ltd, Kochi, India. 4SciGenom Research Foundation, Bangalore, India. 5Molecular Cytogenetics laboratory, Texas A&M University, College Station, TX, USA. 6Ecology and Evolution Unit, Okinawa Institute of Science and Technology, Onna-son, Japan. 7Department of Bioinformatics and Computational Biology, Genentech, Inc., South San Francisco, CA, USA. 8Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. 9Department of Microchemistry Proteomics, and Lipidomics, Genentech, Inc., South San Francisco, CA, USA. 10The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada. 11Division of Genomics, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan. 12Department of Pathology, Genentech, Inc., South San Francisco, CA, USA. 13College of Veterinary Medicine, Flow Cytometry Shared Resource Laboratory, Texas A&M University, College Station, TX, USA. 14Department of Early Discovery Biochemistry, Genentech, Inc., South San Francisco, CA, USA. 15Department of Molecular Biology, SciGenom Labs, Kochi, India. 16Kentucky Reptile Zoo, Slade, KY, USA. 17Department of Biological Sciences, National University of Singapore, Singapore, Singapore. 18Wayanad Wildlife Sanctuary, Sultan Bathery, India. *e-mail: sekar@sgrf.org
components is poorly understood. This is further exacerbated by the lack of access to antivenom and its high cost in many developing countries. Although several alternative approaches have been proposed, large animal-based antivenom production using extracted snake venom as the antigen continues to be the standard practice.

High-quality genomes of venomous snakes, combined with transcriptomics, will enable generation of a comprehensive catalog of venom-gland-specific toxin genes that can be used for the development of synthetic antivenom of defined composition using recombinant technologies. Thus far, only a few snake genomes have been published. A majority of these were generated primarily using short-read sequencing resulting in highly fragmented assemblies, thus limiting their utility for creating a complete catalog of venom-relevant toxin genes. The ‘big four’ medically important snakes from India, including *N. naja*, are no exception and have not been well characterized at either the genome or transcriptome level. Only nine *N. naja* toxin gene sequences and 38 toxins, some of which probably represent the same gene, have been reported using mass spectrometry.

In this study, using a number of genomic technologies, we have generated a de novo near-chromosome level reference genome assembly of *N. naja*, the Indian cobra. This high-quality genome allowed us to study various aspects of snake venom biology, including venom gene genomic organization, genetic variability, evolution and expression of key venom genes. Our integrated genome–transcriptome analysis identified a minimal set of 19 VST genes that constitute the core venom toxins. Targeting these core toxins should lead to the development of a safe and effective humanized antivenom.

**Results**

**Near-chromosomal de novo genome assembly.** Using flow cytometry, we estimated the size of the Indian cobra haploid genome at 1.48–1.77 Gb (Extended Data Fig. 1). Cytogenetic analysis revealed a diploid karyotype of 2n=38, comprising seven pairs of macrochromosomes (MACs), one pair of sex chromosomes (ZZ male orZW female) and 11 pairs of microchromosomes (MICs), consistent with a previous report (Extended Data Fig. 2).

DNA from an adult male Indian cobra (NN01) was used to generate long-read (PacBio and Oxford Nanopore Technologies (ONT)), short-read (Illumina), Chicago28, Hi-C3 and optical mapping (Bionano Genomics (BNG)) data (Supplementary Table 1a,b). Additionally, we generated linked-read 10x Genomics, BNG and short-read illumina sequence data for a female animal (NN05; Supplementary Note and Supplementary Table 1a). Our sequential assembly approach (Fig. 1 and Supplementary Note) resulted in a 1.79-Gb *Nana* y5 genome with a scaffold N50 of 223.35 Mb and a BUSCO30 genome completeness score of 94.3% (Table 1 and Supplementary Table 2a–c).

To assign scaffolds from *Nana* y5 to chromosomes, we used complementary DNA (cDNA) chromosome marker sequences from a colubrid, the Japanese rat snake *Elaphe quadrivirgata*11 (Fig. 1e), synteny information from *Anolis carolinensis* (the green anole lizard genome)12 and single-chromosome sequencing data (SCHROMseq; Fig. 1f, Supplementary Table 3a–c and Supplementary Note). We also generated a hybrid 10x BNG genome assembly to identify a 52.1-Mb female-specific W chromosome-linked scaffold (Supplementary Table 3d and Supplementary Note).

Comparison of the *N. naja* de novo genome assembly to the human13 and goat14 genomes showed that the scaffold N50 of the *N. naja* genome was 2.5x (87.27 versus 223.35 Mb) and 3.3x (67.79 versus 223.35 Mb) greater than the goat and human genomes previously assembled using the involved, expensive, time-consuming physical and cytogenetic maps developed over a decade or more (Table 2). Compared with the king cobra genome16, also an elapid, the Indian cobra genome contained far fewer scaffolds (296,399 versus 1,897, respectively), and had 929-fold better contiguity (scaffold N50 of 0.24 versus 223.35 Mb, respectively). Also, compared to the 7,034 scaffolds and 179.89-Mb scaffold N50 reported recently for the prairie rattlesnake genome (*Crotalus viridis*), the Indian cobra genome had a higher scaffold N50 and fewer scaffolds (Table 2).

**Genome features.** The average DNA base (GC) content of the *N. naja* genome was 40.46%. While MACs, representing 88% of the genome, had a GC content of 39.83%, that of MICs was 43.50% despite their containing only 12% of the genome (Welsh’s two-sample t-test, two-sided P<0.0001; Fig. 2a). Analysis of the repeat content in the Indian cobra genome in relation to other squamate reptile genomes revealed that 43.22% of the genome was repetitive (~760 Mb; Fig. 2a, Supplementary Table 4a,b, Extended Data Fig. 3 and Supplementary Note).

Whole-genome synteny comparison between the Indian cobra and prairie rattlesnake genomes revealed large syntenic blocks between the macro-, micro- and Z-chromosomes (Fig. 2b,c). We observed several fusion/fission events consistent with the difference in chromosome number between these two genomes. In particular, chromosome 4 of the Indian cobra shared syntenic regions with rattlesnake chromosomes 3 and 5, indicating a possible fusion event. In contrast, Indian cobra chromosomes 5 and 6 were syntenic to rattlesnake chromosome 5, indicating a possible fission event (Fig. 2b,c). Comparison of the Indian cobra genome to that of the more distantly related green anole lizard also showed regions of synteny and chromosomal rearrangements (Fig. 2c). Lizard chromosome 2 contained syntenic regions corresponding to chromosomes 4, 5 and 6 in the Indian cobra genome (Fig. 2c). Our synteny analysis also showed that the lizard chromosome 6 is homologous to the Z-chromosome of the Indian cobra (Fig. 2c and Supplementary Table 3b)35. Despite an estimated divergence time of ~280 Ma between snake and chicken (*Gallus gallus*), we observed synteny between several macro- and microchromosomes. Several regions of chicken chromosomes 1 and 2 showed syntenic blocks across Indian cobra macro-, micro- and Z-chromosomes (Fig. 2c). This indicated large-scale changes in macro- and microchromosome organization between squamate and avian genomes during evolution15,25.

**Gene prediction and annotation.** We used the MAKER pipeline46 to annotate the genome using protein homology information, in combination with gene expression data from 14 different tissues (n=26 samples; Figs. 2d and 3, Supplementary Table 1a and Supplementary Note). Overall, we predicted 23,248 protein-coding genes and 31,447 transcripts that included alternatively spliced products encoding 31,036 predicted proteins (Fig. 2d). Of the 23,248 genes, we found 22,116 (95%) on the 191 largest scaffolds corresponding to the numbered chromosomes. A total of 26,216 of the 31,036 predicted proteins (84.4%) contained a canonical start and stop codon. We identified 3,265 of 26,216 proteins (~12.5%) with an N-terminal secretion signal sequence, a feature important for venom gland toxin secretion (Supplementary Table 5). We performed extensive functional annotation of the 31,036 predicted proteins and found that 17,019 (54.8%) had a corresponding ortholog in either the Human Gene Nomenclature Committee database, NCBI’s non-redundant database or the TrEMBL (https://www.ebi.ac.uk/uniprot) database (Supplementary Tables 5 and 7a and Supplementary Note). Comparison of our annotated proteome to that of the king cobra13, prairie rattlesnake15 and green anole lizard genomes17 identified 26,323, 25,505 and 11,820 orthologs in those genomes, respectively.

We comprehensively annotated venom-gland-relevant genes by combining the predicted gene models with long-read sequencing data (Supplementary Table 6a and Supplementary Note), toxin gene hidden Markov models (HMM) and manual curation to identify 139 toxin genes from 33 gene families17 that included 19 three-finger toxins (3FTxs), 8 snake venom metalloproteinases (SVMPS) and 6 cysteine (Cys)-rich secretory venom proteins.
(CRISPs) (Supplementary Table 6b). Near-chromosomal assembly also allowed us to assess the genomic organization of gene families that encoded enzymatic and non-enzymatic toxin proteins involved in venom gland function. In the N. naja genome, 16 major toxin gene families were organized on MACs (Fig. 4a and Supplementary Table 6c). This is in contrast to the genomes of vipers C. viridis (prairie rattlesnake) and Protobothrops flavoviridis (Amami habu), where a majority of the venom gland genes were found on MICs\textsuperscript{23,35}.

**Fig. 1 | Schematic of N. naja genome sequencing and assembly.** a, b, Long-read (PacBio and ONT) and short-read (Illumina) data (a) were used to build contigs that were then combined with Chicago\textsuperscript{28} chromatin interaction data (b) to generate scaffolds. c, Scaffolds from BNG optical mapping de novo assembly were combined with those from the previous step to generate super-scaffolds. d, Hi-C sequencing data were used to refine the assembly. e, Electronic fluorescence in situ hybridization (eFISH) was performed using cDNA FISH marker sequences from the Japanese rat snake, E. quadrivirgata. f, SChrom-seq data were used to assign scaffolds to chromosomes.
Of the 19 full-length 3FTx genes identified in the genome, 14 were clustered within a 6.3-Mb region on chromosome 3 (Fig. 4b). One 3FTx gene (Nana001KS) was located on chromosome 4 and the remaining four were found on an unassigned scaffold, ScVE01q_1072 (Supplementary Table 6c). Additionally, we identified 10 3FTx pseudogenes that lacked parts of the coding region and were not expressed. The second largest toxin gene family encoded by the N. naja genome consisted of eight SVMPs that were clustered on a MIC 1 (Fig. 4c). A cluster of six CRISPs were found on chromosome 4 (Fig. 4d). Other toxin genes, including natriuretic peptide, nerve growth factor (NGF-β), β-hairpin loop that extends from a disulfide bridge46. Functionally, 3FTxs are broadly classified as neurotoxins, cytotoxins, cardiotoxins and anticoagulants.

While expression of all 19 annotated N. naja 3FTxs was detected in the venom gland, 9 were specific to the venom gland. Of these 19 3FTxs, 14 were classified as conventional 3FTxs with 8 conserved Cys residues, while 4 3FTxs contained 10 Cys residues. Homology-based assessment enabled classification of the 3FTxs into seven toxin gene families. Furthermore, differential expression analysis revealed a set of 109 genes from 15 different toxin gene families that were significantly upregulated in the venom gland (fold change >2 and 1% false discovery rate (FDR); Extended Data Fig. 5 and Supplementary Table 7c), and this included 19 toxin genes that were expressed exclusively in the venom gland (Fig. 4g and Supplementary Note). These 19 VST genes are likely to encode the core venom effector toxin proteins, consisting of six neurotoxins, one cytotoxin, one cardiotoxin, one murexicanin toxin, six SVMPs, nerve growth factor (NGF-β), two venom Kunitz serine proteases and a CRISP (Supplementary Table 6b, column L). Additionally, we confirmed the presence of 16 of the 19 VSTs at the protein level using mass spectrometry (Supplementary Table 8).

Functionally diverse 3FTxs. Three-finger toxins are short polypeptides (60–90 amino acids) that belong to a superfamily of non-enzymatic proteins found primarily in elapid snakes43. These small proteins are known to primarily target neuronal receptors including nicotinic acetylcholine receptors (nAChRs), muscarinic acetylcholine receptors, calcium channels and other proteins44. Structurally, they fold into an outstretched, three-finger-like structure where each finger contains a β-hairpin loop that extends from a disulfide bond-stabilized hydrophobic core44,45. 3FTxs typically contain four conserved disulfide bridges, with some containing a fifth disulfide bridge46. Functionally, 3FTxs are broadly classified as neurotoxins, cytotoxins, cardiotoxins and anticoagulants.

While expression of all 19 annotated N. naja 3FTxs was detected in the venom gland, 9 were specific to the venom gland. Of these 19 3FTxs, 14 were classified as conventional 3FTxs with 8 conserved Cys residues, while 4 3FTxs contained 10 Cys residues. Homology-based assessment enabled classification of the 3FTxs into seven neurotoxins, six cytotoxins, four cardiotoxins, one murexicanin toxin and one anticoagulant (Supplementary Table 9). The neurotoxins included two type I short-chain neurotoxins (Nana002KS and Nana005KS), known to interact with muscle nAChR, one type II long-chain neurotoxin (Nana012KS), known to target muscle and neuronal nAChRs, and three putative type III weak neurotoxins containing 10 conserved cysteines (Nana001KS, Nana003KS and Nana004KS), a characteristic feature of both non-conventional and prey-specific 3FTxs47,48. Nana018KS was structurally similar to

---

**Table 1 | N. naja assembly summary statistics**

| Assembly         | Contigs (n) | Scaffolds (n) | Gaps (Gb, %) | Contig N50 (Mb) | Scaffold N50 (Mb) | Assembly size (Gb) |
|------------------|-------------|---------------|--------------|-----------------|-------------------|-------------------|
| Nana_v1 (long-read (LR)) | 13,066      | -             | -            | 0.31            | -                 | 1.66              |
| Nana_v2 (LR + Chicago) | 13,066      | 2,676         | 0.01 (0.6)   | 0.31            | 4.90              | 1.67              |
| Nana_v3 (optical map (OM)) | -          | 1,477         | -            | -               | 16.30             | 1.63              |
| Nana_v4 (v3 + OM) | 13,066      | 2,167         | 0.11 (6.4)   | 0.31            | 157.50            | 1.79              |
| Nana_v5 (v4 + Hi-C) | 13,066      | 1,897         | 0.11 (6.4)   | 0.30            | 223.35            | 1.79              |

**Table 2 | Comparison of Nana_v5 assembly to other high-quality genomes**

| Assembly            | Human GRCh38 | Goat ARS1 | King cobra | Five-pace viper | Indian cobra |
|---------------------|--------------|-----------|------------|-----------------|--------------|
| Total sequence length (Gb) | 3.2          | 2.9       | 1.66       | 1.47            | 1.79         |
| Total assembly gap length (Mb) | 160          | 38        | 210        | 82              | 110          |
| Number of scaffolds | 735          | 29,907    | 296,399    | 160,256         | 1,897        |
| Number of scaffolds >10 Mb (% of assembly) | 23 (99.8) | 31 (90.1) | 35 (2.7) | 412 (72.0) | 19 (94.9) |
| Scaffold N50 (Mb) | 67.79        | 87.27     | 0.24       | 212             | 223.35       |
| Number of chromosomes | 23           | 31        | 18         | 18              | 19           |

---

**Minimal core venom-ome toxin genes.** Analysis of multi-tissue transcriptome data from 26 samples representing 14 different tissues (Supplementary Table 1a) identified 19,426 expressed genes (counts per million (CPM) >1; Fig. 3 and Supplementary Table 7a,b), of which 6,601 common core genes were expressed across all tissues (Supplementary Note).

The venom gland transcriptome (venom-ome) comprised 12,346 expressed genes and included 139 genes from 33 different toxin gene families. Furthermore, differential expression analysis revealed a set of 109 genes from 15 different toxin gene families that were significantly upregulated in the venom gland (fold change >2 and 1% false discovery rate (FDR); Extended Data Fig. 5 and Supplementary Table 7c), and this included 19 toxin genes that were expressed exclusively in the venom gland (Fig. 4g and Supplementary Note). These 19 VST genes are likely to encode the core venom effector toxin proteins, consisting of six neurotoxins, one cytotoxin, one cardiotoxin, one murexicanin toxin, six SVMPs, nerve growth factor (NGF-β), two venom Kunitz serine proteases and a CRISP (Supplementary Table 6b, column L). Additionally, we confirmed the presence of 16 of the 19 VSTs at the protein level using mass spectrometry (Supplementary Table 8).
**Fig. 2 | Genome architecture of N. naja genome.**

**a.** Circos plot of the reference genome assembly (NN01, a male adult (n = 1)) representing N. naja chromosomes (outermost track) from the Nana_v5 assembly, repeat content, gene density and GC content (%). Regions of the genome with GC content higher than average (40.46%) are shown in blue. Regions within the gene density of more than 10 genes are shown as red spikes, while those with 5 to 10 genes are indicated by yellow spikes. Green spikes represent regions with fewer than five genes. The average repeat content is indicated by the red line. All data were plotted in 100-kb windows. The female-specific W-linked scaffold obtained using NN05 DNA is shown on the right.

**b.** Chromosome painting depicting synteny between Indian cobra and rattlesnake genomes.

**c.** Dot-plots showing synteny of the Indian cobra genome with the prairie rattlesnake, chicken or green anole lizard genomes.

**d.** Bar plot of the number of predicted genes and corresponding transcripts observed in Nana_v5. Dashed and solid lines denote average number of genes and transcripts detected in each chromosome along 100-Mb windows, respectively. MICs were combined into one group. Unp, unplaced scaffolds (n = 1,878) containing predicted genes.
haditoxin from the king cobra and known to block $\alpha_7$-nAChRs\(^{19}\). Further assessment by structural modeling classified the Indian cobra 3FTxs into four groups (Fig. 5, Extended Data Fig. 6 and Supplementary Table 10). The aromatic residue (Tyr25 or Phe27), crucial for proper folding\(^{50}\) and stability of the antiparallel $\beta$-sheet structure, was found to be conserved in all 19 3FTxs (Fig. 5a). Three of the four 10-Cys-residue-containing 3FTxs were non-conventional 3FTxs containing a pair of additional cysteines in loop I,

Fig. 3 | *N. naja* expression body map. Heatmap showing log$_2$(CPM) values of differentially upregulated genes (DUGs); FDR < 1% across 14 tissues (sample size $n=6$) as indicated. NN01 and NN02 correspond to *N. naja* specimens obtained from Kerala, India. NN03, NN04, NN05 and NN06 correspond to *N. naja* specimens obtained from the Kentucky Reptile Zoo. Sg, salivary gland.
resulting in a longer N terminus loop that could potentially also play a role in stabilizing this loop and contribute to toxin function\(^4^7\) (Fig. 5a,b). The other 10-Cys 3FTx was a long-chain neurotoxin that contained a pair of additional cysteines in loop II (Nana012KS). The charged amino acid residue Arg39, which typically stabilizes native toxin conformation by forming a salt bridge with the C terminus\(^5^1\),
Fig. 5 | Characterization of N. naja 3FTx gene family. a, Multiple sequence alignment of the 19 3FTx proteins identified in the Indian cobra genome. Protein names in orange in the alignment indicate VSTs identified using RNA-seq. Conserved Cys residues are highlighted yellow in the alignment. b–e, Ribbon representations of representative 3FTxs from four different structural classes. Disulfide bonds are shown as sticks. The hydrophobic packing of Leu39 and surrounding residues is shown in c. Dashed circles highlight the additional disulfide bonds in Nana001KS, Nana012KS and the unpaired Cys in Nana005KS. f, Superimposition of ribbon models of Nana001KS, Nana003KS and Nana010KS highlighting the differences in loop length and conformation between the distinct classes of 3FTx found in the Indian cobra genome. g, Analysis of evolutionary rates on 3FTx venom genes and their non-venom paralogs. Ks and Ka values were calculated according to the Nei–Gojobori method. Ks and Ka with values <1 were not included in further analysis. SNTX, short neurotoxin; LNTX, long neurotoxin; MTLP, muscarinic toxin-like; CTX, cardiotoxin or cytotoxin; Nonc, non-conventional toxin; WTX, weak neurotoxin. SV, snake venom; NV, non-venom.
was conserved in all identified 3FTxs except in Nana012KS, where it was replaced by a leucine residue (Fig. 5a). The homology model of this protein indicated that Leu39 may exhibit van der Waals interactions and form part of a hydrophobic core comprising Ile35, Phe4, Thr22 and Arg68 (Fig. 5c). Two of the cytotoxins, Nana008KS and Nana010KS, had shorter loops I and III compared to the other 3FTxs (Fig. 5d).

Of note, Nana005KS contained nine Cys residues and such unusual 3FTxs have been found in only two other elapids, Micrurus lemniscatus and Micrurus altirostris52. This free Cys residue at position 16 that precedes the conserved second Cys in loop I (Fig. 5b) probably facilitates the formation of covalent homo- or heterodimeric 3FTxs (Supplementary Table 10). Nana005KS was closely related to the short-chain 3FTx Nana002KS and contained a mutation of the residues required for activity against muscle nAChRs. For instance, the presence of the positively charged residues Lys25, Lys26 and Arg32 indicates that these toxins might be crucial for envenomation in mammals. In particular, the guanidyl group of Arg32 mimics acetylcholine forming a cation–π interaction with α-α and α-δ interfaces in the nAChRs53.

Acquired mutations affecting the neurotoxin binding site of nAChR have rendered certain species, such as the Egyptian mongoose (Herpestes ichneumon), immune to snake venom54–56. Comparison of N. naja nAChR sequence and H. ichneumon nAChR and other representative mammalian species showed that the N. naja nAChR carries a key p.Phe189Asn alteration in the α-neurotoxin site, known to result in decreased sensitivity to short- and long-chain neurotoxins54,55 (Extended Data Fig. 7).

In the final group of 3FTxs, Nana013KS was structurally similar to AncTx-I (ref. 57), a synthetic ancestral toxin known to interact with β-adrenergic G-protein-coupled receptors. The 3FTx encoded by Nana006KS was found to be a close homolog of ringhalexin, an inhibitor of the extrinsic tenase coagulation complex, from Hemachatus haemachatus, the African ringhals cobra58 (Fig. 5e and Supplementary Table 9). Additionally, we found Nana017KS to be 97% similar to a recently reported Naja atra (μ-EPTX-Naja)59 Nav1.8 voltage-gated sodium channel inhibitor.

To understand the consequences of genetic variation in the 3FTx family, we assessed the rate of evolution of the major toxin families by computing the numbers of synonymous (Ks) and non-synonymous (Ka) nucleotide substitutions per site for each pairing of toxin gene and its non-toxin paralog. The Ka/Ks substitution ratio for the 3FTx toxin genes was 2.034 (±0.818), while that of the non-toxin paralogs (Ly-6/UPAR domain-containing genes) was 0.894 (±0.103; Fig. 5g). The observed high Ka/Ks ratio (>1) indicated diversifying selection leading to rapid divergence and functional diversification of the venom-gland-specific 3FTx genes.

Indian cobra SVMP, CRISPs, PLA2, CVF and growth factor genes. We detected six venom-gland-specific SVMPs that belonged to the P-III class of metalloproteinases (with metalloproteinase/disintegrin/Cys-rich domains) known to be involved in the induction of hemorrhage, inflammation, apoptosis, prothrombin activation and inhibition of platelet aggregation60. The Indian cobra SVMPs were found to evolve less rapidly than 3FTx genes, because the Ks/Ka ratio for the venom-gland-specific SVMPs was 1.070 (±0.137) when compared to 0.998 (±0.049) observed for metalloprotease domain-containing paralogs (Extended Data Fig. 8). The Indian cobra SVMPs formed a separate cluster compared with the viperid SVMPs (Extended Data Fig. 9). In agreement with this, a seventh Cys residue in domain M12, involved in the disulfide bond exchange for autolysis during secretion or formation of the biologically active disintegrin/Cys-rich domain typical of Viperidae SVMP-P1ILIs61, was absent (Extended Data Fig. 9).

We also detected six CRISPs in the venom gland transcriptome that were highly conserved across different snake species (Supplementary Fig. 1). Venom CRISPs have a wide variety of biological effects, including blockade of K+ and/or Ca2+ currents in neurons and blockage of vascular smooth muscle contraction62. Two of the five CRISPs were homologous to venom CRISPs from the Chinese cobra, Naja atra (natrin), and the monocled cobra, Naja kaouthia (UniProtKB: Q7T1K6). Expression of the N. naja natrin homolog (Nana02866) was venom gland specific, and it probably functions as a K1.3 potassium channel blocker62.

Additionally, two group I secretory acidic PLA2 genes (Nana39244 and Nana39246) that were highly expressed in the venom and salivary glands showed a high degree of similarity to other elapid venom PLA2s, and contained the characteristic calcium-binding (XCGXXG) and catalytic (DXCCXXHD) motifs63 (Supplementary Fig. 2 and Supplementary Note).

In addition to the major elapid toxin families described above, we detected transcripts from other toxin gene families including hyaluronidases, phospholipase B-like genes, calcidinics, ohanins (known to induce hypolocomotion and hyperalgesia64) and 5’ nucleotidases (Supplementary Note). We also detected l-amino acid oxidase (LAAO) (Nana07S85), which is involved in platelet aggregation, edema and hemorrhage. Furthermore, we identified two full-length c-type natriuretic peptide genes, Nana20849 and Nana20852, in the venom-ome. Of the three Kunitz serine protease inhibitors detected in the venom-ome, two were VST gene products (Nana13192 and Nana13193) and these probably function to inhibit serine proteases acting in the hemostatic system65. In addition, two full-length cystatin genes, Nana15338 and Nana35841, were also found expressed in the venom-ome.

Cobra venom factor is a non-lethal protein that resembles the complement C3 protein in structure and function66. Previously, the complete structure of one CVF gene from N. Naja has been reported67. In the present study, we identified three CVF genes (Nana10828, Nana38416 and Nana10826) in the N. naja genome. Nana38416 and Nana10828 contained 40 exons and spanned ~118 and ~75 kb, respectively, on chromosome 2. Isolomforming (Iso-seq) data confirmed the expression of the full-length transcripts corresponding to Nana10828 and Nana38416. Though the S’ genomic structure of Nana10826 was not fully resolved in the currently assembly, the expression of the full-length transcript of Nana10826 was confirmed by iso-seq. Protein sequence alignment showed that Nana38416 was 96% similar to CVF (UniProtKB: Q91133) from N. kaouthia, while Nana10828 was 99% similar to a previously characterized N. naja C3 complement component protein68.

In addition to toxin components, we also identified four full-length PDGF/VEGF-growth factor genes including vascular endothelial growth factor-1 (VEGF1; Nana01393), VEGFC (Nana18254), platelet-derived growth factor (PDGF; Nana34300), placenta growth factor (PGF; Nana05327) and an insulin growth factor (IGF) gene and Nana04360 in the venom-ome. Furthermore, we also identified a venom-gland-specific nerve growth factor (NGF-β; Nana13949) that exhibited high homology to NGFV2 from the spitting cobra Naja sputratrix.

Discussion

Much of our current understanding of snake venom is based on proteomic studies that have provided only a partial picture of its components69–71. A comprehensive catalog of venom proteins, their expression and coding sequence is fundamental to developing a safe and effective antivenom65,72. Also, such a detailed catalog of venom components will be valuable for drug candidate prospecting.

Using next-generation sequencing in combination with emerging genomic technologies, we have generated a de novo high-quality N. naja reference genome73. The near-chromosomal assembly revealed regions of synteny between reptilian and avian genomes, consistent with their evolutionary trajectories. The high contiguity of our genome enabled visualization of the striking differences in venom gene organization between elapid and viperid snake
The locations of the major toxin gene families in the *N. naja* genome on MACs was in contrast to their presence on MICs in *P. flavoviridis* and *C. viridis* genomes, indicative of the differences in their chromosome and venom evolution.

Overall, we found evidence for expression of 12,346 genes that constitute the venom-ome, and this included 139 toxin genes of which 19 were designated as VSTs based on their venom-gland-specific expression. Additionally, well-known modulators of venom function such as CVF, coagulation factors, protein disulfide isomerases, natriuretic peptides, hyaluronidases, PLA2s, phospholipase B-like genes, LAAO, vascular endothelial growth factor (VEGF) and 5’ nucleotidases were also found to be highly expressed in the venom gland. It is likely that these genes, together with the 19 VSTs, form the core toxic effector components of the venom and induce a wide range of symptoms including cardiovascular dysfunction, muscular paralysis, nausea, blurred vision and systemic effects such as hemorrhage (Fig. 6). We propose that neutralization of these core venom effectors using antibodies would be an effective therapeutic strategy. Furthermore, given the variation in venom composition, cataloging the venom gland gene repertoire and its variation (Extended Data Fig. 10 and Supplementary Note), both within and across different snake species, will be important for developing a broadly efficacious antivenom.

Snake antivenom is made by immunizing large mammals, such as horses, with extracted snake venom as antigen. Such horse-derived antibodies show variation in efficacy due to expected differences in the horse antibody response following antigen challenge. Moreover, the heterologous nature of the antibodies leads to adverse treatment-related side effects. Starting with a complete catalog of VSTs, synthetic venom of a defined composition can be generated using recombinant protein expression technologies. Such a cocktail of recombinant core venom VST proteins can be used to raise
specific antibodies in horses, tested for neutralizing activity, rapidly cloned/synthesized and humanized to produce the next generation of antivenoms\textsuperscript{3,4,6,7}. Alternatively, recombinant venom proteins can be used as baits against antibody phage libraries to obtain toxin-neutralizing, activity-tested humanized synthetic antibodies\textsuperscript{8,9,10}.

Genome-guided recombinantly produced venom proteins can be also used to characterize and improve existing horse-derived anti-venom. In this paradigm, antibodies can be raised against recombinant venom toxins that are non-immunogenic in horses and used to supplement current antivenom cocktails to improve their efficacy. Alternatively, single-cell sequencing of B cell repertoires from horses immunized with extracted or synthetic venom can be performed to identify putative neutralizing antibodies\textsuperscript{11}. This information can be used to rapidly synthesize, humanize and identify toxin-neutralizing antibodies for the generation of synthetic antivenom.

As more high-quality snake genomes are completed and venom gland VSTS are cataloged, synthetic antibodies directed against key species-specific toxins identified from such genome initiatives can be combined to create potent broad-spectrum antivenom\textsuperscript{12}. Furthermore, antibodies targeting different epitopes on recombinant toxins can be developed using phage display. These recombinant toxin-directed antibodies can be combined to generate an antivenom that is likely to be more efficacious\textsuperscript{13–16}.

Primary cultures of venom gland cells\textsuperscript{17} or the recently developed venom gland organoid cultures (Post, Y. et al., personal communication), in combination with genomic information, can provide an alternative, viable source of venom antigens for antivenom development (Supplementary Note).

We believe the Indian cobra reference genome and the analysis presented here will facilitate innovations in antivenom development. The genome and the associated predicted proteome will serve as a powerful platform for evolutionary studies of venomous organisms. More importantly, the comprehensive catalog of the venom proteins presented here should enable drug development, in particular to treat hypertension, pain and other disorders\textsuperscript{18–20}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-019-0559-8.

Received: 10 July 2019; Accepted: 22 November 2019; Published online: 6 January 2020

References

1. Hsiang, A. Y. et al. The origin of snakes: revealing the ecology, behavior, and evolutionary history of early snakes using genomics, phenomics, and the fossil record. \textit{BMC Evol. Biol.} \textbf{15}, 87 (2015).
2. Fry, B. G. & Wuster, W. Assembling an arsenal: origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences. \textit{Mol. Biol. Evol.} \textbf{21}, 870–883 (2004).
3. Fry, B. G. From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. \textit{Genome Res.} \textbf{15}, 403–420 (2005).
4. Zaher, H. et al. Large-scale molecular phylogeny, morphology, divergence-time estimation, and the fossil record of advanced caenophidian snakes (Squamata: Serpentes). \textit{PLoS ONE} \textbf{14}, e0216148 (2019).
5. Pyron, R. A., Burbrink, F. T. & Wiens, J. J. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. \textit{BMC Evol. Biol.} \textbf{13}, 93 (2013).
6. Gutierrez, J. M. et al. Snakebite envenoming. \textit{Nat. Rev. Dis. Prim.} \textbf{3}, 17079 (2017).
7. Mohapatra, B. et al. Snakebite mortality in India: a nationally representative mortality survey. \textit{PLoS Negl. Trop. Dis.} \textbf{5}, e0108 (2011).
8. Aird, S. D. et al. Snake venoms are integrated systems, but abundant venom proteins evolve more rapidly. \textit{BMC Genomics} \textbf{16}, 647 (2015).
9. Amazonas, D. R. et al. Molecular mechanisms underlying intraspecific variation in snake venom. \textit{J. Proteomics} \textbf{181}, 60–72 (2018).
10. Casewell, N. R., Wuster, W., Vonk, F. J., Harrison, R. A. & Fry, B. G. Complex cocktails: the evolutionary novelty of venoms. \textit{Trends Ecol. Evol.} \textbf{28}, 219–229 (2013).
11. Chippaux, J. P., Williams, V. & White, J. Snake venom variability: methods of study, results and interpretation. \textit{Toxicon} \textbf{29}, 1279–1303 (1991).
12. Calmette, A. The treatment of animals poisoned with snake venom by the injection of antivenomous serum. \textit{Br. Med. J.} \textbf{2}, 399–400 (1896).
13. de Silva, H. A., Ryan, N. M. & de Silva, H. J. Adverse reactions to snake antivenom, and their prevention and treatment. \textit{Br. J. Clin. Pharmacol.} \textbf{81}, 446–452 (2016).
14. Laustsen, A. H., Johansen, K. H., Engmark, M. & Andersen, M. R. Recombinant snakebite antivenoms: a cost-competitive solution to a neglected tropical disease. \textit{PLoS Negl. Trop. Dis.} \textbf{11}, e005361 (2017).
15. Bermudez-Mendez, E. et al. Innovative immunization strategies for antivenom development. \textit{Toxins} (Basel) \textbf{10}, 452 (2018).
16. Kini, R. M., Sidhu, S. S. & Laustsen, A. H. Biosynthetic oligoclonal antivenom (BOA) for snakebite and next-generation treatments for envenoming victims. \textit{Toxins} (Basel) \textbf{10}, 534 (2018).
17. Harrison, R. A. et al. Research strategies to improve snakebite treatment: challenges and progress. \textit{J. Proteomics} \textbf{74}, 1768–1780 (2011).
18. Richard, G. et al. In vivo neutralization of alpha-cobratoxin with high-affinity llama single-domain antibodies (VHVs) and a VH-Hc antibody. \textit{PLoS ONE} \textbf{8}, e69165 (2013).
19. Vonk, F. J. et al. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. \textit{Proc. Natl Acad. Sci. USA} \textbf{110}, 20651–20656 (2013).
20. Castoe, T. A. et al. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. \textit{Proc. Natl Acad. Sci. USA} \textbf{110}, 20645–20650 (2013).
21. Yin, W. et al. Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer vipers. \textit{Nat. Commun.} \textbf{7}, 13107 (2016).
22. Pasquesi, G. I. M. et al. Squamate reptiles challenge paradigms of genomic completeness with single-copy orthologs. \textit{Mol. Biol. Evol.} \textbf{31}, 2836–2846 (2014).
23. Kerkkamp, H. M. et al. Snake genome sequencing: results and future prospects. \textit{Toxins} (Basel) \textbf{8}, 360 (2016).
24. Kalita, B. & Mukherjee, A. K. Recent advances in snake venom proteomics research in India: a new horizon to decipher the geographical variation in venom proteome composition and exploration of candidate drug prototypes. \textit{J. Proteomics} \textbf{10}, 149–164 (2019).
25. Singh, L. Evolution of karyotypes in snakes. \textit{Chromosoma} \textbf{38}, 185–236 (1972).
26. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. \textit{Genome Res.} \textbf{26}, 342–350 (2016).
27. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. \textit{Science} \textbf{326}, 289–293 (2009).
28. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. \textit{Bioinformatics} \textbf{31}, 3210–3212 (2015).
29. Matsubara, K. et al. Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes. \textit{Proc. Natl Acad. Sci. USA} \textbf{103}, 18190–18195 (2006).
30. Allfeld, J. et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. \textit{Nature} \textbf{477}, 587–591 (2011).
31. International Human Genome Sequencing Consortium. \textit{Finishing the euchromatic sequence of the human genome}. \textit{Nature} \textbf{431}, 931–945 (2004).
32. Rickman, S. et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. \textit{Nat. Genet.} \textbf{49}, 643–650 (2017).
33. Schmidt, D. R. et al. The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. \textit{Genome Res.} \textbf{29}, 590–601 (2019).
34. Srikulnath, K. et al. Karyotypic evolution in squamate reptiles: comparative gene mapping revealed highly conserved linkage homology between the butterfly lizard (\textit{Leiolepis revesui rubrataeniata}, Agamidae, Lacertilia) and the Japanese four-striped rat snake (\textit{Elaphe quadrivirgata}, Colubridae, Serpentes). \textit{Chromosomes Res.} \textbf{17}, 975–986 (2009).
35. Kawai, A. et al. Different origins of bird and reptile sex chromosomes inferred from comparative mapping of chicken Z-linked genes. \textit{Cytogeten. Genome Res.} \textbf{117}, 92–102 (2007).
We estimated genome size by (karyoMax, Thermo Fisher) in GURR buffer (Gibco). At least 30 metaphase spreads on glass slides at room temperature. Chromosomes were stained with 5% Giemsa from both lymphocyte and fibroblast cultures by arresting cells with demecolcine were observed based on the amount of PI absorbed by the cells of either species, genome of 2.5 Gb (EquCab3, GCA_002863925.1)91. Amount of PI absorbed by cobra cells, PI (horse) denotes the median amount of PI absorbed by cobra cells, PI (thermo) denotes the median amount of PI absorbed by cells of either species, and genome size was estimated using the formula: genome size (cobra) = PI (cobra)/PI (horse) × genome size (horse), where PI (cobra) denotes the median amount of PI absorbed by cobra cells, PI (horse) denotes the median amount of PI absorbed by cells, and genome size (horse) is the expected size of the horse genome of 2.5 Gb (EquCab3, GCA_002863925.1)91.

Samples and nucleic acid preps. A total of six animals—two from Kerala, India (NN01 and NN02) and four unrelated animals from the Kentucky Reptile Zoo, KY (NN03, NN04, NN05 and NN06)—were part of the study (Supplementary Table 1a). Two animals from India were fresh road kills that were submitted for post-mortem analysis to the Wayanad wildlife sanctuary forest veterinary officer, and were permitted to be used for this study under order no. WL10–12401/2017 from the Kerala forest and wildlife chief warden. A summary of tissues collected and processed from each animal is shown in Supplementary Table 1a. Genomic DNA was prepared using the Qiagen MagAttract HMW DNA kit (Qiagen). RNA was extracted using the Qiagen RNeasy kit (Qiagen).

Methods

Karyotyping. Briefly, blood from a female animal (NN03) was collected in a sodium-heparin vacutainer (Becton–Dickinson) and used for short-term (72-h) lymphocyte cultures as previously described92. We used phytohemagglutinin (PHA from Phaseolus vulgaris, 20 μg ml⁻¹ Sigma Aldrich) as the mitogen. Additionally, we established primary fibroblast cultures under sterile conditions using small pieces (0.5 mm²) of ovarian tissue from NN03. The fibroblasts were incubated in MEM alpha containing nucleosides and Glutamax (Thermo Fisher), supplemented with 20% fetal bovine serum (Atlanta Biologicals) and antibiotic-antimycotic (Thermo Fisher) at 30 °C with 5% CO₂. Metaphase chromosomes were obtained from both lymphocyte and fibroblast cultures by arresting cells with demecolcine (KaryoMax, Thermo Fisher; final concentration, 0.1 μg ml⁻¹), followed by hypotonic treatment with Optimal Hypotonic Solution (Rainbow Scientific) and fixation in methanol/acetic acid (3/1). Metaphase spreads were prepared on prewetted clean glass slides at room temperature. Chromosomes were stained with 5% Giemsa from both lymphocyte and fibroblast cultures by arresting cells with demecolcine were captured and analyzed for karyotyping using an Axioplan2 microscope (Zeiss) and Ikaros (MetaSystems) software.

Flow cytometry-based genome size estimation. We estimated genome size by flow cytometry using a previously described protocol98. Whole blood cells of the Indian cobra (NN03) were cultured for 4 d in a medium consisting of MEM alpha with nucleosides and Glutamax (Thermo Fisher), 30% fetal bovine serum (Atlanta Biologicals) and 1% PHA (1 mg ml⁻¹ Sigma Aldrich) at 30 °C. Blood cells were washed six times in sterile water and then fixed in cold ethanol. Horse (Equus caballus) lymphocytes from a female were harvested from a lymph node with a Lymphoprep (Stemcell Technologies) before being cleaned and fixed in the same manner as the Indian cobra cells. Next, both cell types were diluted to a concentration of 1×10⁶ cells ml⁻¹ and stained with propidium iodide (PI, Thermo Fisher). After staining, the Indian cobra and horse cells were mixed 1:1 and analyzed on a Becton–Dickinson Accuri C6 personal flow cytometer. Two peaks were observed based on the amount of PI absorbed by the cells of either species, and genome size was estimated using the formula: genome size (cobra) = PI (cobra)/PI (horse) × genome size (horse), where PI (cobra) denotes the median amount of PI absorbed by cobra cells, PI (horse) denotes the median amount of PI absorbed by cells, and genome size (horse) is the expected size of the horse genome of 2.5 Gb (EquCab3, GCA_002863925.1)91.

10x Genomics library generation and sequencing. High-molecular-weight DNA was extracted from a female (NN05) ovary using the 10x Chromium HMW DNA extraction protocol (10x Genomics). With 1 ng of template, DNA whole-genome sequencing libraries were prepared using the Chromium Genome Library and Gel Bead Kit v2 (10x Genomics, no. 120258), Chromium Genome Chip Kit v2 (10x Genomics, no. 120257), Chromium i7 Multiplex Kit (10x Genomics, no. 120262) and Chromium controller (10x Genomics). In total, 761.34 million linked-reads (114 Gb, 2×150-bp) were generated providing a raw coverage of ~65x.

Scaffold to chromosome mapping. We used the synteny between Elaphe chromosomes and the N. naja Nana_v5 genome to assign scaffolds to chromosomes. Briefly, 105,334 DNA sequences corresponding to E. quadrivirgata chromosomes93 were downloaded from the NCBI nucleotide database. The DNA sequences were aligned to the N. naja Nana_v5 genome assembly using Enxonerate94. The chromosome number corresponding to the Elaphe cDNA marker was assigned to the best matching Nana_v5 scaffold based on the highest sequence alignment score.

Snake chromosome laser capture microdissection and ScHrom-seq. Snake chromosome laser microdissection were prepared on Leica 0.9-μm-POL-membrane frame steel frame slides (Leica, no. 11050188) following the manufacturer’s recommended protocol. About 10 μl of fixed and previously tested metaphase cell suspensions from ovary (NN03) fibroblasts was dropped on a dry POL-membrane slide, allowed to spread by gravity and dried overnight. Air-dried slides were placed in 50 ml of Giemsa solution (1 ml of 5% Giemsa staining solution per 50 ml of GURR buffer (Gibco)) for 10 min, washed three times in sterile water and air-dried. Laser focal confocal microscopy-enabled dissection of individual chromosomes was done with the Leica LM16 Laser Microdissection microscope (150x dry-immersion objective; HC PL FLUOTAR ×150/0.9 numerical aperture) and software. A total of 104 dissections were performed. Each sample tube containing microdissected chromosomes was lyzed to release genomic DNA, and whole-genome amplification performed to amplify the DNA and create libraries with the SMARTer PicoPLEX DNA-seq kit (Takara Bio). The resulting 250-bp single-end libraries were further amplified exponentially with primers containing unique illumina dual-index barcodes suitable for illumina sequencing. Data from each library were mapped to a combined reference genome including GRCN38 and snake (Nana_v5) using Burrows–Wheeler aligner (BWA)95 with default options. Mapped reads were sorted and duplicates marked with PicardTools (broadinstitute.github.io/picard/). Reads mapping to multiple loci, GRCh38, a combined bacterial sequence database or those with a mapping quality below 10 were discarded. All remaining mapped reads were then used to calculate total counts per chromosome. Coverage data were generated with GATK96 and binned using 100-kb windows.

Synten mapping and dot-plot analysis. The repeat-masked C. viridis (prairie rattlesnake) and A. carolinensis genomes were aligned to Nana_v5 chromosomes with CoGe’s SynMap program using LAST95. The 10x-BNG hybrid scaffolds and Nana_v5 scaffolds (>1 Mb) were aligned with Sympa (v 4.2) using default parameters97. Chromosome painting of N. naja chromosomes with prairie rattlesnake chromosomes was performed using the SatsumaSynteny2 script with default parameters97.

RNA-sequencing. Ribonuclease acid was extracted from the venom gland, accessory gland, heart, lung, spleen, brain, ovary, testes, gall bladder, pancreas, kidney, liver, eye and tongue (Supplementary Table 1a) using the Qiagen RNeasy kit (Qiagen).
PolyA RNA-sequencing (RNA-seq) libraries were prepared with 1 μg RNA from each tissue using the Illumina TruSeq stranded messenger RNA kit and sequenced on HiSeq (Illumina).

Iso-seq analysis. Total RNA from venom gland was used to generate cDNA and PacBio iso-seq libraries as per the manufacturer’s instructions. Size selection of libraries was performed using the BluePippin system (Sage Science). Sequencing was performed on a PacBio RSII/Sequel (Supplementary Table S5a). Full-length consensus isoform sequences were generated using PacBio’s SMRT portal and SMRTLink for the RSII and Sequel data, respectively. Arrow (https://github.com/PacificBiosciences/GenomicConsensus) was used to polish the consensus isoforms. In total, 101,761 transcript isoforms were processed by the pipeline. Full-length transcripts (that contained a complete open reading frame) were then aligned to the genome using GMAP with parameters specific to long-read alignment15. Complete open reading frames were directly annotated with BLASTX (v.2.2.29+)16 against NCBI NR and TrEMBL using the same pipeline as done with whole-genome annotation. We used these data to manually verify and correct venom-gland-specific toxin gene annotations (Supplementary Note).

Repeat element identification. We identified the repetitive elements in the genome by combining both homology-based and de novo predictions. Next, we used a previously described reptile-specific repeat library with RepeatMasker (v.4.0.7) (http://www.repeatmasker.org) to annotate repetitive elements in the Indian cobra genome. We then used RepeatModeler (v.1.0.11) (http://www.repeatmodeler.org/RepeatModeler.html) to identified repeat sequence libraries for the Indian cobra, and then used these as a query to identify repetitive elements using RepeatMasker. Finally, we retrieved a non-redundant annotation for each species after combining all the annotation results using the reptile-specific libraries and de novo repeat library.

Venom gene synteny analysis. BLASTn was used to identify homologs of all N. naja toxin genes (Supplementary Table 6b) in the Anolis and C. viridis genomes. BLASTn hits were then filtered using the following parameters: query coverage ≥70% and identity ≥80%. Synteny was then plotted using a publicly available Python script for gene synteny visualization (https://github.com/biopython/biopython/blob/master/Doc/examples/Proux_et_al_2002_Figure_6.py).

Differential gene expression analysis. After adapter trimming and quality filtering, reads were aligned to Nana_v5 using STAR2 (v.2.6.0a)17 with default parameters. Gene counts per gene across all samples were calculated using featureCounts101. Raw gene counts were then normalized between samples using the CPM normalization method from the EdgeR package (Supplementary Table 7b–h). Differential expression analysis was then performed using EdgeR (fold change ≥2 and 1% FDR), and differentially upregulated genes (DUGs) were filtered for further analysis for each tissue that had a replicate. Gene Ontology-ontology enrichment analysis of DUGs was performed using EnrichR13,14 (Supplementary Note).

Liquid chromatography–tandem mass spectrometry analysis of venom. Pooled Indian cobra venom samples obtained from Kentucky Reptile Zoo (10 μg per lane) were reduced with DTT (10 mM, 60 min at 37°C), alkylated with iodoacetamide (20 mM, 15 min at room temperature) and separated by SDS–PAGE for ~1 cm on a 4–12% Bis-Tris gel. The gel was Coomassie stained with SimplyBlue (Invitrogen) and each lane cut into two segments based on the ~70–kDa molecular weight marker. Gel pieces were destained in MilliQ water twice and dehydrated with NH4HCO3/50% acetonitrile (ACN) (20 min) and 100% ACN (2×5 min). Digestion was performed overnight at 37 °C with trypsin (20 ng μl–1 in 50 mM NH4HCO3). Peptides were extracted twice in 1% formic acid (FA)/50% ACN, once with 100% ACN and then dried to completion. The sample was then reconstituted in solvent A (2% ACN/0.1% FA) and injected onto a 1.0 × 100-mm² Waters Symmetry 1.7-μm BEH-130 C18 column via an auto-sampler for separation by reverse-phase chromatography on a NanoAcquity UPLC system (Waters). A dual-stage linear gradient was conducted with a flow rate of 1 μl min–1 was applied for peptide separation, where solvent A was 0.1% FA/2% ACN/water and solvent B was 0.1% FA/2% water/ACN. Solvent B was increased from 2 to 25% over 35 min and then from 25 to 60% over 2 min, with a total analysis time of 60 min. Peptides eluting from the column were analyzed on an LTQ Orbitrap Elite mass spectrometer (Thermo Fisher) equipped with an Advance CaptiveSpray ionization source (Michrom). MS/MS precursor ions were scanned in Fourier transform mass spectrometry at 60,000 resolution and tandem mass spectrometry (MS/MS) data acquired on the 15 most abundant ions in LTQ linear ion trap mass spectrometer in data-dependent mode. MS/MS spectra were searched using the Mascot algorithm (Matrix Sciences) against a concatenated target-decoy database comprised of the FASTA sequences from the N. naja predicted proteome, Swiss-Prot entries from the Elapinae family (https://www.uniprot.org/taxonomy/42168) and common contaminant proteins, as well as the reversed versions of each sequence. Peptide spectral matches were filtered using a linear discriminant function to a peptide FDR of 5% to identify those that mapped to the toxin proteins identified from the genome.

3FTX structural modeling. Three-dimensional structural models were generated using MOE software (Chemical Computing Group, v.20101) and homology modeling was performed using the AMBER10 EHT forcefield18. Representative sequences were first searched in the protein database and the closest hits based on percentage sequence identity were chosen as templates for construction of structural homology models and for root mean square deviation calculations (Supplementary Table 10).

Molecular evolutionary analysis. For pairwise comparisons of both venom-expressed and non-venom-expressed genes, the number of nucleotide substitutions per synonymous (Ks) and non-synonymous site (Kt) for each pair of protein-coding genes was computed according to the Nei–Gojobori method19 using Squid Plott (http://www.gen-info.osaka-u.ac.jp/~uhmin/study/sqidiPlot/index.html).

Identification of ZRS limb enhancer deletion. To identify the ZRS enhancer region, sequences corresponding to this region were downloaded from ref. 20. BLASTn (v.2.2.29+)16 was used to identify the orthologous N. naja sequence. Multiple sequence alignment was performed using ClustalW20.

Variant analysis. We first performed phylogenetic analysis using 1,654 polymorphisms identified in the mitochondrial genomes using short-read whole-genome sequence data for each of the six study animals. Briefly, adapter and quality trimmed reads were aligned to the N. naja mitochondrial genome (GenBank: DQ430648.1) using BWA. Variant calling was performed using SAMtools21. The resulting alignments of all six animals were compared to a reference22. Multiple sequence alignment of the consensus mitochondrial sequence from the study animals, generated using MUSCLE23, was used to construct the phylogenetic tree using FigTree v.4.2.

For the genome-wide protein-coding variant analysis, whole-genome sequencing data for each of the six animals was used in each of the six study animals was used to construct respective fasta files using the Nana_v5 reference genome assembly using BWA24 with default options. Mapped reads were sorted and duplicates marked with PicardTools. Local realignment, germline variant calling and joint genotyping were performed using GATK (v.4.1.0.0)25,26. Germline variants were decomposed and normalized with vt27 and functionally annotated with snpEff28 against the annotated genome. Germline variants that had genotype calls in all samples were used to calculate percentage identity values between samples.

Genome heterozygosity estimation. K-mers from short-read sequencing data for each of the six animals used in each of the six study animals was counted using jellyfish (v.2.2.6)29 for several values of k: 15, 17 and 21. Each computed K-mer histogram was then analyzed with GenomeScope30.

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

Data availability
Sequence data (DNA and RNA) can be accessed at NCBI under BioProject accession no. PRJNA527614. The MS data are available under accession no. MSV00084564.

Code availability
All software tools used in this study are provided in the accompanying Nature Research Reporting Summary document.

References
89. Raudsepp, T. & Chowdhary, B. P. FISH for mapping single copy genes. Methods Mol. Biol. 422, 31–49 (2008).
90. Zhu, D. et al. Flow cytometric determination of genome size for eight commercially important fish species in China. In Vitro Cell Dev. Biol. Anim. 48, 507–517 (2012).
91. Wade, C. M. et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326, 865–867 (2009).
92. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31 (2005).
93. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
94. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
95. Hanje-Baltz, A., Stephens, S. A., Davey, S., Scheidegger, C. E. & Lyons, E. SynMap2 and SynMap3D: web-based whole-genome synteny browsers. Bioinformatics 25, 2197–2198 (2009).
96. Soderlund, C., Bomhoff, M. & Nelson, W. M. SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res. 39, e68 (2011).
97. Grabherr, M. G. et al. Genome-wide synteny through highly sensitive sequence alignment: satsuma. *Bioinformatics* **26**, 1145–1151 (2010).
98. Wu, T. D. & Watanabe, C. K. GMAP, a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
99. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
100. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
101. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
102. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
103. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene set enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
104. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
105. Case, D. A. et al. The Amber biomolecular simulation programs. *J. Comput. Chem.* **23**, 2668–1688 (2003).
106. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
107. Kvon, E. Z. et al. Progressive loss of function in a limb enhancer during *Drosophila* melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
108. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **21**, 2947–2948 (2005).
109. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **29**, (15–21 (2013).
110. Yan, J., Li, H. & Zhou, K. Evolution of the mitochondrial genome in snakes: gene rearrangements and phylogenetic relationships. *BMC Genomics* **9**, S69 (2008).
111. Preston, M. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
112. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (2015).
113. Cicogna, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
114. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
115. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).

**Acknowledgements**

We thank Genentech and AgriGenome DNA Sequencing and Bioinformatics groups for their help. We also thank Genentech Department of Pathology for their support. We thank A. Bruce for help with the illustrations. Our thanks to P. Dhanesh Kumar and S. Santhosh Kumar from the Department of Forest and Wildlife, Government of Kerala, India for their study support. We also acknowledge Dovetail Genomics and BGN for their technical assistance.

**Author contributions**

S.S. conceptualized the study. K. Suryamohan, R.C.P., O.K.M., M.D.D., M.S.S., I.K., B.F., R.N.H. and R.M.K. curated the data. K. Suryamohan, M.W., M.S.S., M.J., I.K., O.K.M., R.C.P., S.S., R.G., A.R., L.D.G., B.F., D. Velayutham, G.D.R., G.W., R.N.H., S.D., T.R., and R.K.M. were responsible for formal analysis. K. Suryamohan, S.P.K., J.G., M.J., M.S.S, K. Senger, D. Vargas, A.A., S.M., Y.J.C., P.I., S.C., M.M., Z.M., H.S., J.Z., G.A.W., J.S., D.S.K., R.N.H., S.D., T.R., R.K.M. and S.S. conducted the investigation. K. Suryamohan, M.J., D.S.K., R.N.H., E.W.S. and S.S. created the methodology. S.S. and K. Suryamohan were responsible for project administration. Resources were sourced by S.P.K., B.K., M.B., M.S., J.Z., K.W., A.Z. and Z.M. Software was created by K. Suryamohan, M.S.S, M.W., S.D., E.W.S. and L.D.G. S.S. supervised the project. K. Suryamohan, R.C.P., M.W., R.K.M., I.K. and S.S. completed validation procedures. Visualization was done by K. Suryamohan, M.W., M.S.S, O.K.M., M.D.D., M.J. and A.R. The original draft was written by K. Suryamohan, M.J., T.R., R.H. and S.S. This was reviewed and edited by K. Suryamohan, S.P.K., M.J., D.S.K., R.N.H., D.V., E.W.S., K. Senger, M.W., H.S., A.R., G.D.R., J.Z., T.R., R.M.K., A.Z., Z.M. and S.S.

**Competing interests**

Employees of Genentech hold Roche shares/options, and employees of MedGenome hold MedGenome shares/options.

**Additional information**

Extended data is available for this paper at https://doi.org/10.1038/s41588-019-0559-8.

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41588-019-0559-8.

Correspondence and requests for materials should be addressed to S.S.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Indian cobra genome size estimation by flow cytometry. 

**a**, Gating strategy for Indian cobra (*Naja naja*) genome size estimation showing the propidium iodide (PI) positive sample within the elliptical gate. **b**, PI positive gated population in a histogram showing PI stained *N. naja* blood and *Equus caballus* (horse) lymphocytes. **c**, Table of median fluorescence intensities measured for *Naja naja* and *Equus caballus* and estimated genome size in Gb in 3 replicate experiments. A total of 3000 and 300 measurements were conducted for the Indian cobra and horse, respectively.

| Sample          | Median PI | Genome size (Gb) Replicate 1 | Genome size (Gb) Replicate 2 | Genome size (Gb) Replicate 3 |
|-----------------|-----------|------------------------------|------------------------------|------------------------------|
| *Naja naja*     | 208816    | 1.48                         | 1.59                         | 1.77                         |
| *Equus caballus*| 352615    | 2.50                         | 2.70                         | 3.00                         |
Extended Data Fig. 2 | *N. naja* karyotyping. Representative karyotype obtained from cultured red blood cells from a female animal NN03. A total of N = 15 cells were karyotyped.

**Naja naja** karyotype

2n = 38
7 pairs of macrochromosomes
11 pairs of microchromosomes
Sex chromosomes (ZW – female, ZZ - male)
Extended Data Fig. 3 | Genomic repeat elements identified in the Indian cobra genome. a, Bar plot of the percent distribution of the different classes of repeat elements in the N. naja genome (Nana_v5). b, Comparison of proportion of the repeat content among 4 published snake, green anole lizard genomes and the Indian cobra genome.
Extended Data Fig. 4 | Syntenic comparisons of SVMP gene cluster. Relevant syntenic genomic regions between the Indian cobra (Nana_v5), prairie rattlesnake and green anole lizard genomes are shown. Orthologous gene pairs are indicated by shaded regions across the corresponding genomic regions. Yellow arrows with blue border indicate gene synteny, while those without colored borders represent potential species-specific duplications. SVMP, snake venom metalloproteinase.
Extended Data Fig. 5 | Heatmap of differentially upregulated genes in the *N. naja* venom gland transcriptome. Protein families are indicated in colored bars. NN01 and NN02 correspond to *N. naja* specimens obtained from Kerala, India. NN03, NN04, NN05 and NN06 correspond to *N. naja* specimens obtained from the Kentucky reptile zoo. Expression values plotted as log2 transformed CPM values with FDR cutoff set at 1% used for differential expression analysis.
**Extended Data Fig. 6** | **Pairwise structural comparison of representative *N. naja* 3FTxs.** RMSD matrix for the structural models from 9 representative 3FTxs.

|       | Nana001KS | Nana002KS | Nana005KS | Nana018KS | Nana006KS | Nana013KS | Nana008KS | Nana010KS | Nana012KS |
|-------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Nana001KS | 0.00      | 2.53      | 2.56      | 2.73      | 3.27      | 3.52      | 3.84      | 3.71      | 5.66      |
| Nana002KS | 2.53      | 0.00      | 2.08      | 2.12      | 3.21      | 3.30      | 3.10      | 3.09      | 4.98      |
| Nana005KS | 2.56      | 2.08      | 0.00      | 1.96      | 2.78      | 2.58      | 3.92      | 3.92      | 5.40      |
| Nana018KS | 2.73      | 2.12      | 1.96      | 0.00      | 3.27      | 2.91      | 3.97      | 3.87      | 5.30      |
| Nana006KS | 3.27      | 3.21      | 2.78      | 3.27      | 0.00      | 2.60      | 3.30      | 3.31      | 5.79      |
| Nana013KS | 3.52      | 3.30      | 2.58      | 2.91      | 2.60      | 0.00      | 3.86      | 3.90      | 5.14      |
| Nana008KS | 3.84      | 3.10      | 3.92      | 3.97      | 3.30      | 3.86      | 0.00      | 0.94      | 5.51      |
| Nana010KS | 3.71      | 3.09      | 3.92      | 3.87      | 3.31      | 3.90      | 0.94      | 0.00      | 5.60      |
| Nana012KS | 5.66      | 4.98      | 5.40      | 5.30      | 5.79      | 5.14      | 5.51      | 5.60      | 0.00      |
Extended Data Fig. 7 | nAChR polymorphism in Indian cobra. Multiple sequence alignment showing the region surrounding the alpha neurotoxin binding site in nAChR of seven vertebrate animals and *N. naja* nAChR (Nana03380-RA) identified a SNP at residue 189 in the *N. naja* nAChR indicated by the blue arrow. nAChR – nicotinic acetylcholine receptor; DANRE – *Danio rerio*; CHICK, *Gallus gallus*; HERIC, *Herpestes ichneumon*; HUMAN, *Homo sapiens*; PANTR, *Pan troglodytes*; RAT, *Rattus norvegicus*; MOUSE, *Mus musculus*. 
Extended Data Fig. 8 | Molecular evolution of Indian cobra SVMP genes. $K_A$ and $K_S$ values were calculated according to the Nei-Gojobori method. $K_A$ and $K_S$ with values < 0.1 were not included in further analysis for reliable analysis. NVMP, Non-venomous metalloproteinase genes; SVMP, Snake venom metalloproteinase genes.
Extended Data Fig. 9 | SVMP expression and comparative protein sequence alignment. a, Multiple sequence alignment of SVMP proteins from the Indian cobra, and representative SVMPs from other Elapid and Viperid species and human ADAM28. Arrows indicate additional cysteine residues typically present in the M12 domain of Viperidae SVMPs. b, Phylogenetic tree reveals distinct clusters formed by elapid and viperid SVMPs. The bar indicates 0.03 substitutions per nucleotide position. Elapid species - OPHHA, Ophiophagus hannah; NAJAT, Naja atra; NAJMO, Naja mossambica; NAJK, Naja kaouthia; Viperid species - DABRR, Daboia russelii; AGKCL, Agkistrodon contortrix laticinctus; DEIAC, Deinagkistrodon acutus; BOTJA, Bothrops jararaca; HUMAN, Homo sapiens.
Extended Data Fig. 10 | See next page for caption.
Extended Data Fig. 10 | Genetic polymorphisms in 6 *N. naja* specimens. a–d, Pairwise similarity (PWS) matrices based on (a), all genome-wide protein-altering variants, (b), all venom gland-expressed genes, (c), core venom-ome genes, and (d), all 3FTx genes identified in this study. e, Distribution of protein altering variants in 106 venom gland-specific toxin genes. f, Distribution of protein altering variants in 3FTx genes located on chromosome 3 across all six study animals (NN01-NN06). Within each track in (b–f), homozygous variants are shown as blue vertical lines while heterozygous variants are shown as red vertical lines. NN01 and NN02 correspond to *N. naja* specimens obtained from Kerala, India. NN03, NN04, NN05 and NN06 correspond to *N. naja* specimens obtained from the Kentucky reptile zoo. MIC, microchromosomes.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

No commercial/custom code was used for data collection.

Data analysis

We list here the software URLs of all open-source tools used for data analysis.

R: https://www.R-project.org/
Arrow: https://github.com/PacificBiosciences/GenomicConsensus
Canu: https://github.com/marbl/canu
Flye: https://github.com/fenderglass/Flye
WTDBG https://github.com/ruanjue/wtdbg2
BUSCO: https://github.com/ezlab/busco
Jellyfish: https://github.com/lh3/jellyfish
GenomeScope: http://qb.cshl.edu/genomescope/
Repeatmasker: http://www.repeatmasker.org
Repeatmodeler: http://www.repeatmasker.org/RepeatModeler.html
MAKER: http://www.yandell-lab.org/software/maker.html
SNAP: http://snap.cs.berkeley.edu
AUGUSTUS: http://augustus.gobics.de/
Exonerate: https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate
snpEff: http://snpeff.sourceforge.net/
Enrichr: http://amp.pharm.mssm.edu/Enrichr
GATK: https://software.broadinstitute.org/gatk/download/
BWA: https://github.com/lh3/bwa/releases
EdgeR: https://bioconductor.org/packages/release/bioc/html/edgeR.html
Synten plots: https://github.com/biopython/biopython/blob/master/Doc/examples/Proux_et_al_2002_Figure_6.py
OrthoFinder: http://www.stevekellylab.com/software/orthofinder
STAR: https://github.com/alexdobin/STAR

1 nature research | reporting summary
Corresponding author(s): Somasekar Seshagiri
Last updated by author(s): Nov 19, 2019
For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing data that support this study have been deposited in NCBI under the BioProject accession # PRJNA527614 and can be accessed at "https://dataview.ncbi.nlm.nih.gov/object/PRJNA527614". Mass spectrometry data collected in this study has been deposited at MassIVE and can be accessed via this link - "ftp://massive.ucsd.edu/MSV000084564/".

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences

Sample size A total of 6 animals was used in this study as described in Table S1a and methods section.

Data exclusions No data was excluded in this study.

Replication Biological replicates, where appropriate, have been used and described in the methods section and supplementary tables.

Randomization This is a de novo genome sequencing project and hence, randomization does not apply to this study.

Blinding This is a de novo genome sequencing project and hence, blinding does not apply to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a Involved in the study

- Antibodies
- Eukaryotic cell lines
- Palaeontology
- Animals and other organisms
- Human research participants
- Clinical data

Methods

n/a Involved in the study

- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals: ARRIVE guidelines recommended for reporting animal research

Laboratory animals We obtained horse blood from a horse (Equus caballus) that is a research mare at the Texas A&M Large Animal clinic. Species: Equus caballus Breed: American quarter horse
Sex: female
Age: 6 years old

Wild animals
Two animals used in this study from India were fresh road kills and were approved for use in the study as described in the methods section.

Field-collected samples
This study did not include field-collected samples

Ethics oversight
This study did not require ethics approval.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

---

## Flow Cytometry

### Plots

- Confirm that:
  - The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
  - The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a ‘group’ is an analysis of identical markers).
  - All plots are contour plots with outliers or pseudocolor plots.
  - A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

| Sample preparation | Sample preparation and processing are described in the materials and methods section of the main text. |
|--------------------|-------------------------------------------------------------------------------------------------------|
| Instrument         | BD Accuri™ C6 personal flow cytometer                                                                 |
| Software           | BD Accuri C6                                                                                            |
| Cell population abundance | Described in materials and methods section of main text                                         |
| Gating strategy    | Gating strategy is provided in Supplementary Fig. 1a                                                  |

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.