SUPPORTING INFORMATION

1. SI MATERIALS AND METHODS

1.1 King cobra tissue acquisitions and processing

All animal procedures complied with local ethical approval. Genome sequencing was undertaken on a blood sample obtained from an adult, captive, male king cobra that originated from Bali, Indonesia. Blood was obtained by caudal puncture and frozen in liquid nitrogen. Venom was extracted and four days later (to maximise mRNA production), the venom gland, accessory gland and other tissue samples were dissected (heart, lung, spleen, brain, testes, gall bladder, pancreas, small intestine, kidney, liver, eye, tongue and stomach) from a freshly euthanized second Indonesian adult male specimen and stored in RNAlater.

1.2 Estimation of genome size

Collected king cobra blood was washed with physiological salt and fixed in cold ethanol. Prior to analysis, the cells were collected, resuspended in physiological salt and stained with propidium iodide. After 30 minutes of incubation the cells were analysed by fluorescence-activated cell sorting (FACS) flow cytometry, using chicken blood cells as a size reference (1.05 Gbp haploid). The king cobra genome size was estimated by the following calculation:

= (chicken size) / (mean fluorescence chicken) X (mean fluorescence king cobra).

On the basis of this analysis, the king cobra genome was estimated to be 1.36-1.59 Gbp in size (haploid) (Fig. S1).

1.3 Genomic DNA library preparation

Genomic DNA was isolated from blood using the Qiagen Blood and tissue DNeasy kit according to the manufacturer's description (Qiagen GmbH, Hilden). Paired-end

libraries were prepared from 5 µg of isolated gDNA using the Paired-End Sequencing Sample Prep kit according to the manufacturer's description (Illumina Inc., San Diego). Either a 200 bp band or a 500 bp band was cut from the gel (libraries PE200 and PE500, respectively; see Table S1). After amplification the resulting libraries were analyzed with an Agilent Bioanalyzer 2100 DNA 1000 series II chip according to the manufacturer's description (Agilent, Santa Clara).

Mate Pair libraries were prepared from 10 µg of isolated gDNA using the Mate Pair 2–5 Kb Sample Prep kit according to the manufacturer's description (Illumina Inc., San Diego). Bands from 2–15 Kbp were cut from gel (MP2K, MP7K, MP10K and MP15K libraries, see Table S1). After the first gel purification the fragment length was analyzed by an Agilent Bioanalyzer 2100 DNA 12000 chip. After circularization, shearing, isolation of biotinylated fragments, and amplification, the 400 to 600 bp fraction of the resulting fragments was isolated from the gel. Finally, the libraries were examined with an Agilent Bioanalyzer 2100 DNA 1000 series II chip.

1.4 mRNA-Seq and smallRNA library preparation

Total RNA was isolated using the Qiagen miRNeasy kit according to the manufacturer's instructions and analyzed with an Agilent Bioanalyzer 2100 total RNA Nano series II chip. The RNA used for the venom mRNA-Seq library was obtained from the venom gland and the RNA used for the accessory gland RNA-Seq library from the accessory gland. The RNA used for the 'pooled multi-tissue archive' mRNA-Seq library was obtained by mixing equal amounts of total RNA isolated from heart, lung, spleen, brain, testes, gall bladder, pancreas, small intestine, kidney, liver, eye, tongue and stomach. Transcriptome libraries were prepared from 10 µg total RNA, using the Illumina mRNA-Seq Sample Preparation Kit according to the manufacturer's instructions. The same total RNA was also used to prepare the smallRNA libraries using the Illumina smallRNA v1.5 sample preparation kit according to the manufacturer's instructions.

1.5 Sequencing

Genomic libraries were paired-end sequenced with a read length of 36–151 nucleotides on an Illumina GAIIx instrument according to the manufacturer's description. The mRNA-Seq and smallRNA libraries were single-read sequenced on an Illumina GAIIx with a read length of 51 nucleotides. Image analysis and base calling were done by the Illumina pipeline.

1.6 Genome assembly strategy

In assembling the king cobra genome, we largely followed the strategy pioneered by Li *et al.* (1, 2) for the assembly of the giant panda genome. In summary, this approach consists of four stages:

1. Illumina sequencing of a number of genomic libraries with varying insert sizes;
2. Preprocessing of sequencing reads;
3. De Bruijn graph-based assembly of reads into contig sequences;
4. Orientation of contigs in scaffolds based on large-insert library information.

Sequencing reads from both paired-end libraries were used in building the initial contigs. Both sets were preprocessed to eliminate low quality reads and nucleotides, as well as adapter contamination (mainly caused by insert sizes smaller than the read length). Because of the small insert size of the PE200 library, many read pairs from this library overlap at their 3' ends. When possible, these pairs were merged into longer single reads. This preassembly procedure has the dual advantage of producing long reads (which improve the quality and efficiency of the subsequent assembly) and providing confirmation for the identity of the 3' ends of the reads (which are generally determined with lesser confidence). We merged read pairs that exhibited at least seven nucleotides of unambiguous sequence overlap. Using this criterion, 61% of pairs could be merged, resulting in single reads with a mean length of 108 nt. Seven percent of reads from a 2 × 151 nt run of the PE500 library could be merged into single reads with a mean length of 217 nt.

For initial contig assembly, we employed the CLC Assembly Cell *de novo* assembler (version 3.2, CLC bio, Aarhus, Denmark). This is an efficient implementation of a De Bruijn graph-based assembler (3), which enables the assembly of the king cobra genome on a dual quad-core Xeon workstation with 48 GB of RAM installed in approximately eight hours. A run with a minimum required contig size of 100 bp and a k-mer length of 31 nt resulted in an assembly with a total length of 1.45 Gbp and a contig N50 of 3982 bp (i.e. 50% of the assembly, or 725 Mbp, is in contigs of at least this length).

Initial contigs were oriented in larger supercontigs (scaffolds) using SSPACE (4). Briefly, SSPACE aligns paired reads to the contigs [using Bowtie (5)], and combines contigs if they are connected by at least a specified number of pairs within the limits set for the insert size of the pair library. The insert size is then used to estimate the size of the gap between the contigs. In addition, the algorithm can be forced to extend scaffolds with a contig only if the evidence for its unique placement is above a set threshold, or else abort growth for that scaffold. This allows contigs representing collapsed repeats to be either included or excluded from the final scaffolds. SSPACE was used to scaffold contigs in a hierarchical fashion, employing first links obtained from the PE500 library to generate intermediate supercontigs, which were used as input for subsequent runs with links from individual mate-pair libraries increasing in size. At each stage, a minimum of three non-redundant links was required to join two contigs. This procedure resulted in a final scaffold set with a total length of 1.66 Gbp and an N50 of 225,511 bp.

1.7 Genome annotation

Annotations for the king cobra genome assembly were generated using the automated genome annotation pipeline MAKER (6-8), which aligns and filters EST and protein homology evidence, identifies repeats, produces *ab initio* gene predictions, infers 5' and 3' UTR, and integrates these data to produce final downstream gene models along with quality control statistics. Inputs for MAKER included the king cobra (*Ophiophagus hannah*) genome assembly, a snake specific repeat library constructed using the complete king cobra genome assembly, the complete Burmese python

(*Python molurus bivittatus*) genome assembly (9), with repeats identified using RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) and classified further using Repclass (10). Gene annotations were made using a protein database combining the Uniprot/Swiss-Prot (11, 12) protein database and all sequences for *O. hannah* and *Anolis carolinensis* from the NCBI protein database (13). *Ab initio* gene predictions were created by MAKER using the programs SNAP (14) and Augustus (15). Gene models were further improved by providing MAKER with all mRNAseq data generated in this study for *O. hannah*, which were combined to generate a joint assembly of transcripts using Trinity (16). A total of three iterative runs of MAKER were used to produce the final gene set.

1.8 Transcriptome assembly

Reads for the venom gland, accessory gland and pooled multi-tissue archive were assembled with Abyss (17, 18) with various k values (every even number from 50 to 96). Because Abyss tends to miss highly expressed contigs (19), we have also run the Trinity assembler (16) on the raw data. The resulting assemblies were joined by an iterative BLAST and cap3 assembler (20). Coding sequences were extracted using an automated pipeline, based on similarities to known proteins, or by obtaining coding sequences from the larger open reading frame of the contigs containing a signal peptide. A non-redundant set of the coding and their protein sequences were mapped into a hyperlinked excel spreadsheet, which is presented as File S2. Signal peptides, transmembrane domains, furin cleavage sites and mucin type glycosylations were determined with software from the Center for Biological Sequence Analysis, Denmark (21-24). Detailed bioinformatic analysis of the pipeline utilised can be found in Karim *et al.* (20). To map the raw Illumina reads to the coding sequences and determine their tissue bias, raw reads from each library were blasted to the coding sequences using blastn with a word size of 25 (-W 25 switch) and allowing recovery of up to 3 matches. The 3 matches were used if they had less than two gaps and if their scores were equal to the best score. The resulting blast file was used to compile the number of reads each CDS received from each library and also to count the number of hits at each base of the CDS, allowing for the determination of the average CDS coverage, maximum coverage and minimum coverage. These results can be

statistically tested by a X^2 test (using the number of reads per CDS), the results of which are reported significant if $P < 0.05$ and no CDS had an expected value of 5 or less.

1.9 Gene family evolution

1.9.1. Sourcing gene sequences

King cobra sequences exhibiting homology to toxin families were identified through: (i) annotation in the genome or transcriptome and (ii) by BLAST search of toxin and non-toxin gene homologs against king cobra genome and transcriptome datasets in CLC Main Workbench (version 6.2, CLC bio, Aarhus, Denmark). Burmese python (*P. molurus bivittatus*) (9, 25) and green anole lizard (*A. carolinensis*) (26, 27) gene homologs were obtained in a similar manner. Additional non-venom gland snake sequence data was sourced by mining a pooled tissue transcriptome of the garter snake (*Thamnophis elegans*) using the Bronikowski Lab Data Server (http://eco.bcb.iastate.edu/). Coding regions of genomic loci were extracted and putative orthologs and paralogs from other vertebrate species (with a focus on other snakes) obtained by mining GenBank for BLAST hits and by using the datasets of Casewell *et al*. (28). Gene loci were obtained for the 3FTx, CRISP, cystatin, hyaluronidase, kallikrein, LAAO, lectin, NGF, PLA_2 and SVMP gene families. DNA datasets were trimmed to the open reading frame in MEGA5 (29), with identical sequences and those containing truncations or frameshifts (as the result of insertions or deletions) excluded. Subsequently, each dataset was aligned using MUSCLE (30) and alignments checked manually prior to phylogenetic analysis.

1.9.2. Phylogenetic analysis

DNA gene trees for each toxin family were generated using Bayesian inference. Considering complex models of sequence evolution have been demonstrated to extract additional phylogenetic signal from data (31, 32), we subjected the datasets to codon analysis in MrModelTest v2.3 (33), with the model favoured under the Akaike Information Criterion (AIC) (34) selected for incorporation into Bayesian inference

analyses. Models selected by MrModelTest for each dataset are displayed in Table S6. Bayesian analyses were undertaken using a Markov Chain Monte Carlo (MCMC) algorithm in MrBayes v3.2 (35) on the freely available bioinformatic platform Bioportal (www.bioportal.uio.no) (36). Each dataset was run in duplicate using four chains simultaneously (three heated and one cold) for $5x10^6$ generations, sampling every 500th cycle from the chain and using default settings in regards to priors. Tracer v1.4 (37) was used to estimate effective sample sizes for all parameters and to construct plots of ln(*L*) against generation to verify the point of convergence (burnin); trees generated prior to the completion of burnin were discarded.

1.9.3. Mapping toxin gene family tissue expression

The locations of expression of snake sequences (venom or non-venom) were mapped manually onto the ensuing gene trees. The locations of tissue expression of Burmese python genes were identified using transcriptome expression data constructed for the following tissue types: blood, heart, liver, muscle, ovary, rictal gland, spleen, stomach and testes (9, 25). The location of tissue expression of green anole lizard genes (26, 27) were identified by downloading transcriptome datasets (brain, dewlap, ovaries, testes and pooled organs [gall bladder, heart, kidney, liver, lung, spleen, tongue]) from the UniGene archive via the Anolis genome project webpage (http://anolisgenome.org/?page_id=27) and blasting *A. carolinensis* genomic loci against the transcriptome datasets. For king cobra hits, pie charts were displayed on the phylogenetic trees (Figs. S11, S12 and S15-S22) that visualise the proportion of tags expressed in each tissue location of the co-assembled transcriptome (venom gland, accessory gland and pooled multi-tissue archive).

1.9.4. Toxin family gene duplication

Toxin family gene trees were pruned to contain sequences isolated from the king cobra and Burmese python and a single sequence representing an outgroup. Where transcriptome hits were present (i.e. in the absence of genomic loci), species-specific clades identified in the original gene trees (Figs. S11, S12 and S15-S22) were pruned to remove sequences that could represent potential allelic variants (i.e. removal of one of two sequences present for each clade). The ensuing pruned gene trees were

analysed using the duplication and loss criterion in iGTP (38) with the following species tree: (outgroup, (king cobra, Burmese python)).

1.9.5. Selection analyses

We assembled sets of sequences, which we aligned using MUSCLE (30), of putative orthologs and paralogs by mining GenBank for vertebrate BLAST hits for 3FTx, CRISP, cystatin, hyaluronidase, kallikrein, LAAO, NGF, PLA_2 and SVMP gene loci. We verified no erroneous frame shifts had been introduced by the alignment step by generating the amino acid translation for each sequence from GenBank and filtering out any incorrectly aligned sequences. Subsequently, we removed all "sparse" codons (i.e. those codons that occur in fewer than 50% of all sequences). We then removed all "short" (i.e. those whose length is less than 90% of the length of the longest sequence for that species) and duplicated sequences within each species, with the exception of those from the king cobra (*O. hannah*) and the green anole lizard (*A. carolinensis*). On each of these cleaned alignments we performed Bayesian phylogenetic inference using MrBayes (35) under a GTR+G+I model (39), running the analysis for $2x10^6$ generations with a burn-in of 10%. Convergence was verified using Tracer (40). We subsequently used the majority-rule consensus trees obtained from these analyses as starting trees to obtain fully resolved maximum likelihood trees using PHYML (41) with the BEST tree-searching algorithm.

On the maximum likelihood trees we reconstructed the most parsimonious location of state changes between the venomous and non-venomous state, thereby identifying the two respective classes of branches for the subsequent calculation of $\omega_{venomous}$ and $\omega_{non\text{-}venomous}$. We computed the ω ratios for the two classes of branches using the codeml program from the PAML package (42). We calculated codon frequencies from the average nucleotide frequencies at the three codon positions and assumed (conservatively) no variation of ω among sites. Substitution model parameters (transition/transversion ratio parameter κ and gamma shape parameter α) were estimated from the data.

1.9.6. Stochastic mutational mapping and ancestral state reconstruction

Stochastic mutational mapping and ancestral state reconstruction analyses were performed using SIMMAP v1.5.2 (43) to: (i) explicitly test the timing of lectin gene recruitment in to the venom of snakes and (ii) investigate the evolution of accessory gland expressed king cobra lectin paralogs from venom expressed genes. We first used stochastic mutational mapping (44-46) to trace the pattern of character state changes on the lectin gene family phylogeny reconstructed in SI methods section 1.9.2. Tips on the lectin gene tree were allocated a character (venom = 1 and non-venom = 0) based on the tissue location each sequence was sourced from or expressed in, before the 'map' criterion in SIMMAP was applied. The resulting pattern of predicted character state changes overlaid onto the gene tree is displayed in Fig. S13. Stochastic mutational mapping predicted a single recruitment event of lectins in to snake venom, with king cobra lectins subsequently being repeatedly 'reverse recruited' (28) for expression in the accessory gland. To explicitly test this hypothesis, we used the ancestral state criterion in SIMMAP to predict the posterior probability of the state of a character (venom or non-venom) at key ancestral nodes in a gene tree (28, 43, 46). This method accounts for phylogenetic uncertainty in the gene tree by sampling tree topologies, branch lengths, model parameters and character histories. The posterior probability that an ancestral node has a venom character state was assessed by using the same character allocation described above and using 1000 rooted post-burnin trees sampled from the posterior distribution of the Bayesian analyses for the lectin DNA dataset described in SI methods section 1.9.2. We used a low-rate prior which incorporated a mean of 1 and a standard deviation of 5 of the prior distribution; the number of samples and stochastic draws from this prior distribution was set at 50 (28, 46, 47). The Bayesian posterior probability of key nodes having the 'venom' character state are displayed in Fig. S13.

1.10 Venom proteomics

1.10.1. Constructing the king cobra venom proteome

In-solution digest

Crude lyophilized venom from the same animal from which the transcriptomes were obtained was resuspended in water with 10 mM Tris (pH = 8.0) at 1 mg protein/ml. The venom solution (50 µl) was combined with 2.0 µl of 1% ProteaseMAX (Promega, Madison, WI, USA) in 50 mM NH_4HCO_3 and 41.5 µl 50 mM NH_4HCO_3. Reduction was accomplished by addition of 1.0 µl 0.5 M DTT (incubated at 56°C for 20 min), and alkylation was done by addition of 2.7 µl 0.55 M iodoacetamide and incubation for 15 min at room temperature (~24°C) in the dark. Next, 1.0 µl of 1% ProteaseMAX and 1.8 µl (1 µg/µl) trypsin (Trypsin Gold, Promega) were added, and the mixture was incubated at 37°C for 3 h. Next, 0.5 µl trifluoroacetic acid was added, and the mixture was incubated at room temperature for 5 min, followed by centrifugation at 10,000 rpm for 10 min. The supernatant was drawn off and used for subsequent analysis.

Separation and mass determination

Peptides from trypsin-digested venom were separated via column chromatography with a 50 x 2.1 mm Hypersil Gold C18 column (Thermo Scientific, Waltham, MA, USA) followed by tandem mass spectrometry (MS). The column was run at 200 ul/min with an initial wash of buffer A (0.1% formic acid) followed sequentially by a gradient of 0 to 50% buffer B (80% acetonitrile with 0.1% formic acid) over 32 min and a gradient of 50 to 100% buffer B over 18 min. Mass analysis was conducted using an ESI (electrospray ionization) ion trap mass spectrometer (LCQ Fleet, Thermo Scientific). Parent scans of multiply-charged ions were conducted at 3 µscans/200 ms over a mass range of 500-2000 Da to obtain the charged ions. Subsequently, data-dependent scans of the top 3 most intense ions from each parent scan were performed for amino acid sequencing. Ions were chosen based on the highest charge/mass ratio (a default value of charge state +2 was used, as trypsin-digested peptides usually have at least one arginine or lysine in the sequence) and fragmentation of peptides was accomplished by collision-induced dissociation using helium gas with a normalized collision energy of 35 eV. Following fragmentation, the product ions were scanned for 30 ms (3 µscans/25 ms). Dynamic exclusion was enabled in order to acquire sequences of all possible ions from the parent scan. Two

different tryptic digests were conducted, and two samples from each digest were run on tandem MS (total of four runs).

Proteome construction

Monoisotopic peptide fragments from all four tandem MS runs were compared against databases containing 1) the translated venom gland (VG) and accessory gland (AG) transcriptomes (obtained in-house) as well as a Lepidosaurian subset of the NCBI database, or 2) the VG and AG transcriptomes combined with the human IPI database. All four runs were analyzed using both SEQUEST and Mascot software with a false discovery rate (FDR) of 0.01. For analysis, tandem MS mass tolerances were set to ± 1.0 Da; carboxyamidomethylation of cysteine and oxidation of methionine were both used as variable modifications; and peptides were allowed to have two missed cleavages. Detected peptides (hits) were considered valid if they began immediately after K/R and terminated with a K/R residue. Exceptions were made for peptide hits when they were located at the N- (without previous K/R) or C-terminus (without an end K/R) of the mature protein. Although each analysis resulted in different protein cut-off numbers based on the FDR, SEQUEST scores higher than 2.00 and Mascot scores of 15 or higher were considered significant. Proteins were considered as being present in the proteome (positive protein match) if they had a score above the cut-off in at least one of the analyses, had at least two complete peptide hits, and were secreted proteins (those with signal peptides). Transcriptomic (VG or AG) sequences with hits were searched against the full NCBI protein database for toxin family designation. All peptide hits were mapped on to the full protein sequences, and all these protein sequences within each family were aligned using Kalign and/or ClustalW (EMBL-EBI) for comparison. In the cases where multiple protein sequences had identical peptide hits without an additional unique peptide hit, only the protein with the highest SEQUEST or Mascot score was included. In the case of SVMPs, which are relatively large proteins, there were twelve partial protein sequences (from the VG transcriptome) without overlap. It was difficult to determine whether these sequences represented twelve or fewer isoforms. Sequence alignment showed a minimum of four unique isoforms, which were enumerated in the proteome. Overall, this proteome covers a conservative, minimum estimate of the secreted proteins present in the venom. The results of these analyses are presented in File S3.

1.10.2. Two-dimensional gel electrophoresis of king cobra venom

Two-dimensional gel electrophoresis

Two-dimensional electrophoresis (2-DE) and venom sample preparation was performed according to the method of Görg *et al.* (48). Precast immobilised pH gradient (IPG) strips (7 cm, pH 3–10 nonlinear; BioRad) were allowed to rehydrate in a total of 130 µl of the pre-treated venom sample at 50 V for 12 h. Strips were covered with mineral oil to avoid dehydration. First dimension isoelectric focusing was carried out in a BioRad Protean IEF, where the separation used a three-phase electrophoresis program: linear increase to 100 V over 1 h, then a constant gradient at 100 V for 1 h and finally a linear increase to 4000 V for over 5 h until the program reached at least 10,000 Vh. Prior to running the second dimension, immobilised pH gradient IPG strips were placed in a rehydration tray and equilibrated in two stages (reduction and alkylation) of 15 min each. Both were undertaken in equilibration buffer (50 mM Tris-HCl pH 8.8, 6 M urea, 30% glycerol, 2% sodium dodecyl sulphate) containing 1% (w/v) DTT for reduction and 2.5% (w/v) iodoacetamide for alkylation. IPG strips were then loaded on the top of vertical 0.75mm 15% polyacrylamide gels and sealed with dyed agarose. Samples were separated using the mini-Protean II electrophoresis system according to the manufacturer's recommendations (BioRad) at 50 V until the dye front reached the resolving gel, then at 100 V until the end of dye migration. Upon completion, 2-DE gels were stained with Coomassie Blue R-250 and destained until no background staining was observed. The Gel Doc^{TM} EZ system (BioRad) was used to record and edit gel images. Spot detection and further dissection was carried out manually.

In-gel tryptic digestion

In-gel tryptic digestion was performed on major protein bands of sizes less than 25 kDa according to the method of Hayter *et al.* (49). Peptides contained in the supernatant following trypsin digestion were recovered and dried using a Speed Vac (Heto DNA mini, ThermoScientific). Eluted peptides were reconstituted in 50µl 0.1% (v/v) formic acid, sonicated for 15 min and centrifuged at 13,000 rpm for 20 min.

Liquid chromatography-MS/MS

Peptides were initially separated by reversed-phase chromatography using a DIONEX UltiMate™ 3000LC chromatography system. Peptides (10 µl) were injected onto a C18 Dionex Acclaim® Pepmap100 Nano Trap reversed-phase Column (2 µm particle size, 75 µm diameter x 150 mm length) at nanoflow rate (0.3 µl*min^{-1}) and separated over linear chromatographic gradients. The gradients employed for chromatographic separation were composed of buffer A (2.5% acetonitrile, 0.1% formic acid) and buffer B (90% acetonitrile, 0.1% formic acid) over a 60 minute linear chromatographic gradient. Following chromatographic separation, MS analysis was performed on an LTQ Orbitrap Velos mass spectrometer using Xcalibur (Version 3.0) software (Thermo Scientific, UK). A data-dependent Top20 CID data acquisition method was used with intact peptides detected in the Orbitrap at a mass resolution of 30,000. Ions were scanned between 350-2000 *m\z* in positive polarity mode. The ion-trap operated with CID MS/MS (with wide band activation) on the 20 most intense ions. Dynamic exclusion was enabled to avoid repeatedly selecting intense ions for fragmentation and this was set at 500 with an exclusion duration of 20.0 s. The minimum MS signal threshold was set at 500 counts and the MS/MS default charge state was 2 with a 1.2 *m/z* isolation width, normalised CID at 35 V and an activation time of 10 ms.

Protein identification

Proteins were identified by screening LC-MS sequence data against the translated king cobra transcriptome database using Proteome Discover 1.0.0 (Thermo Scientific) software incorporating both Sequest and Mascot search algorithms. A parent mass tolerance of 1.5 Da and fragment mass tolerance of 1 Da were used, allowing for 1 missed cleavage. Carbamidomethylation of cysteine and oxidation of methionine were the fixed and variable modifications, respectively. For Sequest, high confidence peptide matches were filtered with an XCorr cut-off (+1>1.5, +2>2 and +3>2.5) and for Mascot, an Exp value of less than 0.01 indicated a confident peptide match. The results of these analyses are presented in Fig. S8 and Table S3.

1.10.3. Searching for lectins in the king cobra venom proteome

Two milligrams of king cobra venom proteins were separated by reverse-phase HPLC using a Teknokroma Europa C_{18} (0.4 cm x 25 cm, 5 mm particle size, 300 Å pore size) column and an Agilent LC 1100 High Pressure Gradient System equipped with DAD detector and micro-Auto-sampler. The flow-rate was set to 1 ml/min and the column was developed with a linear gradient of 0.1% TFA in water (solution A) and acetonitrile (solution B), isocratically (5% B) for 10 min, followed by 5-25 % B for 20 min, 15-45% B for 60 min, and 45-70% for 10 min. Protein detection was carried out at 215 nm with a reference wavelength of 400 nm. Isolated fractions were subjected to SDS-PAGE (on 15% polyacrylamide gels). Protein bands exhibiting apparent molecular masses compatible with the isotope-averaged masses calculated for the putative C-type lectin-like protein sequences identified in the king cobra genome and transcriptome were excised from a Coomassie Brilliant Blue-stained SDS-PAGE gel (Fig. S9) and subjected to automated in-gel digestion and tryptic peptide mass fingerprinting (PMF) (recorded with an Applied Biosystems Voyager-DE Pro™ TOF instrument). Peptide ion sequencing was performed by collision-induced dissociation MS/MS with a Waters' nanoAquity uPLC-SYNAPT™ G2 mass spectrometry system. CID spectra were interpreted manually or using MassLynx™ searches against the king cobra transcriptome database. Mass tolerance was set to ± 0.6 Da, and carbamidomethyl cysteine and oxidation of methionine were fixed and variable modifications, respectively. In all cases, analysis of digested bands revealed the identification of proteins other than C-type lectins, demonstrating the absence of this toxin type from king cobra venom. The results of these analyses are presented in Fig. S9 and Table S4.

1.11 *Hoxd12* comparison

After the final assembly of the king cobra *HoxD* cluster, a gap of an estimated size of 3 kb was detected between *Hoxd13* and *Hoxd11*. In order to confirm that *Hoxd12* was indeed absent from the sequence, we used primers flanking the gap to amplify this region. (Primer sequences: FW 5'AAGCTGCGAAGCTTTGGCTG; REV 5'GTCAGAGAGGCACTTGATCC) The PCR fragment was then cloned and

sequenced. The king cobra *Hoxd12* sequence was aligned with sequence data sourced from previously published *Hoxd12* genes from other squamate reptiles using default settings of the shuffle-LAGAN alignment program from vista server (http://genome.lbl.gov/vista/index.shtml). The genome contig containing the Burmese python *Hoxd12* gene was provided by Castoe *et al.* (9).

1.12 microRNAs

1.12.1. microRNA sequencing and analysis

Total RNA was extracted from the venom gland, accessory gland, and from a variety of other tissues (heart, lung, spleen, brain, testes, gall bladder, pancreas, small intestine, kidney, liver, eye, tongue and stomach - representing the pooled multi-tissues archive), of an adult king cobra (*O. hannah*) of Indonesian origin. Sample preparation for sequencing was undertaken following the small RNA sample preparation kit v1.5 (Illumina) guidelines and small RNAs sequenced using the Illumina GAIIx platform. We generated a total of 3.0 Gbp of sequencing data for the RNA samples: venom gland - 1.0 Gbp; accessory gland - 0.8 Gbp; pooled multi-tissue archive (see Table S1 for details). The small RNA sequences were analyzed using CLCBio Genome Workbench version 5 (http://www.clcbio.com). Briefly, the small RNA sequences were filtered for quality and size, and reads of low quality and lengths less than 17 nucleotides or greater than 26 nucleotides were discarded. The remaining pool of small RNAs was compared to miRBase release 18 (http://www.miRBase.org) to extract orthologous mature microRNA sequences from each RNA sample of *O. hannah*. The orthologous mature microRNA sequences were then mapped to the genome of *O. hannah* and 70 bp upstream and downstream of the mature sequence was extracted as the potential precursor microRNA (pre-miRNA) sequence using PHP scripts and Blast (50) 2.2.26+ running on a Linux Ubuntu 12.04, with intel Xeon CPU and 16GB of memory.

The secondary structures of the pre-miRNA sequences were predicted using CLCBio and aligned with orthologous precursor sequences (usually *A. carolinensis*, *Gallus gallus*, and *Mus musculus*) using MacVector version 10.0. Orthologous *O. hannah*

pre-miRNA sequences that were not conserved outside of the mature gene product (either 5p or 3p) and/or did not fold into a hairpin were discarded.

The expression level of each microRNA was assessed using CLCBio and compared with data available at the miRNA targets and expression database (http://www.microRNA.org, release August 2010) for the expression profiles of orthologous microRNA genes in mouse and human (e.g. *miR-375*).

Targets for *O. hannah* microRNAs were predicted using our in-house program sTARsearch, designed based on targeting principles outlined in Bartel (51) to identify potential 3'UTR binding in the following order of priority: i) to the seed of the mature microRNA (nucleotide positions 2-7, 3-8, or 2-8); ii) to an adenine at position 1 of the target mRNA and potential complementary binding at this position; iii) to 3' supplementary binding at nucleotide positions 13-15 (52) and iv) to any and all other complementary nucleotide pairs. sTARsearch allows for a maximum of a 5 nucleotide loop (a.k.a bulge) between the seed and 3' supplementary binding in either the mature microRNA or the 3'UTR. When the exact 3' UTR was not known, 2000 nucleotides downstream of the coding sequence was used as the potential 3'UTR. All targets predicted are listed in File S1. All miRNA sequences have also been submitted to http://www.miRBase.org (53) and are available in File S1.

1.12.2. microRNA in situ hybridization

Whole mount *in situ* hybridisations for *miR-375* detection using 5' digoxigenin (Dig) labeled LNA (locked nucleic acid, Exiqon) probes were performed following the protocol from Darnell *et al*. (54). Standard tissue section *in situ* protocol from Jostarndt *et al*. (55) for paraffin embedded tissues was followed for *miR-375* detection in the adult king cobra venom gland. For whole mount *in situ* hybridisations in late stage snake embryos (27 days post oviposition or older), embryos were skinned and the abdominal wall was cut open, followed by an extended probe hybridization for approximately 36 hours. All *miR-375* LNA *in situ* hybridisations were carried out at 57˚C (TM = 79˚C) as was a no-probe control. *miR-196* LNA *in situ* was carried out at 47˚C as an additional negative control in the adult venom gland. Detailed protocols are available upon request. Images of whole embryos were taken with an Olympus

MVX10 Dissector. Tiled images of tissue section *in situs* were taken with an Olympus dotSlide, higher magnifications were obtained using an Olympus IX83 with CellSens Dimension imaging software.

1.13 Lectin *in situ* hybridisations

1.13.1. Probe production

An RNA template was obtained from the venom apparatus of an *O. hannah* specimen (provided by Venom Supplies Pty Ltd., Tanunda, South Australia). The venom system was removed in a sterile environment and placed in RNAlater® (Life Technologies, Victoria, Australia). Total RNA was extracted using an RNeasy Lipid Mini Kit (Qiagen, Victoria, Australia) with the obtained concentration estimated using UV spectrophotometry (NanoDrop Lite Spectrophotometer, Thermo Scientific, Victoria, Australia). Sequences of interest were amplified and converted to cDNA using a One-Step RT-PCR kit (Qiagen, Victoria, Australia) as per manufacturer's instructions. The in-house designed primers used for the lectin transcriptome contig Oh-516 (see File S2; corresponds to genome scaffold s8808 gene 2 in Fig. S12) (forward primer 5'-GGGCAATTCCTCTTGGTGAG-3' and reverse 5'-CCACTTCCAGGTGCGAGTAT-3') and Oh-3509 (see File S2; corresponds to genome scaffold s8157 gene 2 in Fig. S12) (forward primer 5'-GCGATTCCTCTTTGCAAGC-3' and reverse 5'-GACAGCACAGTGCAGAACTCC-3') generated products of 316bp and 393bp respectively.

The PCR products were ligated into the pGEM-T Easy Vector System which incorporates both T7 and SP6 RNA polymerase sites (Promega, New South Wales, Australia). Plasmids were introduced into competent DH5α cells (New England Biolabs, Ipswich, MA, USA) which were incubated overnight at 37°C. Positively transformed colonies were screened using colony PCR and grown in LB broth containing ampicillin. Plasmids were extracted using a QIAprep® Miniprep Kit (Qiagen, Victoria, Australia) and sequenced (Flinders Sequencing Facility, Adelaide, South Australia) using T7 and SP6 primers allowing the identification of the sequence

and determination of the insert orientation in the plasmid. The plasmids were linearized in separate reactions with Nde 1 and Nco 1 restriction enzymes (New England Biolabs, Ipswich, MA, USA) to provide both antisense (positive) and sense (negative) probes. Lastly, probes were labelled with digoxigenin using a DIG RNA labelling kit (SP6/T7) (Roche, New South Wales, Australia) as per the manufacturer's instructions.

1.13.2. Sample preparation

The complete venom apparatus was obtained from an *O. hannah* specimen which had been milked 4 days prior to removal. The excised tissue was immediately placed in freshly prepared 4% paraformaldehyde and fixed overnight at 4°C. The specimen was bisected lengthwise then processed to paraffin wax via a routine processing schedule. Sections were cut at 4μm, mounted onto silanised slides (HD Scientific, NSW, Australia) and placed at 60°C for 60 minutes.

1.13.3. In situ hybridisation protocol

All *in situ* hybridisation protocols were conducted in an RNase and DNase free environment. Sections were deparaffinised in xylene and rehydrated to water through a graded ethanol series. DEPC treated PBS (0.1mol/L phosphate buffer in 0.75% saline, 0.1% DEPC (Sigma-Aldrich, Missouri, USA); pH 7.2) was applied to the sections for 10 minutes. Sections were then treated with DEPC treated PBS containing 100mmol/L glycine for 10 minutes following which DEPC treated PBS with 0.3% Triton X-100 (Sigma-Aldrich, Missouri, USA) was applied for 15 minutes. Sections were washed with DEPC treated PBS for 10 minutes before 0.1mol/L triethanolamine (Sigma-Aldrich, Missouri, USA) containing 0.3% acetic anhydride (BDH Laboratories, Kampala, Uganda) buffer was applied for 10 minutes. The sections then underwent a prehybridisation step in 4x SSC (saline-sodium citrate) (150nmol/L NaCl, 15mmol/L trisodium citrate), 50% deionised formamide (Sigma-Aldrich, Missouri, USA) in DEPC-treated water) for 30 minutes at 37°C.

Hybridisation buffer (40% deionised formamide (Sigma-Aldrich, Missouri, USA), 10% dextran sulphate (Sigma-Aldrich, Missouri, USA), 1x Denhardt's solution (Sigma-Aldrich, Missouri, USA), 20x SSC, 10mmol/L DTT (Sigma-Aldrich, Missouri, USA), 1mg/ml tRNA (Sigma-Aldrich, Missouri, USA) and 1mg/ml salmon sperm (Sigma-Aldrich, Missouri, USA)) containing 100ng/µl of labelled probe was applied to the sections which were then overlaid with a sterile coverslips and sealed using a rubber sealant. Slides were incubated at 52°C overnight in a Dako Hybridiser (Dako, Victoria, Australia) to allow probes to hybridise.

Following hybridisation, sections were washed in a 2x SSC followed by 1x SSC at 37ºC for 30 minutes duration each. The sections were then rinsed in buffer (0.1mol/L maleic acid, 0.15mol/L NaCl; pH 7.5; 0.3% (w/v) Tween 20) and covered with Dako Dual Enzyme Block Solution (Dako, Victoria, Australia) for 15 minutes. Sections were washed in DEPC treated PBS and an anti-DIG alkaline phosphatase antibody was applied for 1 hour at room temperature. Sections were again washed in buffer (0.1mol/L maleic acid, 0.15mol/L NaCl; pH 7.5; 0.3% (w/v) Tween 20) twice then rinsed in detection buffer (0.1mol/L Tris-HCl (Research Organics, Ohio, USA), 0.1mol/L NaCl, pH 9.5 for 5 minutes. This was followed by application of NBT-BCIP developing solution (Sigma Aldrich, Missouri, USA) for 2 hours. Development was stopped by rinsing in distilled water after which sections were mounted in 90% glycerol (Fronine Laboratory Supplies, Australia) in PBS (0.1mol/L phosphate buffer in 0.75% saline, 0.1% DEPC (Sigma-Aldrich, Missouri, USA); pH 7.2) solution and sealed with nail varnish.

1.13.4. Alcian blue – PAS stain

Sections of venom apparatus were deparaffinised in xylene and rehydrated through a graded series of ethanol. Sections were stained with 1% alcian blue for 5 minutes then rinsed in distilled water. The slides were placed in 1% aqueous periodic acid for 10 minutes. Following oxidation sections were washed in distilled water then placed in Schiff's reagent for 5 minutes. Sections were then washed, dehydrated and mounted in a synthetic mountant.

2. SI FILES

File S1. The microRNA dataset. An Excel spreadsheet displaying (i) microRNA identification, annotation and expression and (ii) *Hox* gene and venom toxin gene targets.

File S2. The assembled king cobra transcriptome. An Excel spreadsheet displaying the annotated, co-assembled king cobra transcriptomes from the venom gland, accessory gland and pooled multi-tissue archive.

File S3. The king cobra venom proteome. An Excel spreadsheet displaying the protein identification of HPLC-fractioned venom components by tandem mass spectrometry

3. SI FIGURES

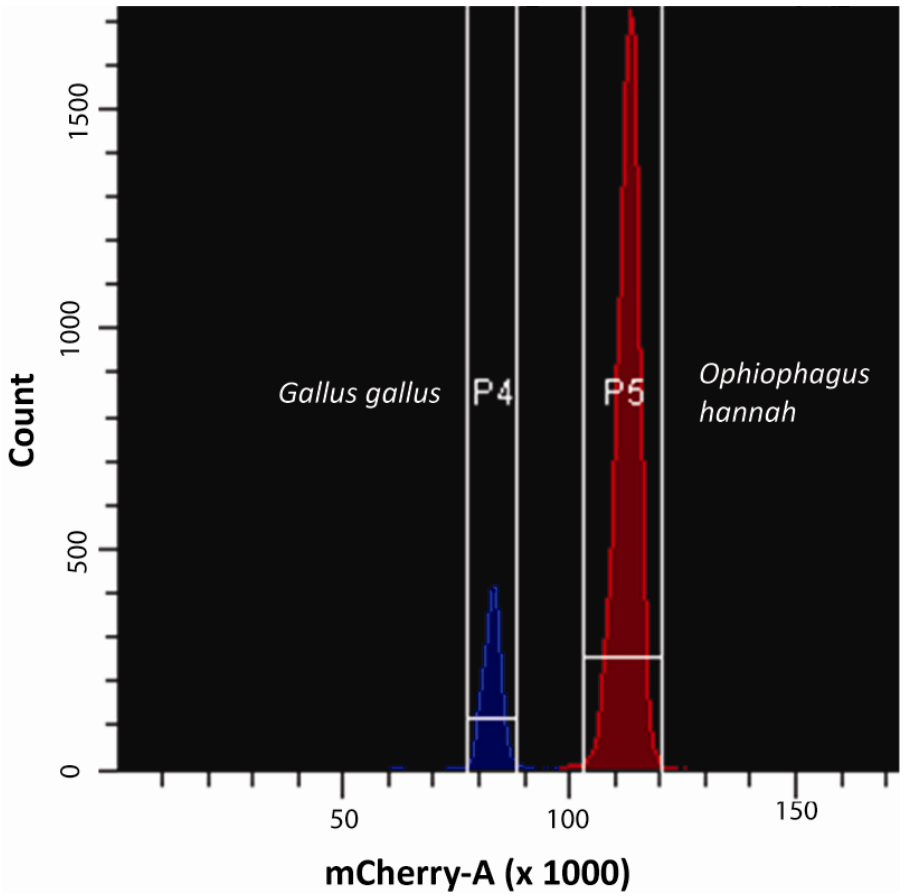

Fig. S1. Size estimation of the king cobra genome. Fluorescence-activated cell sorting analysis used chicken (*Galllus gallus*) erythrocytes as a reference. The size of the chicken genome is 1.05 Gbp and the king cobra (*O. hannah*) haploid genome was estimated to be 1.36-1.59 Gbp (see methods above). mCherry fluorescent proteins were used as marker.

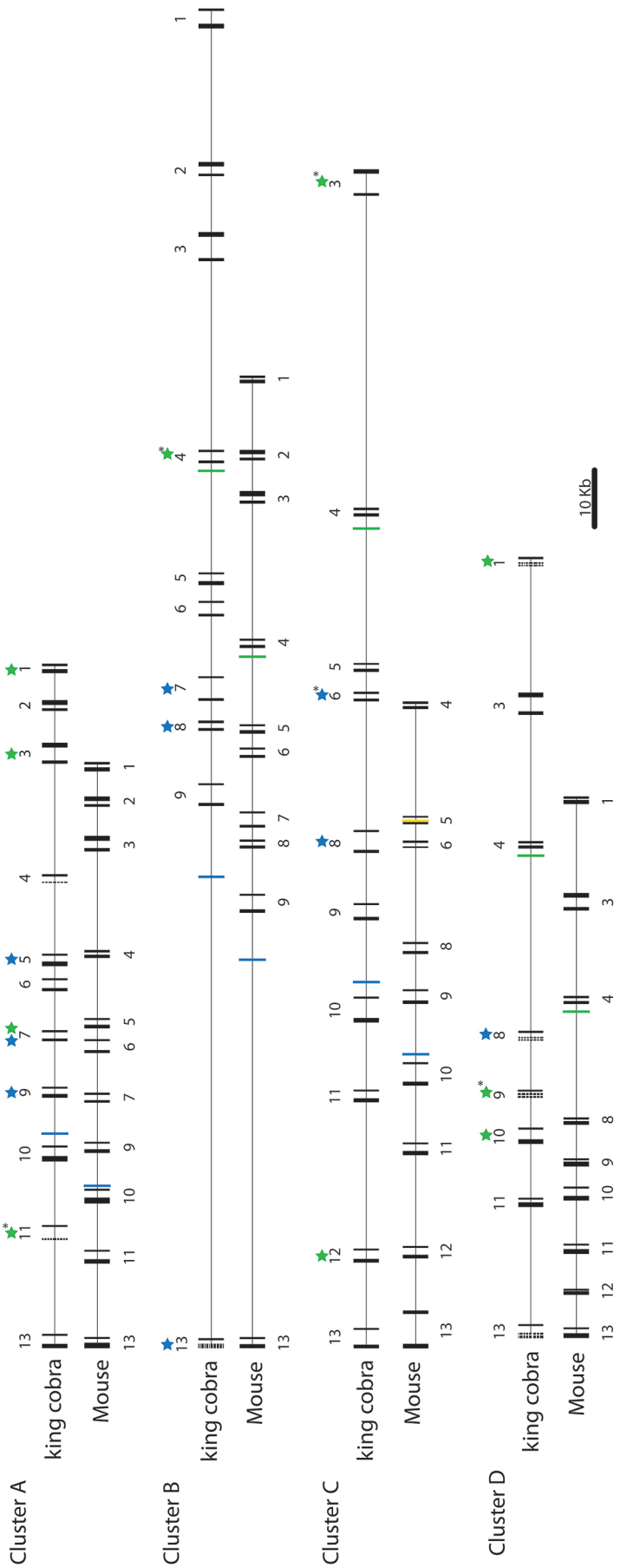
Cluster A
king cobra
Mouse
Cluster B
king cobra
Mouse
Cluster C
king cobra
Mouse
Cluster D
king cobra
Mouse
10 Kb

Fig. S2 (page above). Schematic representation of the king cobra (*Ophiophagus hannah*) and the mouse (*Mus musculus*) *HoxA*, *B, C* and *D* gene clusters. Each black bar represents a successfully annotated exon and for every *Hox* gene, the number of the paralogy group is indicated either above (king cobra) or below (mouse). The dashed lined boxes illustrate exons with unknown boundaries, either due to gaps in sequence or to possible errors in the assembly. The microRNA genes located within the *Hox* clusters are indicated by coloured bars. In blue is *miR-196*, in green *miR-10* and in orange *miR-615* (mouse only). Predicted targets for the 5p mature sequences of *miR-196* and *miR-10* are indicated by a star directly above the target *Hox* genes. Blue stars are targets of *miR-196* and green stars are targets of *miR-10* (see File S1 for details). Asterisks denote targets in the king cobra, which are not predicted for mouse or human. The king cobra *Hox* genes, as in other vertebrates, were found clustered at four distinct genomic loci, but the gene clusters are substantially larger than in mammals, with a 10 to 40 percent increase in size for the *HoxA* and *HoxD* clusters, respectively. This expansion in size was mainly due to the presence of repeated elements, a peculiarity that seems to be a genomic synapomorphy of the squamate reptiles, as similar observations have been described in the corn snake and *Anolis* lizard and not been reported in other vertebrate taxa. In addition, the king cobra contains a *Hoxc3* gene that is found in other squamate reptiles and amphibians, but absent in mammals. Finally, the *HoxD* cluster lacks *Hoxd12* (see also Fig. S3), a gene involved in tetrapod limb morphogenesis.

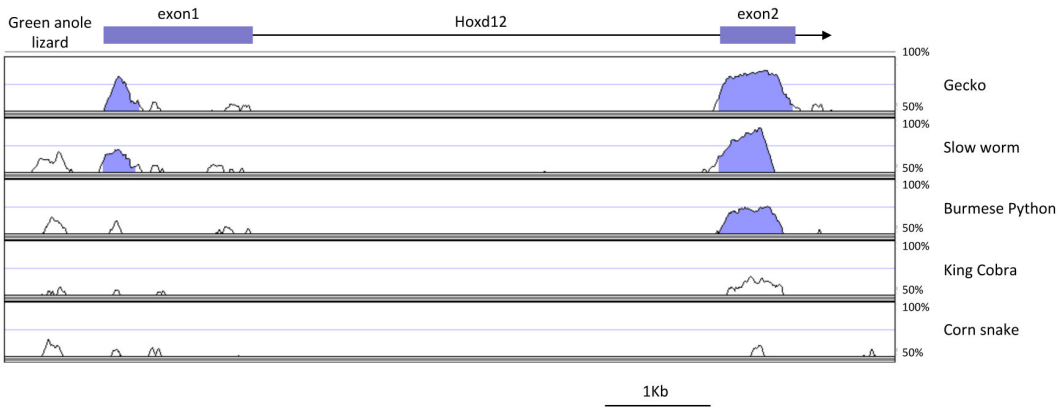

Fig. S3. Conservation plot of six squamate *Hoxd12* sequences. A schematic representation of the green anole lizard (*Anolis carolinensis*) *Hoxd12* gene was used as a reference and is depicted above the plot. Blue boxes represent exons. The peaks represent percentage of conservation (50 to 100%) to the reference sequence and blue peaks represent levels of conservation higher than 70% in exons. Note the loss of sequence similarity in exon 1 of the three snake species. Gecko, *Gekko ulikovskii*; slow worm, *Anguis fragilis*; Burmese python, *Python molurus bivittatus*; king cobra, *Ophiophagus hannah*; corn snake, *Pantherophis guttatus*.

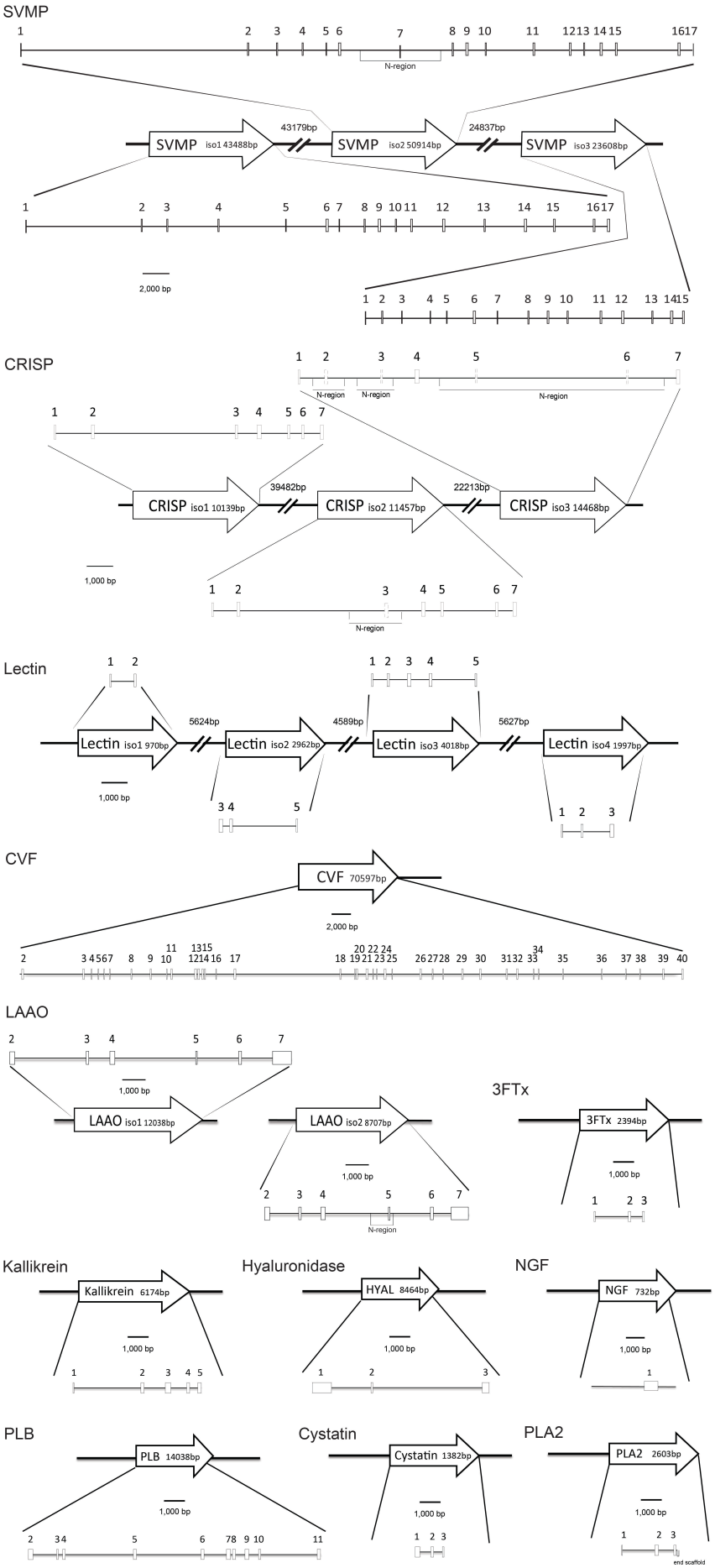
SVMP
N-region
SVMP iso1 43488bp
43179bp
SVMP iso2 50914bp
24837bp
SVMP iso3 23608bp
2,000 bp
CRISP
N-region
N-region
N-region
CRISP iso1 10139bp
39482bp
CRISP iso2 11457bp
22213bp
CRISP iso3 14468bp
1,000 bp
N-region
Lectin
Lectin iso1 970bp
5624bp
Lectin iso2 2962bp
4589bp
Lectin iso3 4018bp
5627bp
Lectin iso4 1997bp
1,000 bp
CVF
CVF 70597bp
2,000 bp
LAAO
1,000 bp
LAAO iso1 12038bp
LAAO iso2 8707bp
1,000 bp
N-region
3FTx
3FTx 2394bp
1,000 bp
Kallikrein
Kallikrein 6174bp
1,000 bp
Hyaluronidase
HYAL 8464bp
1,000 bp
NGF
NGF 732bp
1,000 bp
PLB
PLB 14038bp
1,000 bp
Cystatin
Cystatin 1382bp
1,000 bp
PLA2
PLA2 2603bp
1,000 bp
end scaffold

Fig. S4 (page above). Annotation of the macrostructures of venom gene families found in the king cobra genome. The intron-exon boundaries of each toxin family are displayed schematically, where arrows represent genes found on genome scaffolds and cut-outs (represented by lines and white blocks) demonstrate the genomic arrangement of representative genes from each toxin family. In the cut-outs, white blocks represent exons and black lines introns. Numbers highlight the number of exons present in each gene. Where available, toxins from the same gene family that are found downstream of each other on the same genomic scaffold are displayed (SVMP, CRISP, lectin). SVMP, snake venom metalloproteinase; CRISP, cysteine-rich secretory protein; CVF, cobra venom factor; LAAO, L-amino acid oxidase; 3FTx, three-finger toxin; NGF, nerve growth factor; PLB, phospholipase-B; PLA2, phospholipase A_2.

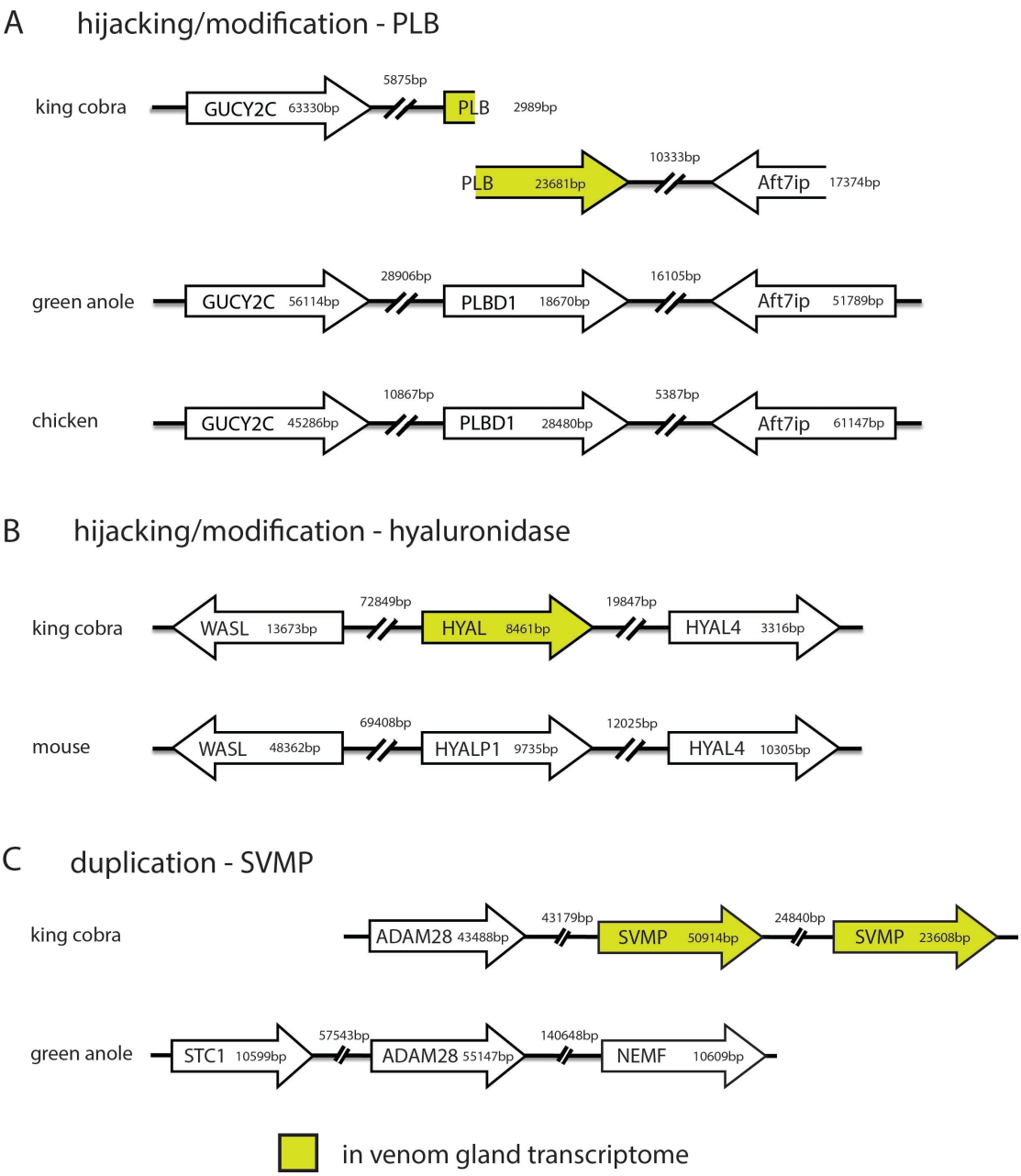

Fig. S5. Syntenic comparisons of venom genes in the king cobra with other vertebrates revealing toxin recruitment by hijacking/modification and gene duplication. (A) modification of PLBD1 gene found in the green anole lizard (*Anolis carolinensis*) and the chicken (*Gallus gallus*) results in the venom gland expressed phospholipase-B (PLB). Note that PLB is found split across two king cobra genome scaffolds. (B) modification of HYALP1 gene found in the mouse (*Mus musculus*) results in the venom gland expressed hyaluronidase (HYAL). (C) duplication of the non-venom gland expressed ADAM gene in the king cobra results in a venom gland expressed snake venom metalloproteinase (SVMP) gene. The ADAM gene in the

green anole (*A. carolinensis*) is flanked on both sides by non-SVMP genes, demonstrating the absence of gene duplication in this species. Note that subsequent downstream duplication of the SVMP gene in the king cobra results in multiple venom gland expressed SVMP isoforms.

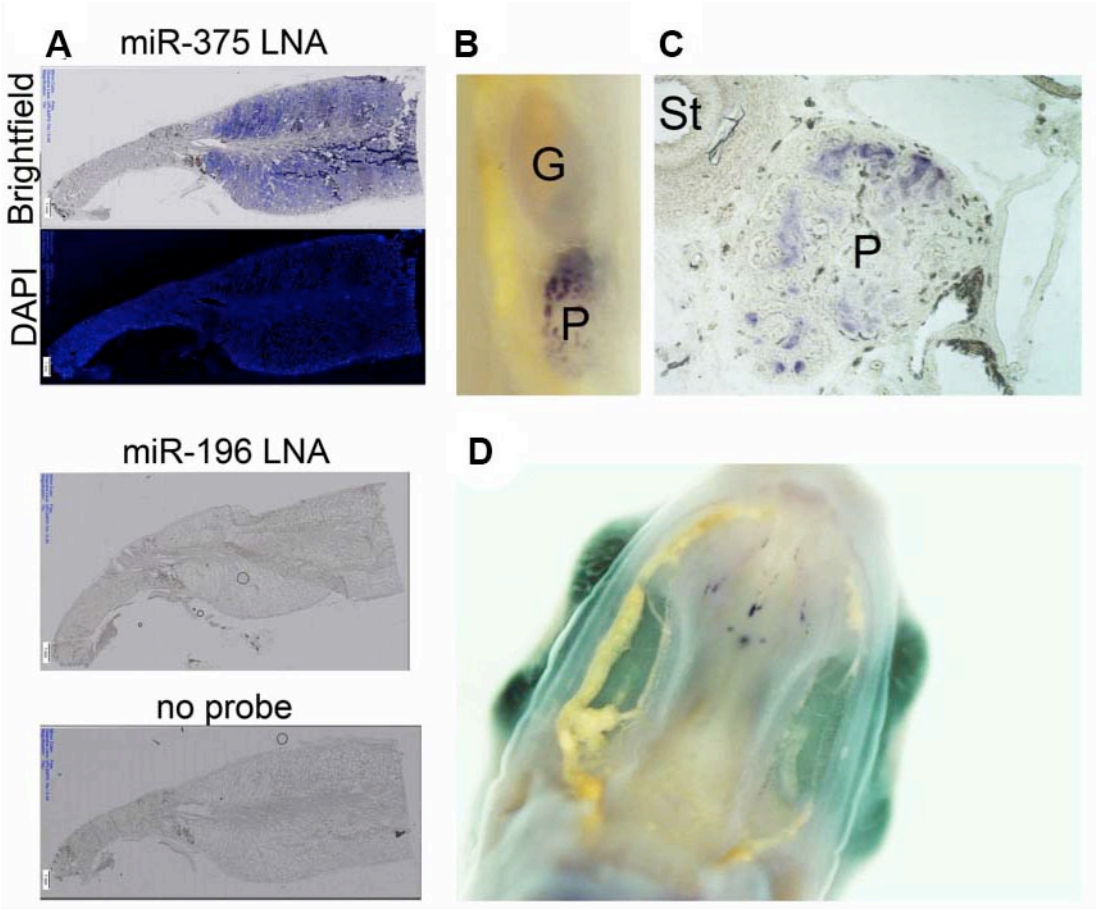

Fig. S6. LNA *in situ* detection of *miR-375* in the adult king cobra venom apparatus and *Coelognathus radiatus* 27 dpo embryo. (A) *miR-375* is detected in the main venom gland of the king cobra whereas neither *miR-196* LNA probe or "no probe" resulted in alkaline phosphatase activity. *miR-196* is not detected in the miRNA libraries from the venom gland or accessory gland; however, it is detected in the pooled multi-tissue library. (B) Higher magnification of the pancreas and spleen from *C. radiatus* 27 dpo whole mount *in situ* shown in Fig. 2B. (C) Tissue section of pancreas shown in B. (D) Detection of *miR-375* in the palate of the *C. radiatus* 27 dpo embryo shown in Fig. 2B. G, gall bladder; P, pancreas; S, spleen; St, stomach.

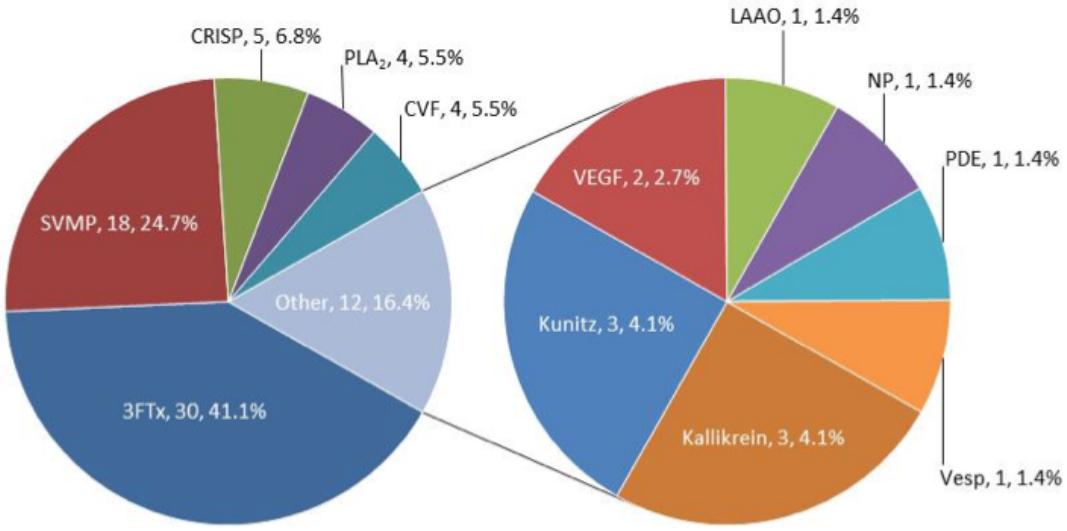

Fig. S7. Schematic representation of the proteome of the king cobra, *Ophiophagus hannah*. Overall, a total of 73 protein sequences were detected by matching at least two peptides. Any redundancies in matches were removed to give the most conservative proteome possible. In the case of the SVMPs, seven partial sequences were combined to yield three more complete sequences without peptide match overlaps. For each segment, the identification is given followed by the number of unique proteins and the overall percentage of the proteome covered by that toxin family. Matched proteins without a putative signal peptide were not included in this analysis because they are not secreted into the venom and represent structural proteins released from lysed cells during venom extraction. 3FTx, three-finger toxin; CRISP, cysteine-rich secretory protein; CVF, cobra venom factor; Kunitz, Kunitz-type serine protease inhibitor; LAAO, L-amino acid oxidase; NP, natriuretic peptide; PDE, phosphodiesterase; PLA_2, phospholipase A_2, kallikrein, kallikrein-type serine proteases; SVMP, snake venom metalloproteinase; VEGF, vascular endothelial growth factor; Vesp, vespryn.

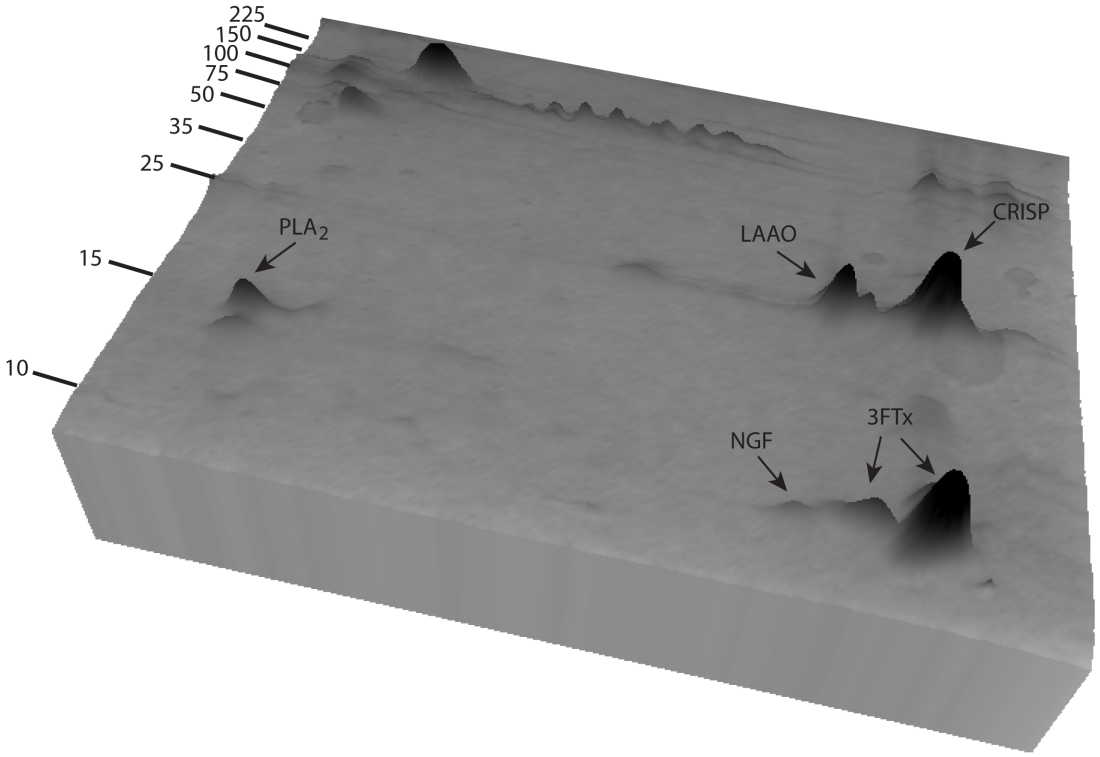

Fig. S8. 2D gel electrophoresis profile of the king cobra venom proteome. Major protein stained bands estimated to be smaller than 25 kDa in size were excised, in-gel trypsin digested and identified by tandem mass spectrometry. Comparative estimates of protein abundance were calculated by the intensity of protein staining and are indicated by the height of protein peaks. Bands to the left of the gel indicate molecular weight in kDa. 3FTx, three finger toxin; CRISP, cysteine rich secretory protein; LAAO, L-amino acid oxidase; NGF, nerve growth factor; PLA_2, phospholipase A_2.

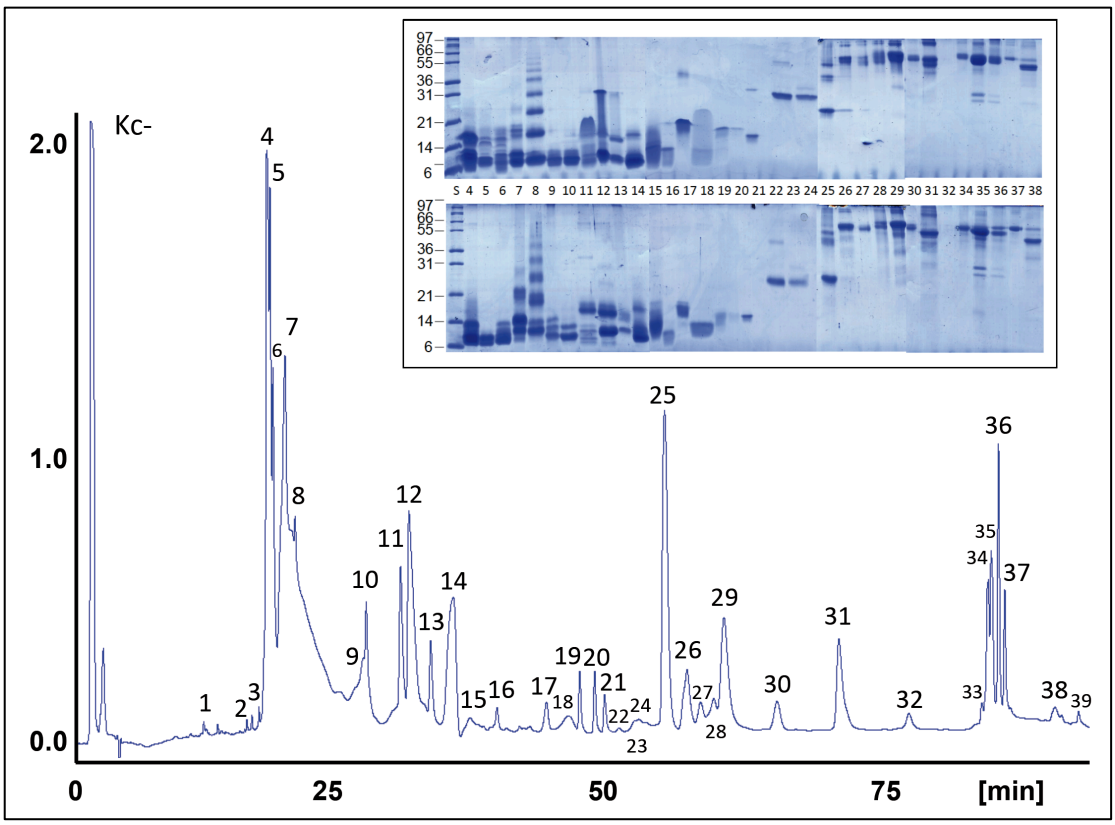

Fig. S9. Reverse-phase HPLC separation of king cobra venom proteins. Insert, SDS-PAGE analysis of the isolated, numbered, protein fractions.

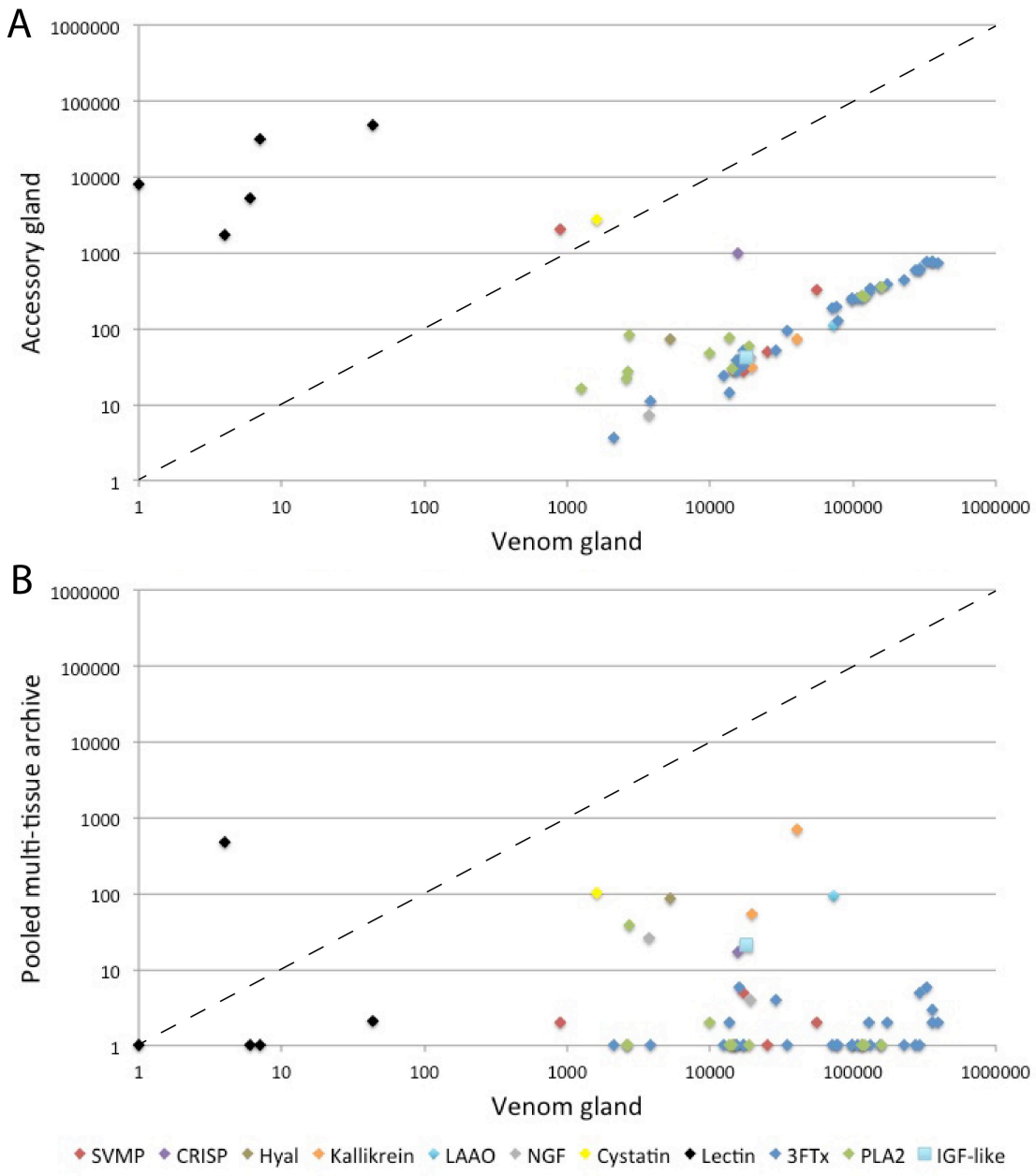

Fig. S10. **Comparisons of expressed toxin genes in different king cobra tissues.** (A) Venom gland and accessory gland. (B) Venom gland and pooled multi-tissue archive. Normalised tag counts from the venom gland, accessory gland and pooled multi-tissue transcriptomes were plotted against each other on logarithmic axes. Where expression was absent, toxins were given an arbitrary value of 1 to permit logarithmic plotting. The expression levels of toxin-related genes is typically at least ten-fold greater in the venom gland than the accessory gland. The exception to this is the lectin toxin family where expression is significantly greater ($P < 0.05$) in the accessory gland (black diamonds). The dotted line indicates 1:1 expression.

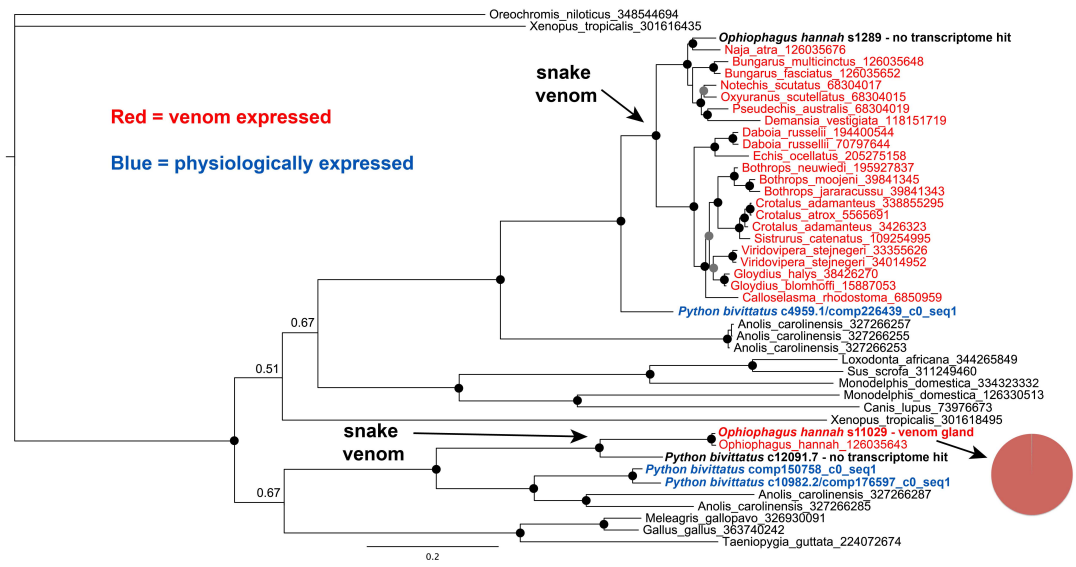

Fig. S11. Bayesian DNA gene tree of the L-amino acid oxidase toxin family. King cobra and Burmese python genes are labelled bold. The tips of the tree coloured in red indicate sequences sourced from the snake venom gland and blue those sourced from non-venom gland 'physiological' tissues. Circles placed at internal tree nodes indicate the Bayesian posterior probabilities (bpp) for that node – black: bpp = 1.00; grey: bpp $\geq$ 0.95. Pie charts display the proportional, normalised, transcriptomic expression profile of king cobra genes, where red = venom gland; orange = accessory gland; blue = pooled multi-tissue archive. Note the two occasions that L-amino acid oxidase genes have been co-opted for a role in snake venom.

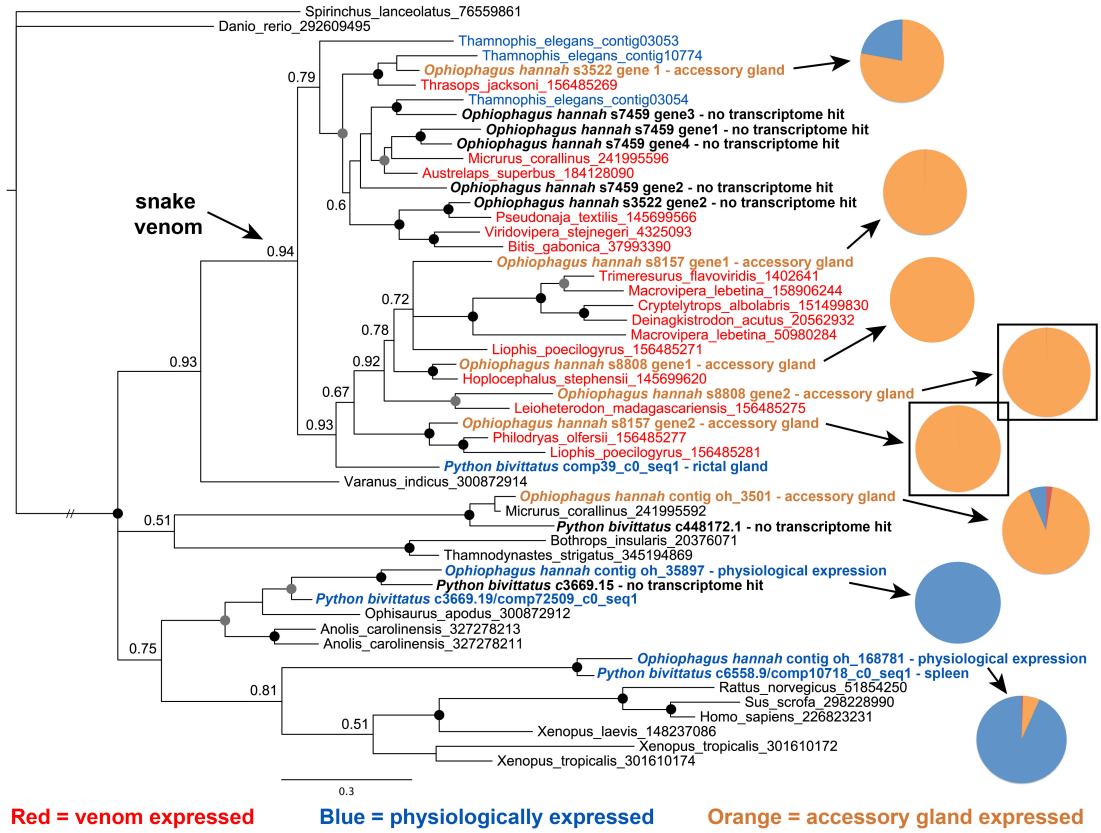

Fig. S12. Bayesian DNA gene tree of the lectin toxin family. King cobra and Burmese python genes are labelled bold. The tips of the tree coloured in red indicate sequences sourced from the snake venom gland, orange those sourced from the accessory gland and blue those sourced from non-venom gland 'physiological' tissues. Where coexpression was identified, the location of highest expression was used for colouring. Circles placed at internal tree nodes indicate the Bayesian posterior probabilities (bpp) for that node – black: bpp = 1.00; grey: bpp ≥ 0.95. Pie charts display the proportional, normalised, transcriptomic expression profile of king cobra genes, where red = venom gland; orange = accessory gland; blue = pooled multi-tissue archive. Pie charts highlighted by boxes indicated those genes where the spatial expression of lectins in the accessory gland was validated by *in situ* hybridisation (see Fig. S14). See Fig. S13 for explicit tests of character state changes in the lectin gene family.

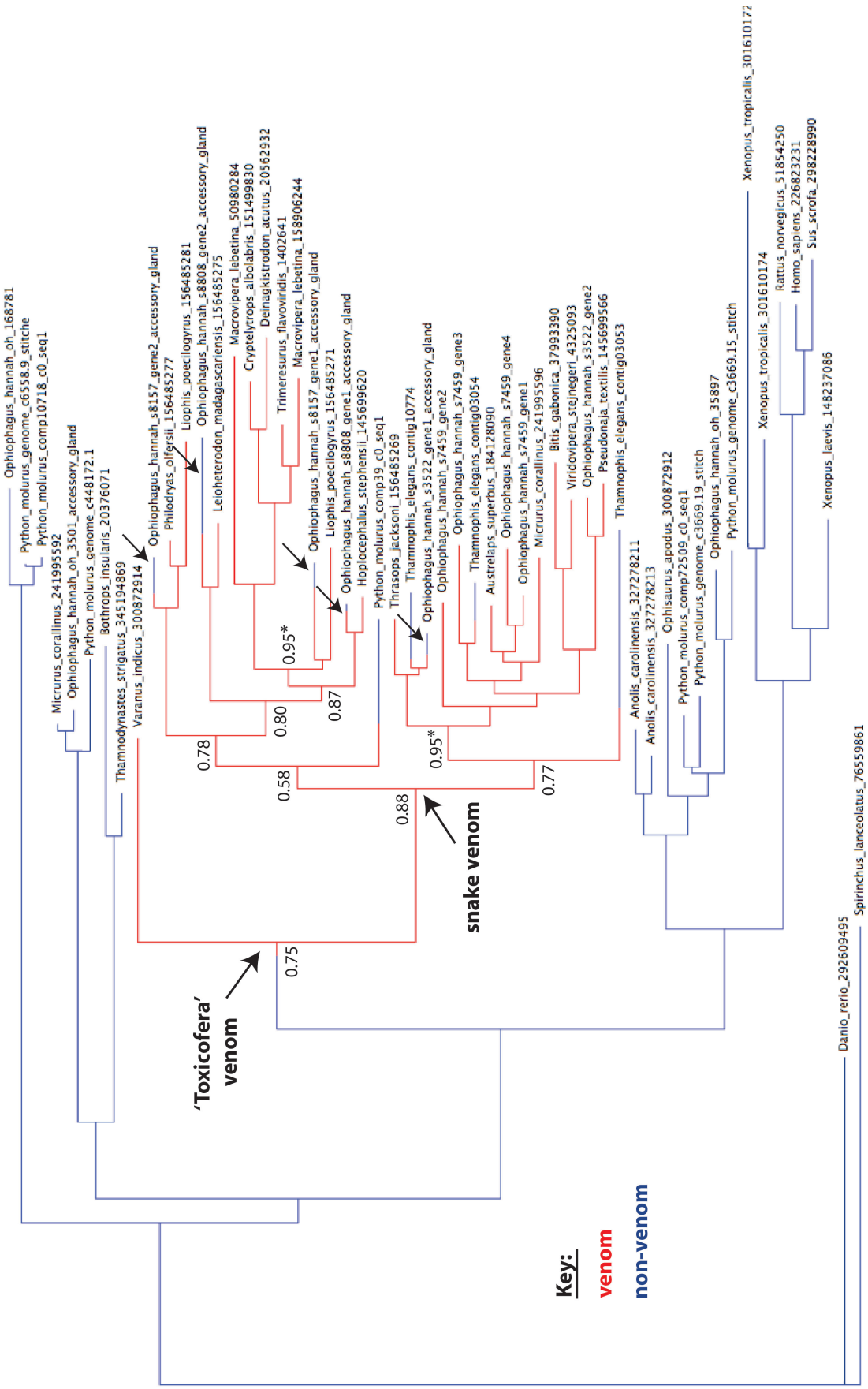
Ophiophagus_hannah_oh_168781
Python_molurus_genome_c6558.9_stitche
Python_molurus_comp10718_c0_seq1
Micrurus_corallinus_241995592
Ophiophagus_hannah_oh_3501_accessory_gland
Python_molurus_genome_c448172.1
Bothrops_insularis_20376071
Thamnodynastes_strigatus_345194869
Varanus_indicus_300872914
Ophiophagus_hannah_s8157_gene2_accessory_gland
Philodryas_olfersii_156485277
Liophis_poecilogyrus_156485281
Ophiophagus_hannah_s8808_gene2_accessory_gland
Leioheterodon_madagascariensis_156485275
Macrovipera_lebetina_50980284
Cryptelytrops_albolabris_151499830
Deinagkistrodon_acutus_20562932
Trimeresurus_flavoviridis_1402641
Macrovipera_lebetina_158906244
Ophiophagus_hannah_s8157_gene1_accessory_gland
Liophis_poecilogyrus_156485271
Ophiophagus_hannah_s8808_gene1_accessory_gland
Hoplocephalus_stephensii_145699620
Python_molurus_comp39_c0_seq1
Thrasops_jacksonI_156485269
Thamnophis_elegans_contig10774
Ophiophagus_hannah_s3522_gene1_accessory_gland
Ophiophagus_hannah_s7459_gene2
Ophiophagus_hannah_s7459_gene3
Thamnophis_elegans_contig03054
Austrelaps_superbus_184128090
Ophiophagus_hannah_s7459_gene4
Ophiophagus_hannah_s7459_gene1
Micrurus_corallinus_241995596
Bitis_gabonica_37993390
Viridovipera_stejnegeri_4325093
Ophiophagus_hannah_s3522_gene2
Pseudonaja_textilis_145699566
Thamnophis_elegans_contig03053
Anolis_carolinensis_327278211
Anolis_carolinensis_327278213
Ophisaurus_apodus_300872912
Python_molurus_comp72509_c0_seq1
Python_molurus_genome_c3669.19_stitch
Ophiophagus_hannah_oh_35897
Python_molurus_genome_c3669.15_stitch
Xenopus_tropicalis_301610172
Xenopus_tropicalis_301610174
Rattus_norvegicus_51854250
Homo_sapiens_226823231
Sus_scrofa_298228990
Xenopus_laevis_148237086
Danio_rerio_292609495
Spirinchus_lanceolatus_76559861
0.78
0.80
0.95*
0.87
0.58
0.95*
0.88
0.77
0.75
'Toxicofera' venom
snake venom
Key:
venom
non-venom

Fig. S13 (page above). Mapping changes in the sites of lectin expression to the gene family phylogeny. Changes in character states (red, venom; blue, non-venom) were mapped onto the lectin gene tree revealing a single recruitment event of lectins into venom. Support for the timing of this event was provided by ancestral state analyses, with an origin prior to the Toxicoferan (56) diversification reasonably supported (0.75), and a single origin prior to the snake diversification (0.88) approaching the significance of the test ($\geq$0.95). Small arrows indicate the five occasions that the expression of king cobra lectin paralogs have been transferred from the venom gland to the accessory gland. In all cases, nodes within the venom clade that contained king cobra accessory gland expressed lectins approached the significance of the test for the 'venom' character state. Two of these 'reversal events' (28) are significant ($\geq$0.95). This demonstrates that lectins with a venom origin appear to have been repeatedly recruited for expression in the accessory gland. Asterisks indicate significance at the node and small arrows highlight the relevant 'reversed' king cobra branches.

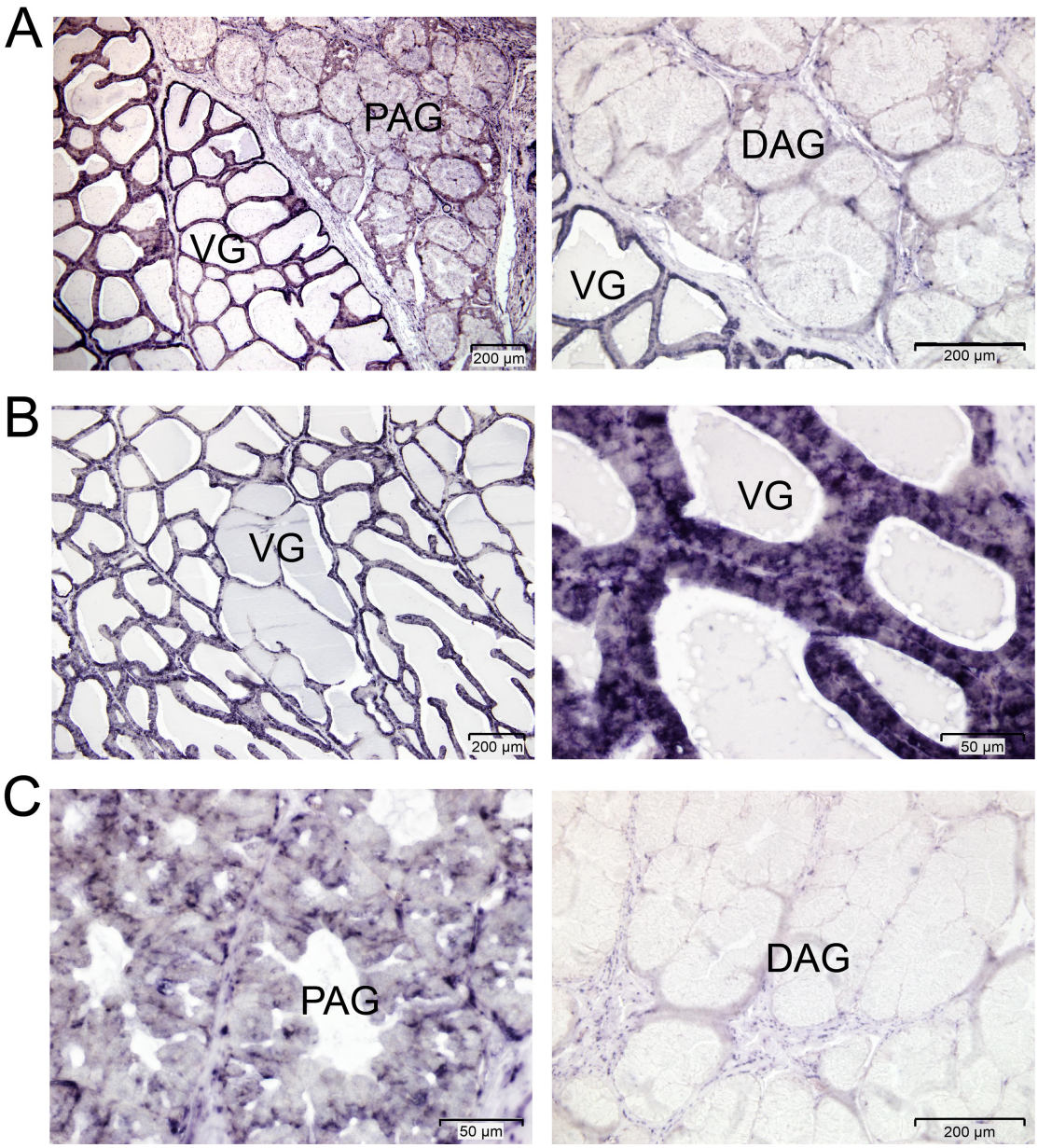

Fig. S14. *in situ* hybridisation of lectin genes in the venom system of the king cobra. (A) *in situ* hybridisation of lectin gene Oh-3509 (genome id s8157 gene 2) showing extensive staining in the venom gland (VG) and the proximal portion of the accessory gland (PAG) (left) and absence of staining in the distal portion of the accessory gland (DAG) (right). Staining in the VG is consistent with the low levels of lectin transcripts observed in the king cobra venom gland transcriptome. Staining in the VG was observed in the apical regions of the secretory epithelial cells, with the columnar epithelium more prominently stained compared with cuboidal epithelial cells. Staining in the epithelium of the PAG was confined to the basal regions of the cell. Unstained mucous vacuoles comprise the bulk of the cytoplasm. (B) *in situ*

hybridisation of lectin gene Oh-516 (genome id s8808 gene 2) showing staining in the secretory epithelial cells in the VG. (C) *in situ* hybridisation of lectin gene Oh-516 showing granular staining in the epithelium of the PAG (left) and absence of staining in the DAG (right).

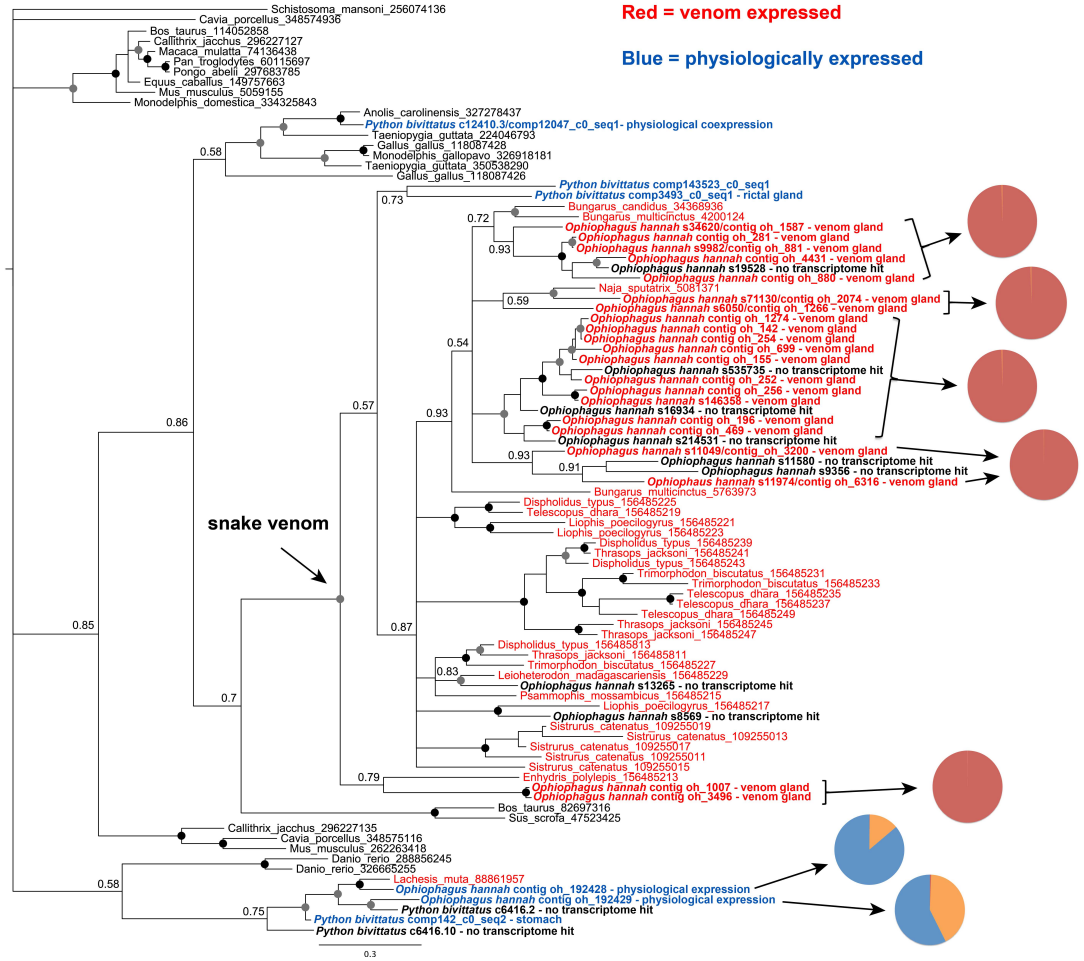

Fig. S15. Bayesian DNA gene tree of the three-finger toxin family. King cobra and Burmese python genes are labelled bold. The tips of the tree coloured in red indicate sequences sourced from the snake venom gland and blue those sourced from non-venom gland 'physiological' tissues. Circles placed at internal tree nodes indicate the Bayesian posterior probabilities (bpp) for that node – black: bpp = 1.00; grey: bpp ≥ 0.95. Pie charts display the proportional, normalised, transcriptomic expression profile of king cobra genes, where red = venom gland; orange = accessory gland; blue = pooled multi-tissue archive. Burmese python gene comp12047 was identified as coexpressed physiologically (ovary, rictal gland and testes).

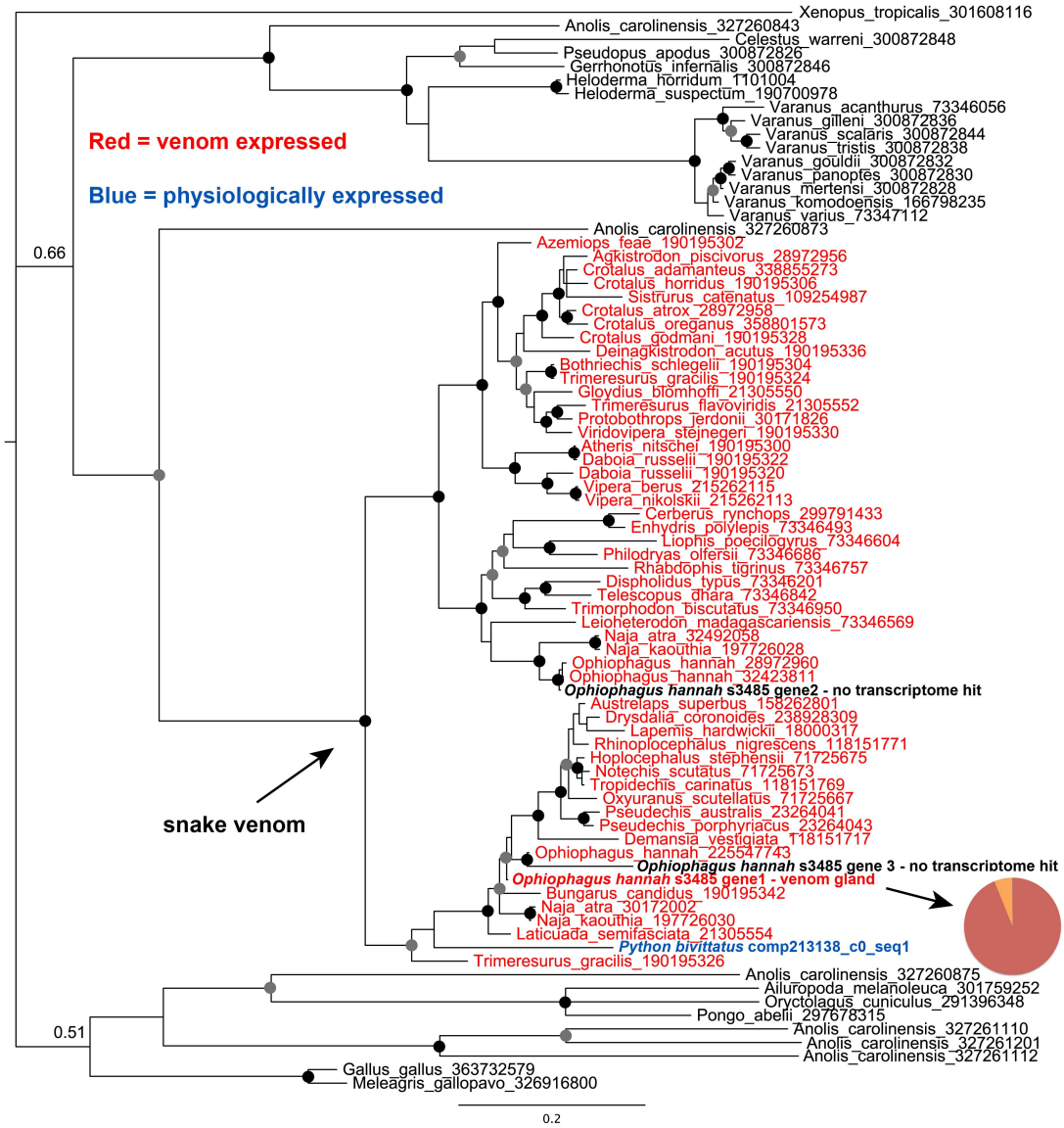

Fig. S16. Bayesian DNA gene tree of the CRISP toxin family. King cobra and Burmese python genes are labelled bold. The tips of the tree coloured in red indicate sequences sourced from the snake venom gland and blue those sourced from non-venom gland 'physiological' tissues. Circles placed at internal tree nodes indicate the Bayesian posterior probabilities (bpp) for that node – black: bpp = 1.00; grey: bpp ≥ 0.95. Pie charts display the proportional, normalised, transcriptomic expression profile of king cobra genes, where red = venom gland; orange = accessory gland; blue = pooled multi-tissue archive.

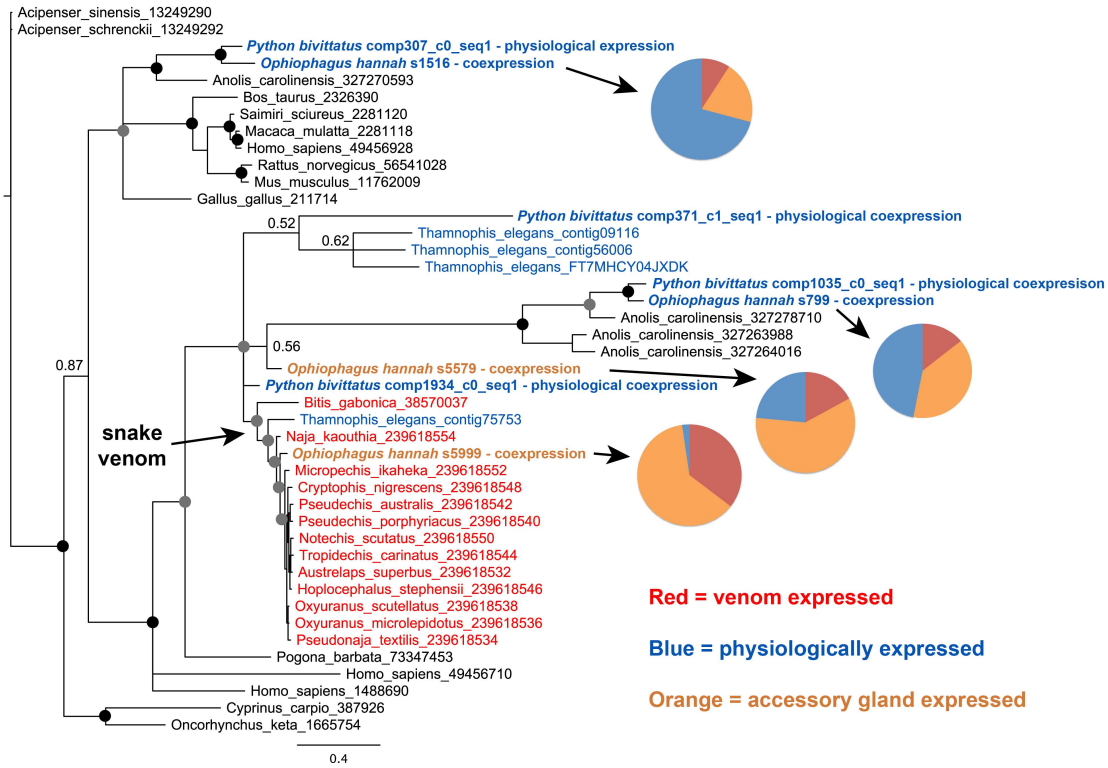

Fig. S17. Bayesian DNA gene tree of the cystatin toxin family. King cobra and Burmese python genes are labelled bold. The tips of the tree coloured in red indicate sequences sourced from the snake venom gland, orange those sourced from the accessory gland and blue those sourced from non-venom gland 'physiological' tissues. Where coexpression was identified, the location of highest expression was used for colouring. Circles placed at internal tree nodes indicate the Bayesian posterior probabilities (bpp) for that node – black: bpp = 1.00; grey: bpp $\geq$ 0.95. Pie charts display the proportional, normalised, transcriptomic expression profile of king cobra genes, where red = venom gland; orange = accessory gland; blue = pooled multi-tissue archive. Burmese python genes were identified as coexpressed physiologically (*Python bivittatus* comp371, comp1035 and comp307 in blood, muscle, ovary, rictal gland, spleen, stomach and testes; comp1934 in ovaries, rictal gland, stomach and testes).

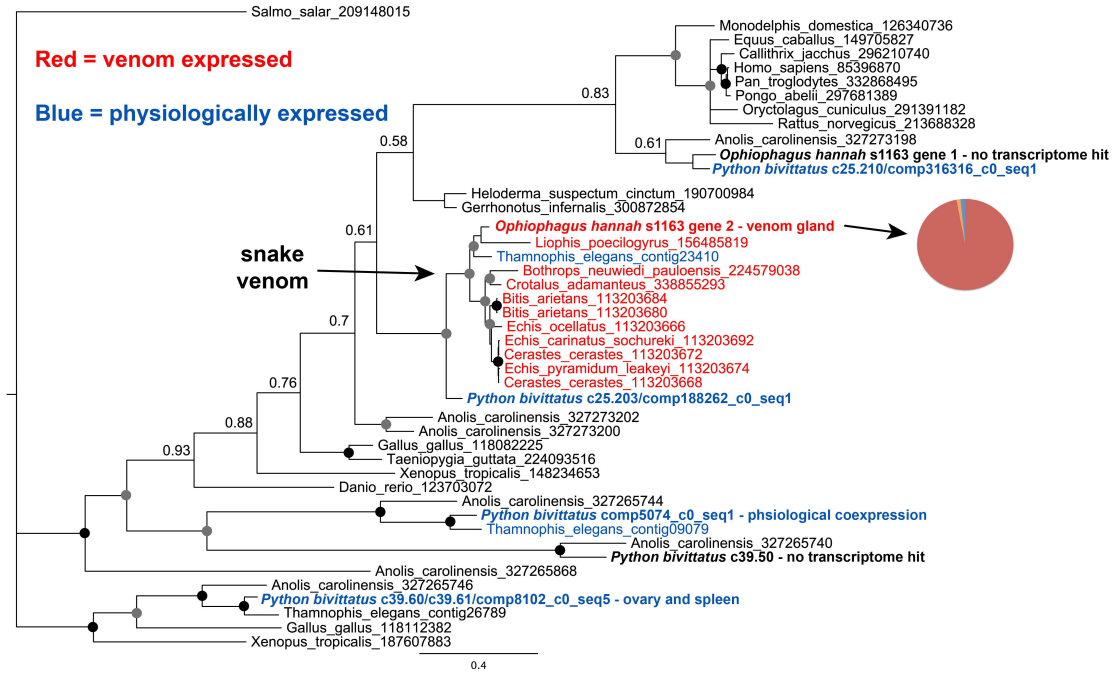

Fig. S18. Bayesian DNA gene tree of the hyaluronidase toxin family. King cobra and Burmese python genes are labelled bold. The tips of the tree coloured in red indicate sequences sourced from the snake venom gland and blue those sourced from non-venom gland 'physiological' tissues. Circles placed at internal tree nodes indicate the Bayesian posterior probabilities (bpp) for that node – black: bpp = 1.00; grey: bpp $\geq$ 0.95. Pie charts display the proportional, normalised, transcriptomic expression profile of king cobra genes, where red = venom gland; orange = accessory gland; blue = pooled multi-tissue archive. Burmese python gene comp5074 was identified as coexpressed physiologically (ovary, rictal gland, stomach and testes).

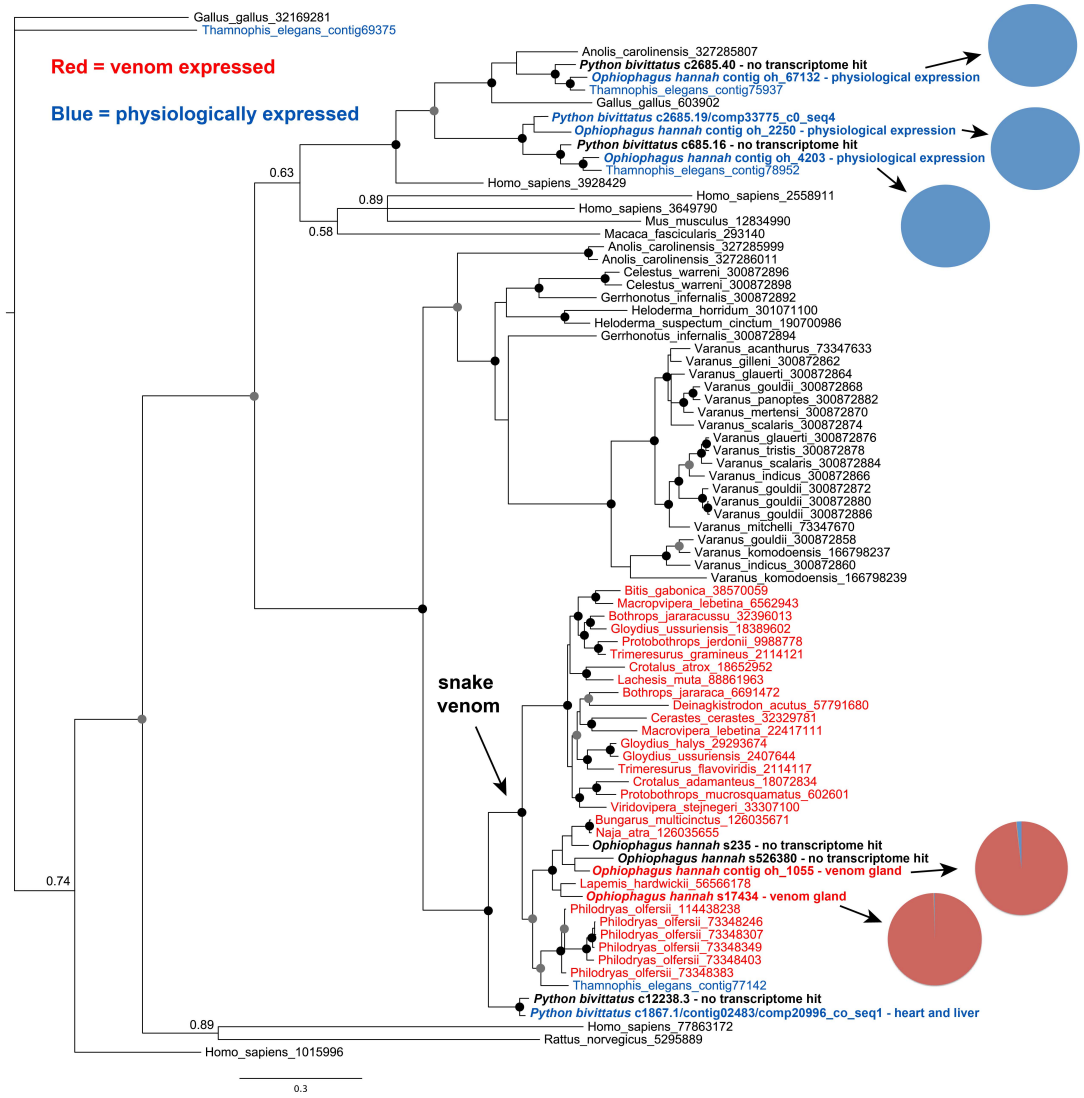

Fig. S19. Bayesian DNA gene tree of the kallikrein toxin family. King cobra and Burmese python genes are labelled bold. The tips of the tree coloured in red indicate sequences sourced from the snake venom gland and blue those sourced from non-venom gland 'physiological' tissues. Circles placed at internal tree nodes indicate the Bayesian posterior probabilities (bpp) for that node – black: bpp = 1.00; grey: bpp ≥ 0.95. Pie charts display the proportional, normalised, transcriptomic expression profile of king cobra genes, where red = venom gland; orange = accessory gland; blue = pooled multi-tissue archive.

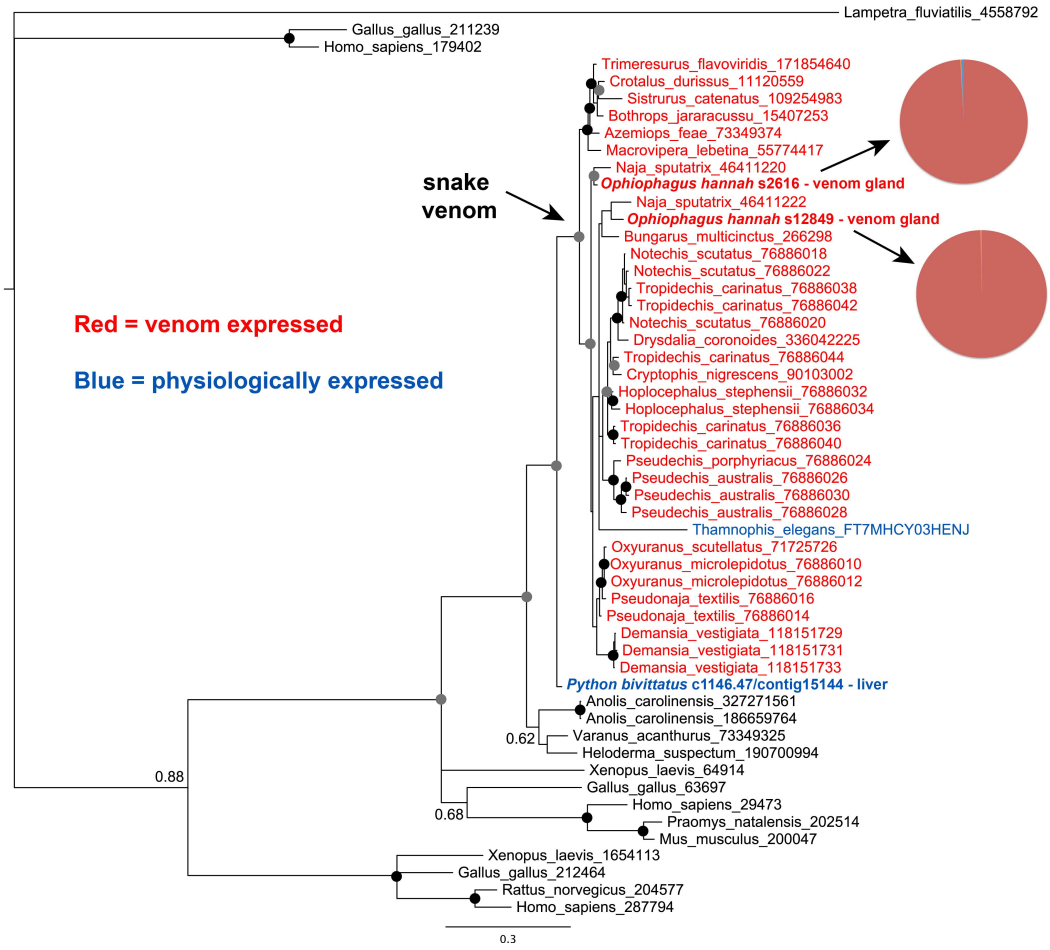

Fig. S20. Bayesian DNA gene tree of the nerve growth factor toxin family. King cobra and Burmese python genes are labelled bold. The tips of the tree coloured in red indicate sequences sourced from the snake venom gland and blue those sourced from non-venom gland 'physiological' tissues. Circles placed at internal tree nodes indicate the Bayesian posterior probabilities (bpp) for that node – black: bpp = 1.00; grey: bpp $\geq$ 0.95. Pie charts display the proportional, normalised, transcriptomic expression profile of king cobra genes, where red = venom gland; orange = accessory gland; blue = pooled multi-tissue archive.

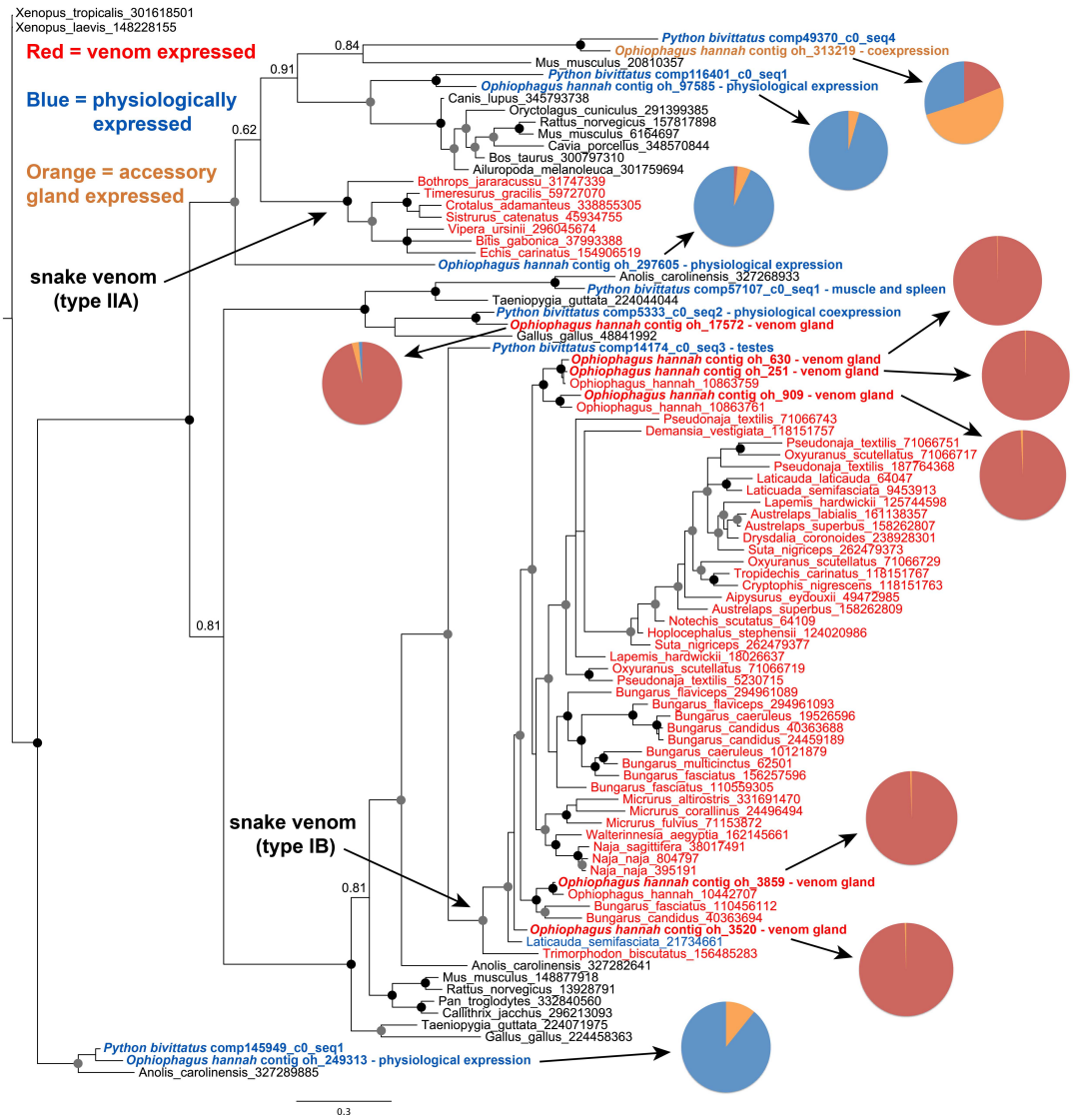

Fig. S21. Bayesian DNA gene tree of the phospholipase A_2 toxin family. King cobra and Burmese python genes are labelled bold. The tips of the tree coloured in red indicate sequences sourced from the snake venom gland, orange those sourced from the accessory gland and blue those sourced from non-venom gland 'physiological' tissues. Where coexpression was identified, the location of highest expression was used for colouring. Circles placed at internal tree nodes indicate the Bayesian posterior probabilities (bpp) for that node – black: bpp = 1.00; grey: bpp ≥ 0.95. Pie charts display the proportional, normalised, transcriptomic expression profile of king cobra genes, where red = venom gland; orange = accessory gland; blue = pooled multi-tissue archive. Burmese python gene comp5333 was identified as coexpressed physiologically (blood, ovary, rictal gland, spleen, stomach and testes).

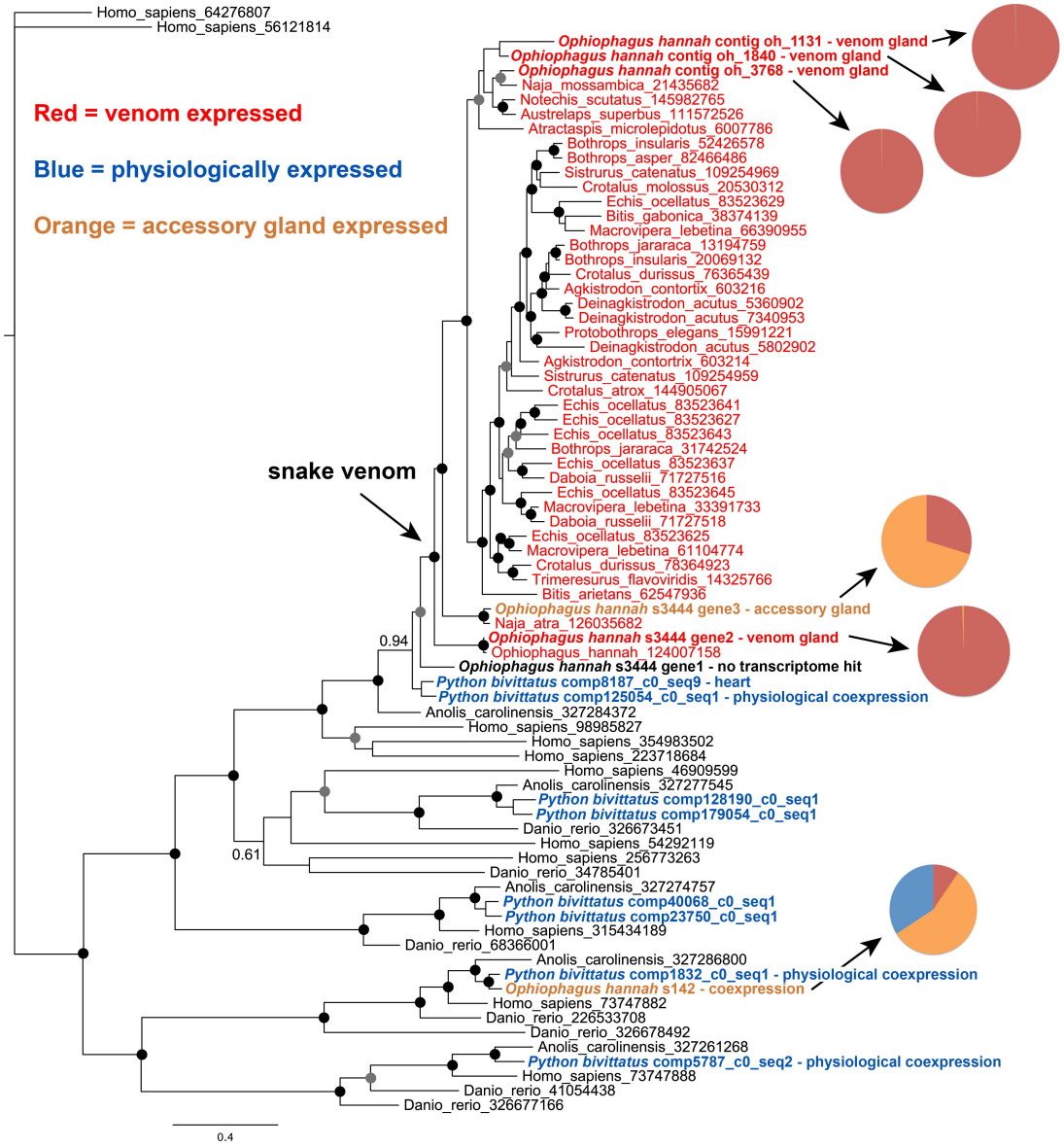

Fig. S22. Bayesian DNA gene tree of the snake venom metalloproteinase toxin family. King cobra and Burmese python genes are labelled bold. The tips of the tree coloured in red indicate sequences sourced from the snake venom gland, orange those sourced from the accessory gland and blue those sourced from non-venom gland 'physiological' tissues. Where coexpression was identified, the location of highest expression was used for colouring. Circles placed at internal tree nodes indicate the Bayesian posterior probabilities (bpp) for that node – black: bpp = 1.00; grey: bpp ≥ 0.95. Pie charts display the proportional, normalised, transcriptomic expression profile of king cobra genes, where red = venom gland; orange = accessory gland; blue = pooled multi-tissue archive. Burmese python genes comp12504, comp1832 and comp5787 were identified as coexpressed physiologically (12504: blood, rictal gland,

spleen and testes; 1832: blood, muscle, ovary, rictal gland, spleen and stomach; 5787: blood, ovary, rictal, spleen and stomach).

4. SI TABLES

Table S1. Sequencing statistics for libraries used for genome and transcriptome construction.

Library name	Library type	Insert size (bp)[1]	Read length	Raw sequence	Clean sequence[2]	Scaffolding links[2]
PE200	Paired-end	60-127	2x76 nt	21.9 Gbp	16.8 Gbp	n.a.
PE500	Paired-end	122-478	2x50 nt	8.5 Gbp	7.9 Gbp	43M
			2x151 nt	10.8 Gbp	9.8 Gbp	
MP2K	Mate pair	1600-2400	2x36 nt	5.4 Gbp	n.a.	3.4M
MP7K	Mate pair	2500-6000	2x51 nt	2.3 Gbp	n.a.	181K
MP10K	Mate pair	6500-10000	2x51 nt	5.3 Gbp	n.a.	1.4M
MP15K	Mate pair	9000-13000	2x51 nt	3.8 Gbp	n.a.	1.2M
Venom gland	mRNA-Seq	n.a.	51 nt	0.77 Gbp	n.a.	n.a.
Accessory gland	mRNA-Seq	n.a.	51 nt	0.57 Gbp	n.a.	n.a.
Pooled organs	mRNA-Seq	n.a.	51 nt	0.91 Gbp	n.a.	n.a.
Venom gland	Small RNA	n.a.	50 nt	1.04 Gbp	n.a.	n.a.
Accessory gland	Small RNA	n.a.	36 nt	0.80 Gbp	n.a.	n.a.
Pooled organs	Small RNA	n.a.	50 nt	1.17 Gbp	n.a.	n.a.

[1] Actual insert sizes were first determined by alignment of reads against an initial *de novo* assembly. For PE200 and PE500, 99% of aligned pairs had an insert size in this interval. Mate pair insert size distributions are based on inspection of a histogram.

[2] Clean sequences are those filtered for adapter sequences and low quality nucleotides, and subsequently preassembled. Scaffolding links are pairs of which both reads align to different initial contigs at unique positions.

n.a., not applicable.

Table S2. The location of expression of non-toxin gene homologues related to venom toxin genes.

Toxin type	Location of non-toxin gene homolog expression		
	Previously inferred (57)	This study	
		Burmese python	Green anole
3FTx	-	Co-expressed (testes, ovaries, rictal gland)	Brain
CRISP	'Exocrine tissues'	-	-
cystatin	-	Co-expressed (stomach, ovaries, testes, rictal gland)	Co-expressed (brain, ovaries, testes, pooled organs)
hyaluronidase	-	Co-expressed (stomach, ovaries, testes, rictal gland)	-
kallikrein	'Exocrine tissues'	Heart and liver	Ovaries and pooled organs
LAAO	'Exocrine and immune tissues'	-	Testes
lectin	'Variety of tissues'	Spleen	Ovaries and pooled organs
NGF	'Variety of tissues'	Liver	-
PLA_2	Pancreas	Testes*	-
SVMP	'Variety of tissues'	Co-expressed (blood, heart, rictal gland, spleen)	-

The locations of expression of related non-toxin genes identified by phylogenetic analyses are presented in Figs. S11, S12 and S15-S22. The Burmese python (*P. molurus bivittatus*) transcriptomes analysed (9, 25) were: liver, spleen, heart, blood, muscle, ovary, rictal gland, stomach and testis. The green anole (*A. carolinensis*) transcriptomes analysed (26, 27) were: brain, dewlap, ovary, testis, and pooled organs (tongue, liver, gall bladder, spleen, heart, kidney, lung). *A sequence isolated from the pancreas of *Laticauda semifasciata* (GenBank: 21734661) was found nested within the venom PLA_2 radiation.

Table S3. LC-MS/MS identification of king cobra (*Ophiophagus hannah*) venom proteins excised and trypsin-digested from a 2D SDS-PAGE gel.

Transcriptome contig	Protein family	Peptide sequence	Charge	m/z	Sequest		Mascot	
					Prob	X-Corr	Ion Score	Exp value
Oh-675	LAAO	MSANNPENFGYQLNPNER	2	1047.97021	95.75	4.10	85	1.18E-07
		SASQLFDETLDKVTDDCTLQK	3	805.38495	124.24	4.64	67	7.02346E-06
		SASQLFDETLDK	2	677.33087	35.43	2.89	62	4.01283E-05
		MSANNPENFGYQLNPNER (+1)	2	1055.96765	34.84	4.46	62	2.6085E-05
		EDGWYVDVGPMR	2	712.31909	53.00	2.96	61	4.74909E-05
		QTDENAWYLIK	2	690.84357	38.07	3.07		
Oh-2904	CRISP	NMLQMEWNSNAAQNAK	2	925.41980	75.65	4.49	108	7.87606E-10
		NMLQMEWNSNAAQNAK (+2)	2	941.41669	92.22	5.09	103	2.50295E-09
		NMLQMEWNSNAAQNAK (+1)	2	933.41876	54.16	4.95	85	1.35827E-07
		YLYVCQYCPAGNIR	2	888.91400	81.60	4.32	81	3.80702E-07
		YKDDFSNCQSLAK	2	788.36011	41.59	3.98	79	6.62167E-07
		VIQSWYDENKK	2	705.35663	43.08	3.03	64	2.57274E-05
		FVYGVGANPPGSVIGHYTQIVWYK	3	884.79279	69.09	4.19	60	2.88107E-05
		CSFAHSPPHLR	2	654.81958	62.01	3.31	55	0.000208957
		FSCGENLFMSSQPYAWSR (+1)	2	1091.97131			69	4.58677E-06
		VIQSWYDENK	2	641.30811			60	6.99477E-05
		SGPPCGDCPSACVNGLCTNPCK	3	803.65247	61.72	4.87		
		FSCGENLFMSSQPYAWSR	2	1083.97449	67.37	4.78		
		SKCPASCFCR	3	424.84811	24.16	3.97		
		CPASCFCR	2	529.20508	37.02	3.04		

Oh-2022	NGF	ALTMEGNQASWR	2	682.32452	71.90	3.13	85	1.99207E-07
		IDTACVCVISR	2	647.31866			62	4.09153E-05
Oh-142	α-neurotoxin (3FTx)	ANPGVDIICCSTDNCNPFPTR	2	1204.52576	56.75	5.99	81	2.52667E-07
		VNLGCAATCPK	2	595.78632	52.25	2.98	80	7.85306E-07
		RVNLGCAATCPK	2	673.83722	19.75	3.97	58	0.000121174
Oh-1007	3FTx	ITCSAEETFCYK	2	754.82373	85.20	3.53	84	2.59665E-07
		ISNDRWYGCAK	2	685.32019	52.95	3.57		
Oh-630	PLA_2	YSYDCSEGTLTCK	2	792.32202	102.33	3.46	96	1.56776E-08
		RYSYDCSEGTLTCK	2	870.37244	79.77	4.93	69	6.5624E-06
		ADNDECAAFVCDCDR	2	909.33044			97	9.59511E-09
		CCQVHDNCYTQAQQLTECSPYSK	3	959.72906			78	4.39212E-07
		VAAICFAR	2	454.24448			57	0.000196214
		CCQVHDNCYTQAQQLTECSPYSKR	4	759.07373	141.69	5.67		

The 2D SDS-PAGE gel that was used for in gel trypsin-digestion is displayed in Fig. S8. Identification of peptides was undertaken using the Mascot and Sequest search algorithms. +1 and +2 represent the number of oxidized methionine residues. Transcriptome contigs refer to File S2.

Table S4. Identification of king cobra venom proteins exhibiting apparent molecular masses compatible with the isotope-average masses calculated for putative C-type lectin toxins.

Fraction	Molecular mass (SDS-PAGE)	Peptide ion		MS-derived sequence	Transcriptome contig	Relative abundance (%)	Protein family
Kc-		m/z	z				
15	9 kDa	534.2	2	IE(Mox)GCGCPK	Oh_4431	1.1	3FTx weak toxin DE-1 homolog-1
		724.8	2	TEPYTNLYCCK			
	9 kDa	459.7	2	TCPIGQDK			CTX15_OPHHA [Q53B46]
		799.3	2	SSADVEVLCCDTNK			
	9 kDa	1102.6	2	CLNTPLPLIYTTCPIGQDK			CTX27_OPHHA [Q69CK0]
		799.3	2	SSADVEVLCCDTNK			
16	9 kDa	760.9	2	GFYFSKPAGYGGNR	Oh_2851	0.8	Insulin-like growth factor
		1871.9	1	GIVEECCFQSCDLVR			
		2432.2	1	AGHETLCGAELVDALQFVCGER			
18	12 kDa	643.4	2	VVEAQSQVVAGAK	Oh_3322	1.3	cystatin

21	14 kDa	428.7	2	GVPDSPER	Oh_1836	0.7	Ohanin (vespryn)
		1071.5	2	ADVTFDSNTAFESLVVSPDK			
		757.4	3	ADVTFDSNTAFESLVVSPDKK			
		1497.7	1	FSSSPCVLGSPGFR			
23	23 kDa	435.8	2	ERILAIR	Oh_118903	0.4	Transferase-like protein *
24	23 kDa	597.3	2	QIVDKHNALR	Oh_2904	0.4	CRISP
		705.4	2	VIQSWYDENKK			
		2166.9	1	FSCGENLFMSSQPYAWSR			
		1776.8	1	YLYVCQYCPAGNIR			
		1849.8	1	NMLQMEWNSNAAQNAK			
		1308.6	1	CSFAHSPPHLR			

Identification was undertaken using a combination of CID-MS/MS and PMF. Cysteine residues are carbamidomethylated. HPLC fractions are displayed in Fig. S9. Transcriptome contigs refer to File S2. Relative abundance reflects the estimated percentage of total venom proteins.

* Transferase-like protein, Protein-cysteine N-palmitoyltransferase protein HHAT-like.

Table S5. Test statistics for directional selection analyses undertaken on venom toxin gene families.

Gene family	N sequences	Bayesian tree resolution	Mean posterior of nodes	Stdev posterior of nodes	N transitions	dN/dS non-venomous lineages	dN/dS venomous lineages
3FTx	122	0.742	0.953	0.112	7	0.312	**1.548**
CRISP	109	0.913	0.967	0.098	4	0.339	0.991
cystatin	53	0.864	0.969	0.087	3	0.223	0.585
hyaluronidase	45	0.946	0.979	0.080	4	0.164	0.393
kallikrein	83	0.855	0.957	0.110	5	0.253	0.907
LAAO	47	0.854	0.973	0.084	4	0.205	0.825
NGF	60	0.893	0.954	0.115	7	0.160	0.696
PLA_2	158	0.869	0.947	0.127	21	0.204	**1.268**
SVMP	78	0.589	0.939	0.125	10	0.105	**1.236**

Gene family loci analysed for evidence of directional selection. Bayesian tree resolution shows how well-resolved the Bayesian majority rule consensus gene trees were (1.0 is fully bifurcating), and the subsequent two columns indicate how well supported the nodes on these trees were. N transitions shows the number of transitions between the venomous and non-venomous state as reconstructed on fully resolved maximum likelihood gene trees obtained with the Bayesian consensus trees as starting point. The final two columns show the dN/dS ratio as constructed on non-venomous and venomous lineages, respectively. Values > 1 in the last column (highlighted in bold and underlined) indicate directional selection acting on a locus subsequent to its recruitment in a venom-producing pathway.

Table S6. Estimated models of sequence evolution for DNA toxin family phylogenetics as determined by MrModelTest (33).

Dataset	Codon Position	Model
3FTx	1	HKY + I + Γ
	2	GTR + I + Γ
	3	GTR + Γ
CRISP	1	GTR + I + Γ
	2	GTR + I + Γ
	3	GTR + I + Γ
cystatin	1	HKY + I + Γ
	2	GTR + Γ
	3	HKY + Γ
hyaluronidase	1	GTR + I + Γ
	2	GTR + I + Γ
	3	GTR + I + Γ
kallikrein	1	GTR + I + Γ
	2	GTR + I + Γ
	3	HKY + I + Γ
LAAO	1	GTR + I + Γ
	2	GTR + Γ
	3	HKY + Γ
lectin	1	GTR + I + Γ
	2	GTR + Γ
	3	GTR + I + Γ
NGF	1	GTR + I + Γ
	2	GTR + I + Γ
	3	GTR + I + Γ
PLA_2	1	GTR + I + Γ
	2	GTR + Γ
	3	GTR + Γ
SVMP	1	GTR + I + Γ
	2	GTR + I + Γ
	3	GTR + I + Γ

5. SI REFERENCES

1. Li R, *et al*. (2010) The sequence and *de novo* assembly of the giant panda genome. *Nature* 463:311–317.

2. Li R, *et al*. (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20:265–272.

3. White paper on *de novo* assembly in CLC Assembly Cell 3.0, March 16, 2010 (http://www.clcbio.com/files/whitepapers/white_paper_on_de_novo_assembly_on_the_CLC_Assembly_Cell.pdf).

4. Boetzer M, *et al.* (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27:578–579.

5. Langmead B, *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

6. Cantarel BL, *et al.* (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18:188-196.

7. Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491-505.

8. Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* 13:329-342.

9. Castoe TA, *et al.* The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Jointly submitted to *Proc. Natl. Acad. Sci. USA* with this submission.

10. Feschotte C, *et al.* (2009) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol. Evol.* 1:205-220.

11. O'Donovan C, *et al.* (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform.* 3:275-284.

12. UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. *Nucleic Acids Res.* 38:D142-D148.

13. Pruitt KD, *et al.* (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* 37:D32-D36.

14. Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:59-68.

15. Stanke M, *et al.* (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7:62.

16. Grabherr MG, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29(7):644-652.

17. Birol I, *et al.* (2009) De novo transcriptome assembly with ABySS. *Bioinformatics (Oxford, England)* 25(21):2872-2877.

18. Simpson JT, *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19(6):1117-1123 (2009).

19. Zhao QY, *et al.* (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* 12(Suppl 14):S2.

20. Karim S, Singh P, Ribeiro JM (2011) A Deep Insight into the Sialotranscriptome of the Gulf Coast Tick, *Amblyomma maculatum*. *PLoS ONE* 6(12):e28525.

21. Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6:175-182.

22. Nielsen H, Brunak S, von Heijne G (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Engineering* 12(1):3-9.

23. Duckert P, Brunak S, Blom N (2004) Prediction of proprotein convertase cleavage sites. *Protein Eng. Des. Sel.* 17(1):107-112.

24. Julenius K, *et al.* (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology* 15(2):153-164.

25. Castoe TA, *et al.* (2011) Sequencing the genome of the Burmese python (*Python molurus bivittatus*) as a model for studying extreme adaptations in snakes. *Genome Biol.* 12:406.

26. Alföldi J, *et al.* (2011) The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* 477:587-591.

27. Eckalbar WL, *et al.* (2013) Genome reannotation of the lizard *Anolis carolinensis* based on 14 adult and embryonic deep transcriptomes. *BMC Genomics* 14:49.

28. Casewell NR, Huttley GA, Wüster W (2012) Dynamic evolution of venom proteins in Squamate reptiles. *Nat. Commun.* 3:1066.

29. Tamura K, *et al.* (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol. Biol. Evol.* 28:2731-2739.

30. Edgar R (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792-1797.

31. Castoe TC, Parkinson CL (2006) Bayesian mixed models and the phylogeny of pitvipers (Viperidae: Serpentes). *Mol. Phylogenet. Evol.* 39:91-110.

32. Castoe TC, Sasa M, Parkinson CL (2005) Modelling nucleotide evolution at the mesoscale: the phylogeny of the Neotropical pit vipers of the Porthidium group (Viperidae: Crotalinae). *Mol. Phylogenet. Evol.* 37:881-898.

33. Nylander JAA (2004) MrModeltest v2. Program distributed by the author, Evolutionary Biology Centre, Uppsala University.

34. Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53:793-808.

35. Ronquist F, *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61:539-542.

36. Kumar S, *et al.* (2009) AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. *BMC Bioinformatics* 10:357.

37. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214.

38. Chaudhary R, *et al*. (2010) iGTP: a software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics* 11:574.

39. Tavaré S (1986) Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Am. Math. Soc. Lect. Math. Life Sci.* 17:57-86.

40. Rambaut A, Drummond A (2007) Tracer v1.4. Available from: http://beast.bio.ed.ac.uk/Tracer

41. Guindon S, Gascuel A (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52:696-704.

42. Yang Z, Nielsen R (2002) Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Mol. Biol. Evol.* 19:908-917.

43. Bollback JP (2006) SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* 7:88.

44. Nielsen R (2001) Mutations as missing data: inferences on the ages and distributions of nonsynonymous and synonymous mutations. *Genetics* 159:401-411.

45. Nielsen R (2002) Mapping mutations on phylogenies. *Syst. Biol.* 51:729-732.

46. Huelsenbeck JP, Nielsen R, Bollback JP (2003) Stochastic mapping of morphological characters. *Syst. Biol.* 52:131-158.

47. Couvreur TLP, *et al.* (2010) Insights into the influence of priors in posterior mapping of discrete morphological characters: a case study in Annonaceae. *PLoS ONE* 5(5):e10473.

48. Görg A, *et al.* (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* 21(6):1037-1053.

49. Hayter JR, *et al.* (2003) Proteome Analysis of Intact Proteins in Complex Mixtures. *Mol. Cell. Proteomics* 2:85-95.

50. Altschul S, *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

51. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136:215-233.

52. Wheeler BM, *et al.* (2009) The deep evolution of metazoan microRNAs. *Evol. Dev.* 11:50-68.

53. Griffiths-Jones S, *et al.* (2008) miRBase: tools for microRNA genomics. *Nucleic Acid Res.* 36:D154-158.

54. Darnell DK, *et al.* (2006) MicroRNA expression during chick embryo development. *Dev. Dynamics* 235:3156-3165.

55. Jostarndt K, *et al.* (1994) The use of 33P-labelled riboprobes for in situ hybridizations: localization of myosin alkali light-chain mRNAs in adult human skeletal muscle. *Histochem. J.* 26:32-40.

56. Fry BG, *et al.* (2006) Early evolution of the venom system in lizard and snakes. *Nature* 439:584-588.

57. Fry BG (2005) From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. *Genome Res.* 15:403-420.