Snake evolution and prospecting of snake venom

Vonk, F.J.

Cover Page

Universiteit Leiden

Author: Vonk, Freek Jacobus
Title: Snake evolution and prospecting of snake venom
Date: 2012-09-06

# Chapter 5: Massive Evolutionary Expansion of Venom Genes in the King Cobra Genome

Vonk FJ, Henkel CV, Casewell NV, Kini RM, Kerkkamp HM, Wuster W, Castoe TA, Ribeiro JMC, Spaink HP, Jansen HJ, Hyder SA, Arntzen JW, Pollock DD, van den Thillart GEEJ, Boetzer M, Pirovano W, Dirks RP and MK Richardson.

Snake venom is a complex mixture of proteins and peptides evolved to immobilize prey and deter predators[166]. The rapid evolution of venom toxins is part of a predator-prey 'arms race' that represents a classic model for studying molecular evolution[13,25,27,167,168]. Snake toxins are thought to evolve from normal physiological proteins through gene duplication and recruitment to the venom gland[169-172]. However, in the absence of genomic resources, these hypotheses remain speculative. Using Illumina sequencing technology we have produced a draft genome of an adult male Indonesian king cobra (*Ophiophagus hannah*), and have deep-sequenced its venom gland transcriptome. Comparative genomics revealed evidence of tandem duplication of genes encoding physiological L-amino acid oxidase, cysteine-rich secretory proteins and metalloproteinases, followed by recruitment through selective expression in the venom gland. By contrast, nerve growth factor toxins appear to have evolved by duplication and dual recruitment, while hyaluronidase and phospholipase B evolved by recruitment of existing physiological genes without further duplication, similar to acetylcholinesterase[173]. We also identify 21 different three-finger toxin (3FTX) genes in the genome, suggesting a massive expansion of this family. We find a significant variation in the expression levels of these 3FTX genes in the venom. These data show that venom proteins originate and evolve through multiple distinct mechanisms. These sequences provide a valuable resource for studying rapid evolution of gene sequences and the evolution of recruitment of genes to different tissues.

Many advanced snakes (Caenophidia[174]) use venom, with or without constriction, to subdue their prey. Venom is produced in a post-orbital venom gland that probably evolved from an ancestral gland in the posterior part of the mouth[2]. There is evidence that the evolutionary history of venom in reptiles may be traced as far back as Triassic lizards[175]. The pharmacologically-active proteins and peptides in venom target a wide variety of proteins including receptors, ion channels and enzymes[176]. Their actions include disruption of the central and peripheral nervous systems, the blood coagulation cascade, the cardiovascular and neuromuscular systems, and homeostasis. Only in recent years has the remarkable variability of venom composition at the taxonomic, population, and individual levels been fully appreciated, and a start made to investigate the underlying ecological drivers and molecular mechanisms[176].

It is thought that toxin gene families have evolved through duplication of normal physiological genes[170-172,177], followed by recruitment and expression in the venom gland. However, this hypothesis has not been verified with genomic data. Therefore, we have produced a draft genome of an adult male Indonesian king cobra (*Ophiophagus hannah*) and deep-sequenced the transcriptome of its venom gland using Illumina technology. The sequence data were first assembled *de novo* into contigs, which were subsequently oriented and merged in scaffolds (see Methods summary). Haploid genome size was estimated using flow cytometry to be around 1.36-1.59 Gbp. Our assembled draft has an N50 contig size of 6533 bp, and an N50 scaffold size of 242 Kbp. The contigs sum to 1.34 Gbp, and the scaffolds (which contain gaps) to 1.59 Gbp.
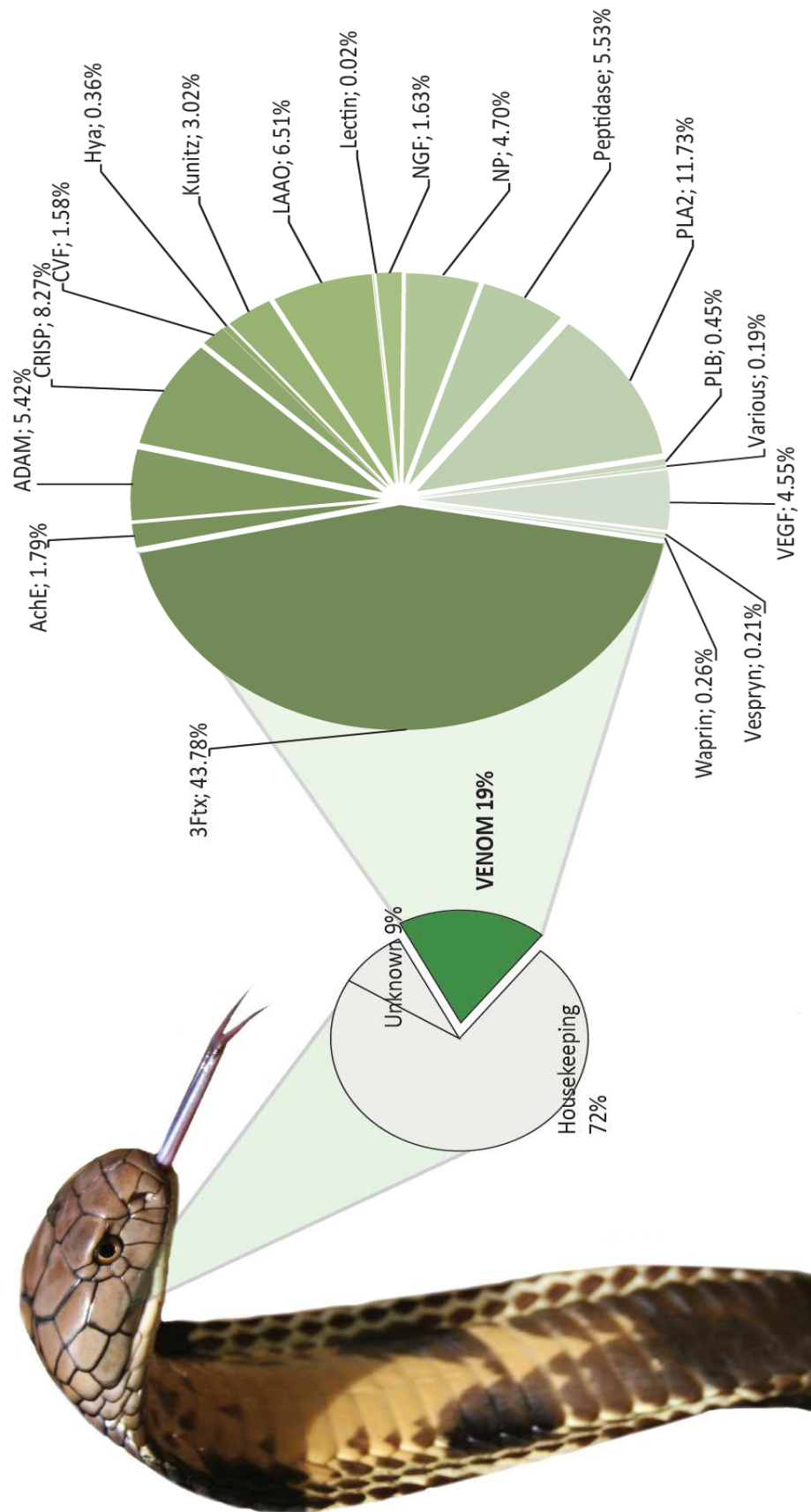
Mitochondrial genome phylogeny confirms that the male specimen we used for genome sequencing clusters in the *Ophiophagus* group with other king cobras (Fig. 18). Using Augustus gene prediction[178], and our transcriptome data we estimate that the king cobra has approximately 22,183 protein-coding genes (data not shown). Although some of the predicted genes will be either part of a gene that spans multiple scaffolds, or will represent mispredictions, the values suggest that the total number of genes in snakes and other amniotes is similar[179-181].
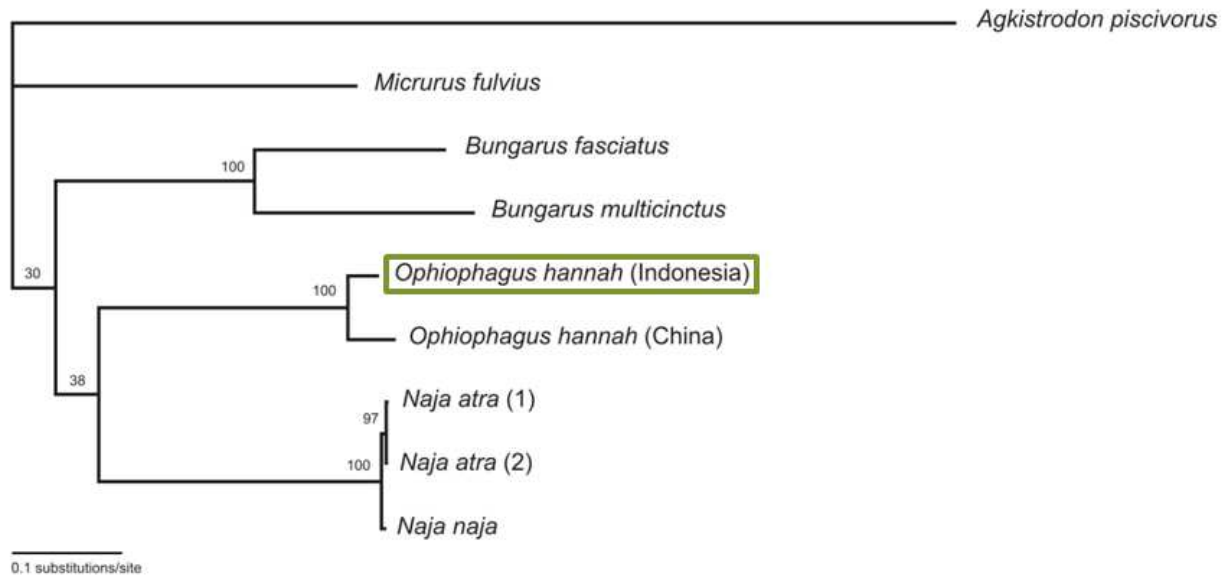
Figure 18 | Mitochondrial genome phylogeny showing that our male Indonesian king cobra (*Ophiophgus hannah*) groups with the king cobra from China (Yi Chuan. 2010 Jul;32(7):719-25). Note that the *Naja naja* is in fact a taxonomy error because the sample was taken from a *Naja atra* (Wolfgang Wuster, pers comm).

We identified 17 different toxin families in the venom gland transcriptome by blasting against reference sequences (from www.ncbi.nlm.nih.gov, and see Fig. 19) and annotated nine of them in the genome (Figs. 20-25). These nine include: three-finger toxins (3FTXs), L-amino acid oxidase (LAAO), phospholipase A_2 (PLA$_2$), phospholipase-B (PLB), cysteine-rich secretory protein (CRISP), metalloproteinases (ADAM), nerve growth factor (NGF), hyaluronidase (HYA), cobra venom factor (CVF). Three of these (NGF, PLB and CVF) have not previously been reported in king cobra venom.

Figure 19 | (page 64) Relative abundance of the venom toxins in the venom gland transcriptome. The percentages are calculated based on the expression value of the transcripts sequenced from the venom gland transcriptome. The most abundant family is the three finger toxins (31,54% of all toxin transcripts identified), represented in the genome by at least 21 different isoforms (see also Fig. 20).

Hya; 0.36%

Kunitz; 3.02%

LAAO; 6.51%

Lectin; 0.02%

NGF; 1.63%

NP; 4.70%

Peptidase; 5.53%

CVF; 1.58%

CRISP; 8.27%

PLA2; 11.73%

ADAM; 5.42%

PLB; 0.45%

Various; 0.19%

AchE; 1.79%

VEGF; 4.55%

Waprin; 0.26%

Vespryn; 0.21%

3Ftx; 43.78%

VENOM 19%

Unknown 9%

Housekeeping 72%

Proteins in two of these families (3FTX and PLA2), are known to exhibit a wide variety of toxic and pharmacological effects including neurotoxicity, cardiotoxicity and hemolysis[182,183]. We find evidence for massive expansion in the genome in both these families. We found seven different exons-2 that belong to PLA2 (Fig. 20). These genomic sequences do not contain premature stop codons or frameshifts indicating that they do not contain pseudogenes.

Figure 20 | Massive expansion of PLA2 genes. Alignment of multiple PLA2 genomic hits. Note no stopcodons or frameshifts in the sequences.

3FTXs are three-exon genes, of which the second exon is most readily identified. We found 21 of these exons-2 in the genome (Fig. 21). However, some of these are on small contigs and covered by relatively many sequencing reads, indicative of high copy numbers. Therefore, the actual diversity of full-length 3FTX genes may be even higher. Most exons-2 are expressed in the venom gland, although the
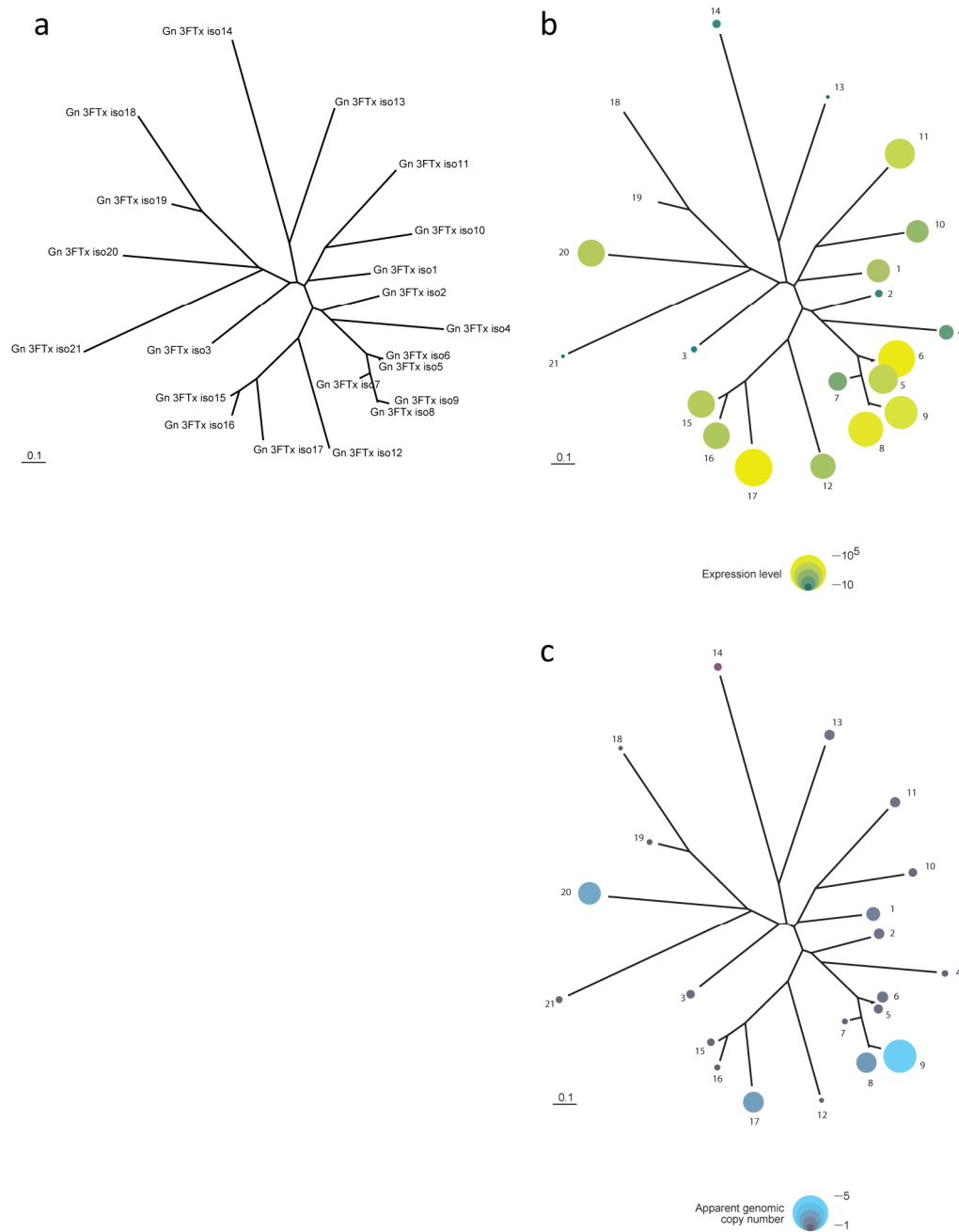
Figure 21 | Unrooted phylogenetic tree constructed from all different exon 2 sequences of the three finger toxin genes. Isoform 19 contains a premature stop codon, thus most likely is a pseudogene. Green circles indicate relative expression levels (on a logarithmic scale), blue circles apparent genomic copy numbers, both based on local coverage by venom.

expression levels differ by five orders of magnitude (Fig. 21). One non-expressed isoform (isoform 19) contains a premature stop codon and may be part of a pseudogene (Fig. 22). The presence of multi-copy and highly expressed exons is clustered in several 'successful' branches of the 3FTX gene family, and genomic copy number and expression level in the venom gland appear to be correlated (Fig. 20).

```
                                      20
                                      |
Gn_3FTx iso1   GYTLTCLTHE  SLFFETTETC  SDGQNLCYAK  -WFAVFPG 37
Gn_3FTx iso2   GYTRIC--HK  SSFI--SETC  PDGQNLCYLK  SWCDIF-- 32
Gn_3FTx iso3   GYTLTCITSA  RNF---ETC   PPGQNLCFLK  SWYEA--S 32
Gn_3FTx iso4   -----XXXYK  TGERIISETC  PPGQDLCYMK  TWCDVF-- 31
Gn_3FTx iso5   GYTTKCYVTP  DA---TSQTC  PDGENICYTK  SWCDGF-- 33
Gn_3FTx iso6   GYTTKCYVTP  DA---TSQTC  PDGENICYTK  SWCDVF-- 33
Gn_3FTx iso7   GYTTKCYITP  DV---KSQTC  PDGENICYTK  TWCDVW-- 33
Gn_3FTx iso8   GYTTKCYVTP  DV---KSETC  PDGQDICYTE  TWCDVW-- 33
Gn_3FTx iso9   GYTTKCYVTP  DV---KSETC  PAGQDICYTE  TWCDAW-- 33
Gn_3FTx iso10  GHTRICLTDY  SKVSETIEIC  PDGQNFCF-K  KFPKGIPF 37
Gn_3FTx iso11  GYTMKCLTKY  SRVSETSQTC  HVWQNLCFKK  -----WQK 33
Gn_3FTx iso12  GYTTKCYNHQ  STTPETTEIC  PDSGYFCYKS  SWIDG--R 36
Gn_3FTx iso13  GYTLICHRVH  GL----QTC   EPDEKFCFRK  TTM-FFPN 32
Gn_3FTx iso14  GYTRKCLNTP  ----------  ---LPLIYXK  MTIKKLPS 25
Gn_3FTx iso15  XYTRICLKQE  PFQPETSTTC  PDGEDACYST  FWSDN--- 35
Gn_3FTx iso16  XYTRICLKQE  PFQPETTTTC  PEGEDACYNL  FWSDH--- 35
Gn_3FTx iso17  GYSLICFNQE  TYRPETTTTC  PDGEDTCYST  FWNDHH-- 36
Gn_3FTx iso18  AQTKTCYSCT  GAFCSNRQKC  SGGQVICF-K  SWKNTLLI 37
Gn_3FTx iso19  AHTLTCYSCN  GLLCSDREQC  PDG*DICF-K  RWNDTDWS 37
Gn_3FTx iso20  GYSLTCLNCP  EQYCKRIHTC  RDGENVCF-K  RFYEGKLL 37
Gn_3FTx iso21  GYTLLCCKCN  QTVCDLNSYC  SAGKNQCYIL  Q-----NN 33
```

Figure 22 | Alignment of multiple 3FTX isoforms. Note that only isoform19 (Gn_3FTx iso19) contains a stopcodon and may thus be a pseudogene.

There is a substantial difference in expression levels of each of the 3FTX isoforms (Fig. 20). Isoform diversity and toxin expression levels are thought to be important in optimization of the prey-specificity of the venom — more so than differences in the representation of entire toxin families and the recruitment of novel toxin families[184]. In general, a high genomic copy number is associated with a high relative expression value (Fig. 20). All relatively 'successful' branches of 3FTX genes (the ones that are expressed heavily) share sequence similarities (Fig. 22), indicating conservation of important functions.

Reptile venom CRISPs act as regulators of several types of ion channels[185]. We find three CRISP genes in tandem in the king cobra genome (Fig. 23), only two which are represented in our venom gland transcriptome (Fig. 23). Together with the comparative genomic data (Fig. 23), this is consistent with an evolutionary scenario in which the two venom genes have been derived by tandem duplication from the non-venom expressed (physiological) CRISP gene.
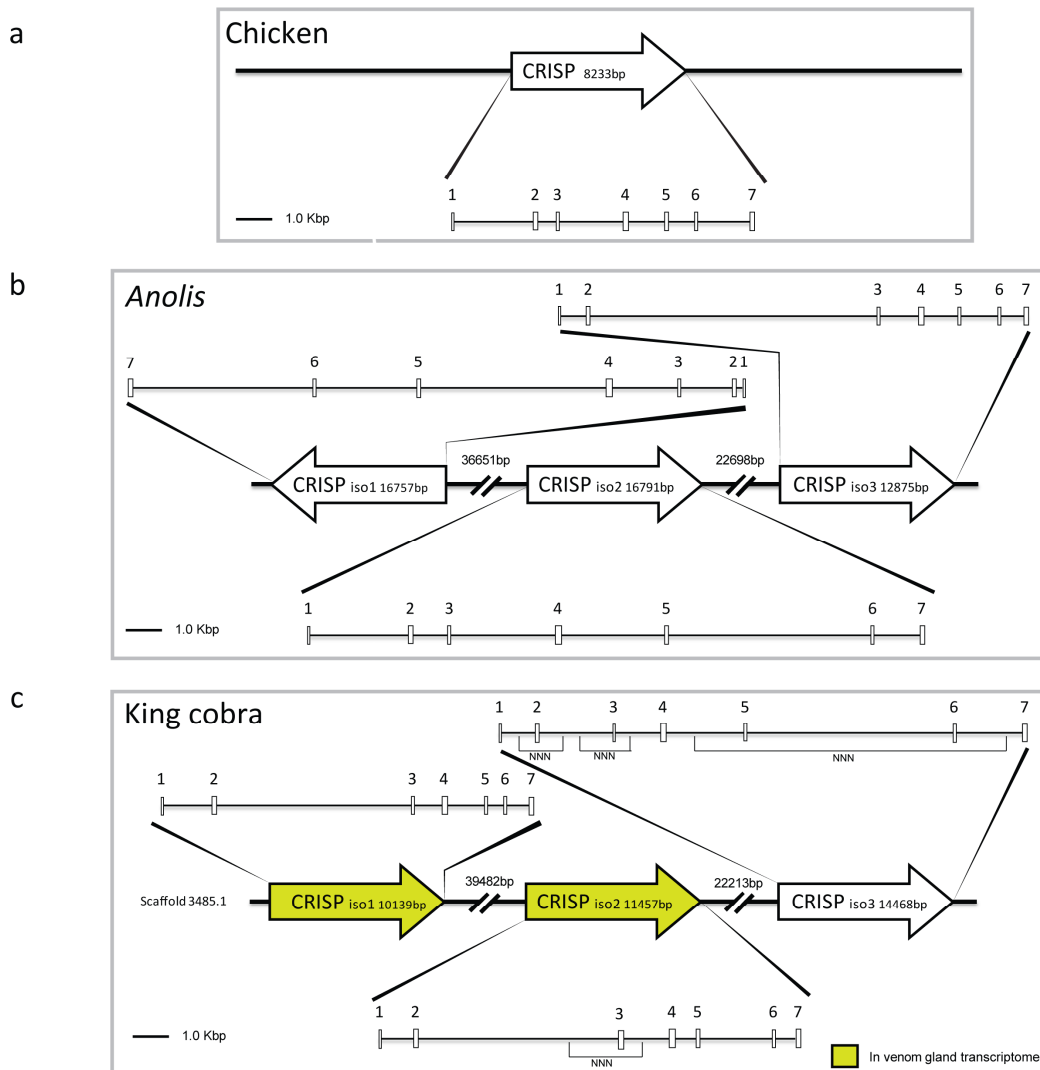
Figure 23 | Comparative genomic architecture of the CRISP genes. a, chicken (*Gallus gallus*); b, anole lizard (*Anolis carolinensis*); and c, king cobra (*Ophiophagus hannah*). Chick and Anolis sequences are from www.ensembl.org. The exploded views show scale diagrams of the exons and introns. Scale bar refers to the exploded views. NNN, unresolved sequence. In the

Anolis genome we annotated three CRISP genes with different orientations. Based on the relative sizes of the second introns the two 'venom' CRISP genes are comparable to isoform 3 in Anolis. In chicken we could only find one CRISP gene.
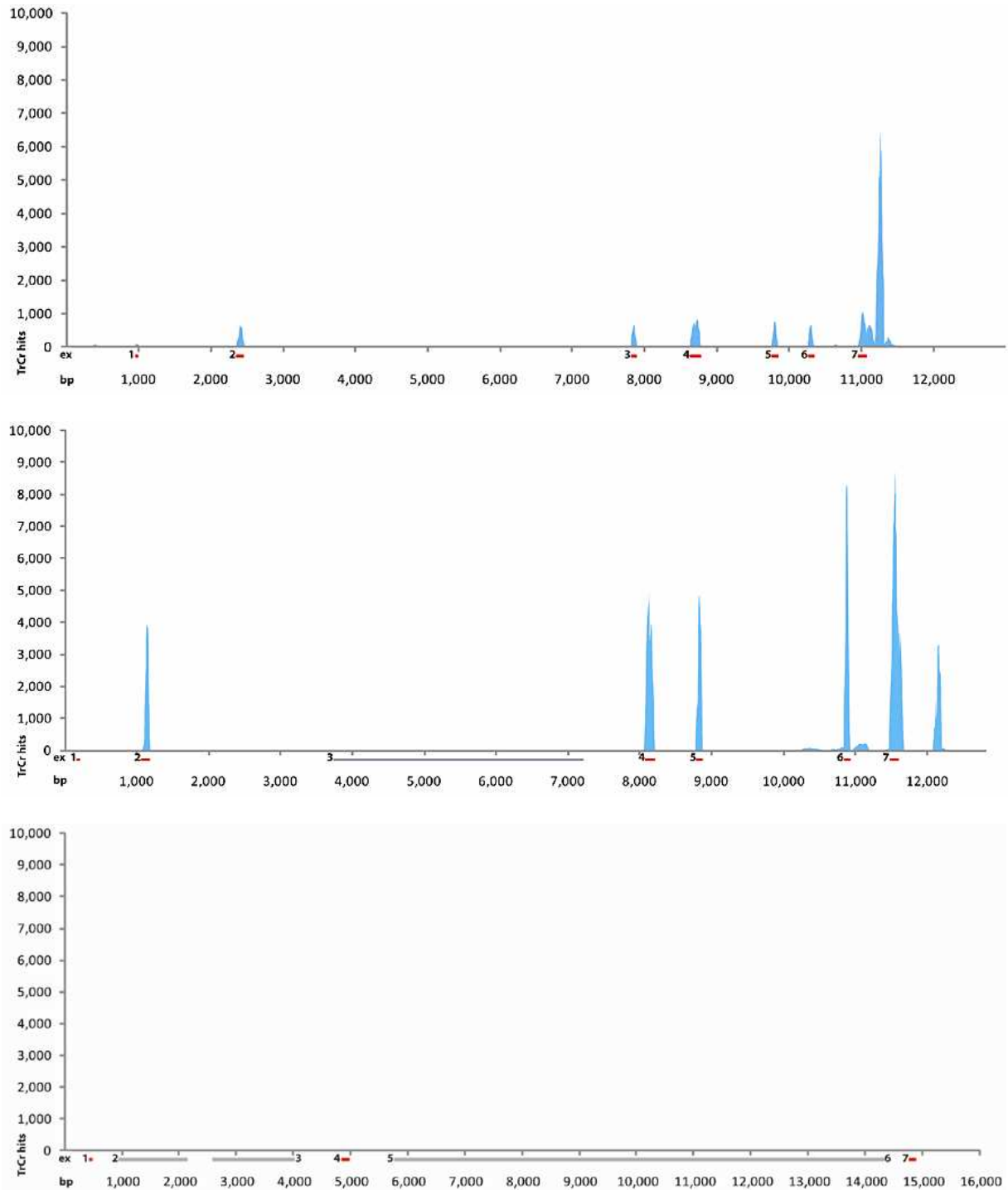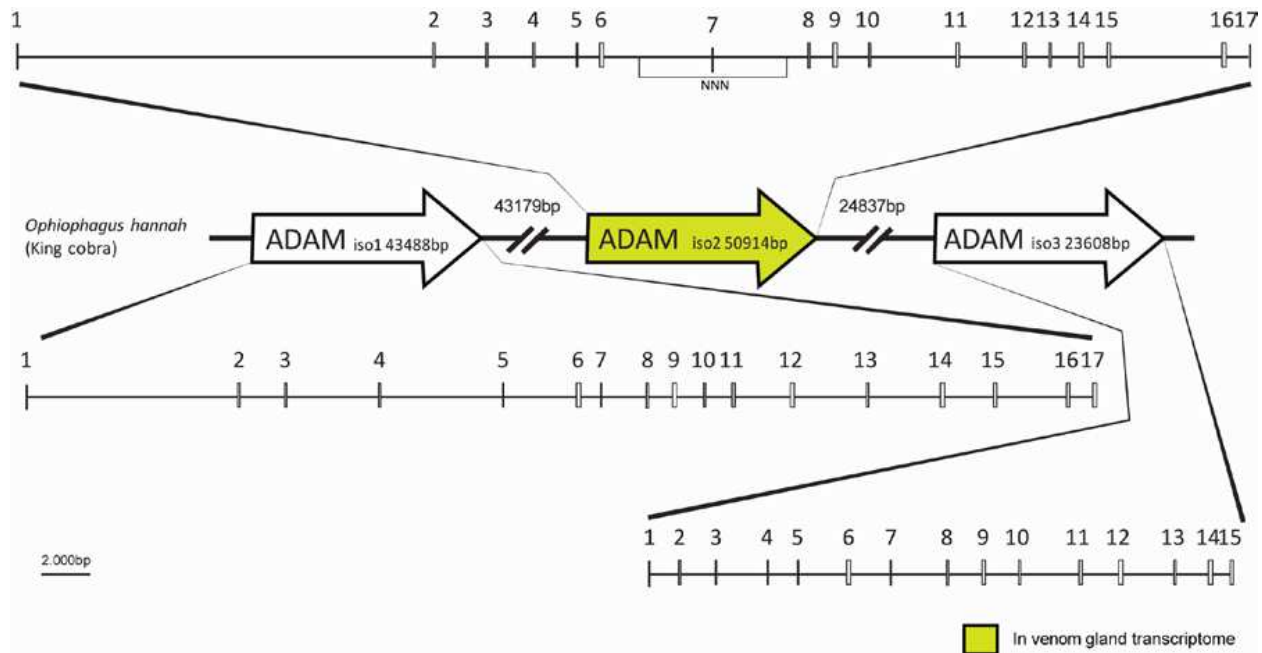

Figure 24 | The scaffold containing three CRISP genes with different isoform transcripts (see Figure 23c) mapped on as follows: a) isoform 1; b) isoform 2; c) isoform 3. As can be seen, only the first two isoforms are expressed in the venom gland; d)

alignment of the three CRISP genes with reference sequences showing that our identified genes belong to the CRISP family. Isoform1 is opharin and isoform 2 is ophanin.

Venom metalloproteinases belong to the ADAM family and target various stages of blood coagulation and platelet aggregation and are responsible for hemorrhage[186]. We also find three ADAM genes in tandem, only one of which was expressed in the venom gland transcriptome (Fig. 25).

Figure 25 | Scaffold containing three ADAM genes; b) isoform 1; c) isoform 2; d) isoform 3. Only isoform 2 is expressed in the venom gland (data not shown). Amino acid alignments of these three metalloproteinase genes with the single transcriptome sequence (not shown) shows that one gene is identical and confirms its expression. Isoform 1 has a longer C-terminal tail. In *O. hannah* isoform 2 is expressed in the venom gland, while in *Naja atra* (a different elapid snake) isoform 3 appears to be expressed, since we find that *N. atra* metalloproteinase sequence is more similar to isoform 3 than isoform 2

LAAO produces H_2O_2 during oxidation of amino acids leading to cytotoxicity and inhibition of platelet aggregation (and is responsible for the yellow color of the venoms)[187]. We find two LAAO genes on two different scaffolds (Fig. 26 a). Based on the mapping of venom gland transcriptome reads (not shown), only one LAAO gene appears to be expressed in the venom gland; the other is presumably the

70

non-venom, physiological gene. To the best of our knowledge, non-venom LAAO proteins have not been found in reptiles before, although they are found widely among vertebrates.
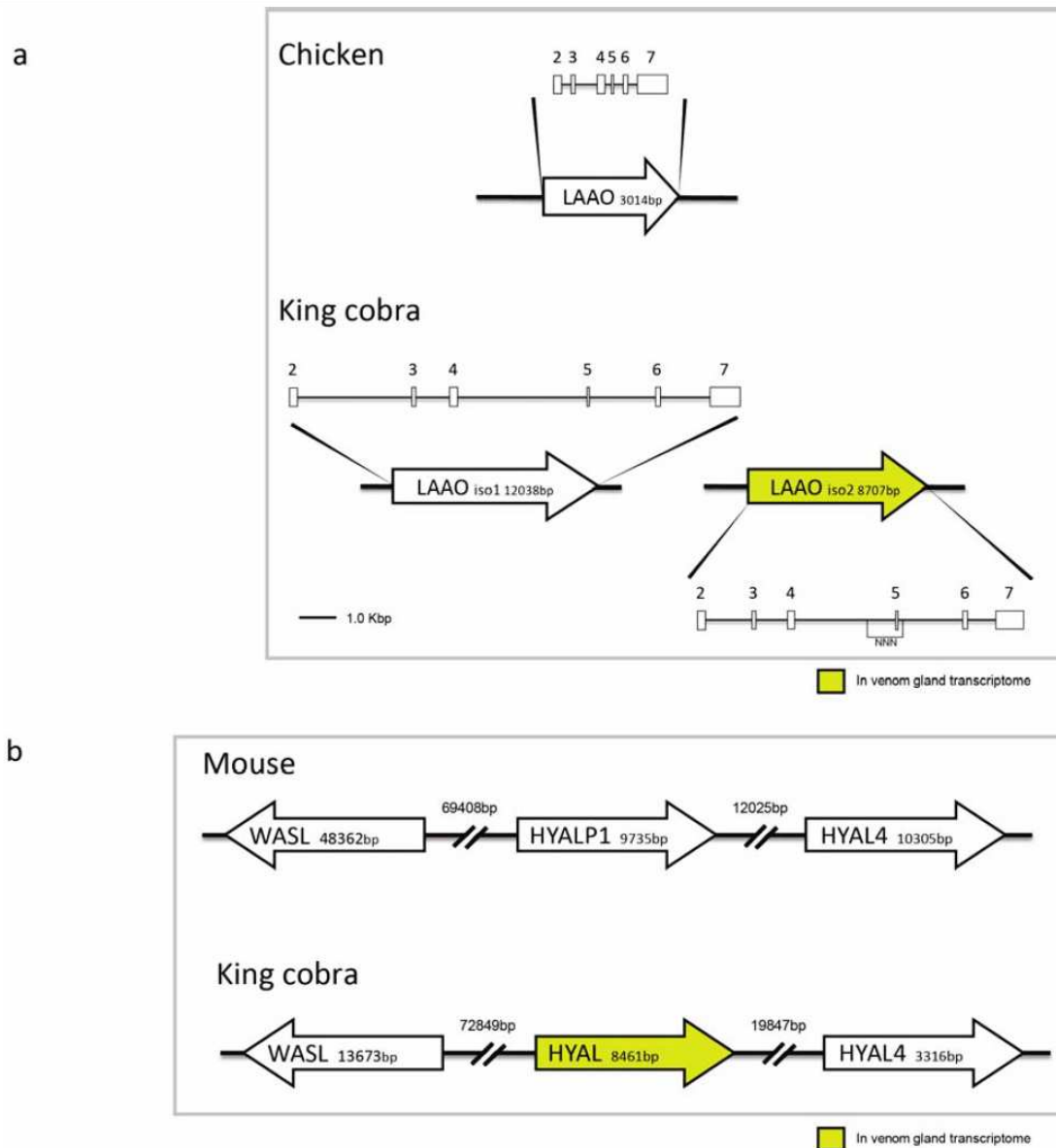

Figure 26 | a, Genomic architecture of l–amino acid oxidase (LAAO) genes in the chicken and king cobra. b, scheme of the genomic context of the hyaluronidase genes in the mouse (*Mus musculus*) and king cobra. Mouse genomic sequences from www.ensembl.org. Scale bar refers to the exploded views. NNN, unresolved sequence.

The role of venom NGF is not clear[98]. We find two different NGF genes, both of which are

encoded by a single exon; and both of them are expressed in the venom gland (Fig. 27). Presumably, one

or both of these has duplicate functions (in both venom-gland and in other tissues). Venom hyaluronidase

plays a key role as the venom 'spreading factor', making tissue more permeable[188]. We annotated two

hyaluronidase genes in the king cobra genome, both lie downstream of the WASL gene, and we find the

same arrangement in the mouse genome (Fig. 26 b). Only the gene corresponding to HYALP1 is

expressed in the venom gland (data not shown), which is interesting because in the mouse this gene

appears to be inactive[189]. This synteny is consistent with a scenario in which the duplication of the

hyaluronidase gene took place long before one of the copies was recruited to the venom gland.

a
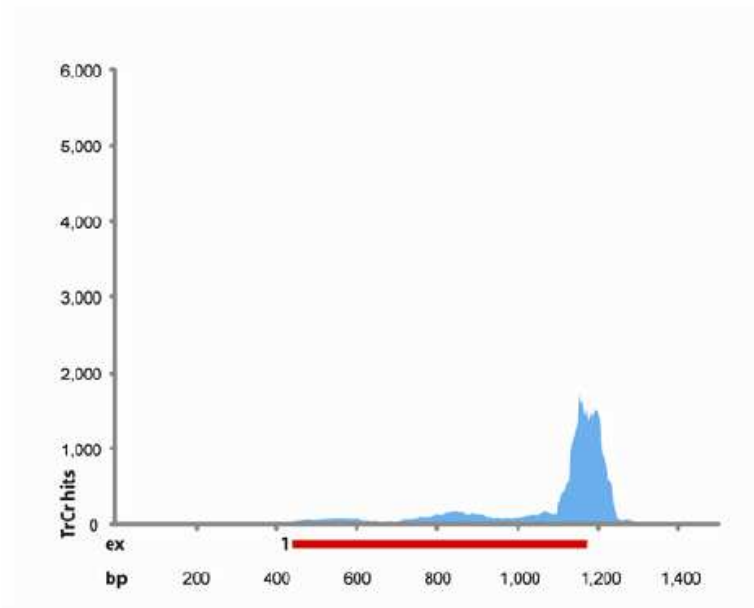
b
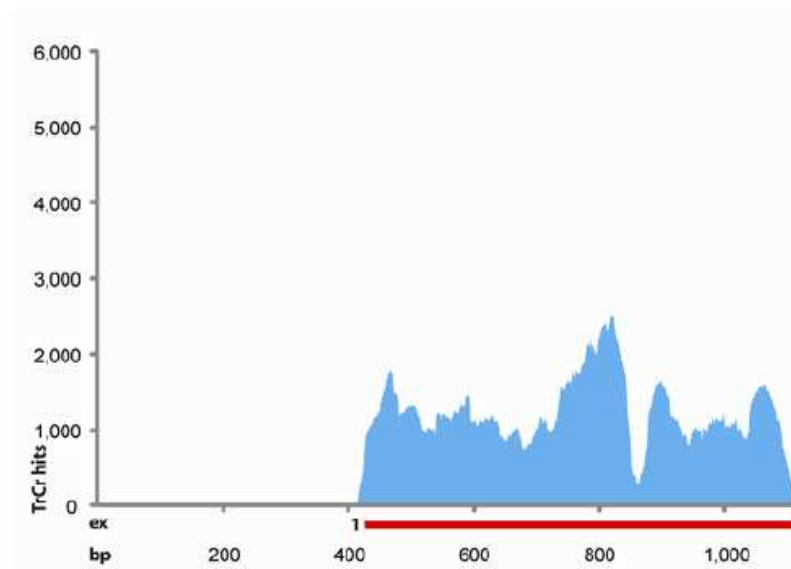
Figure 27 | Mapping of the transcriptome reads onto the two scaffolds containing two NGF genes shows that both of these genes are expressed in the venom gland; a) isoform 1; b) isoform 2; c) Alignment of the two NGF genes with reference sequences showing that our identified genes belong to the NGF gene family (data not shown).

Recently, PL-B was also found to be expressed in the venom gland[190] but its role in toxicity is yet unclear. We could only find one PL-B gene (Figure 28). This indicates that an existing PL-B gene was recruited to the venom gland. Thus HYA, NGF and PL-B genes appear to be recruited for expression in the venom gland without gene duplication being involved. In the case of the Asian krait (*Bungarus fasciatus*) Acetylcholinesterase toxin, it was shown that both the neuronal and the venom enzymes are encoded by the same gene, although alternatively spliced[173].
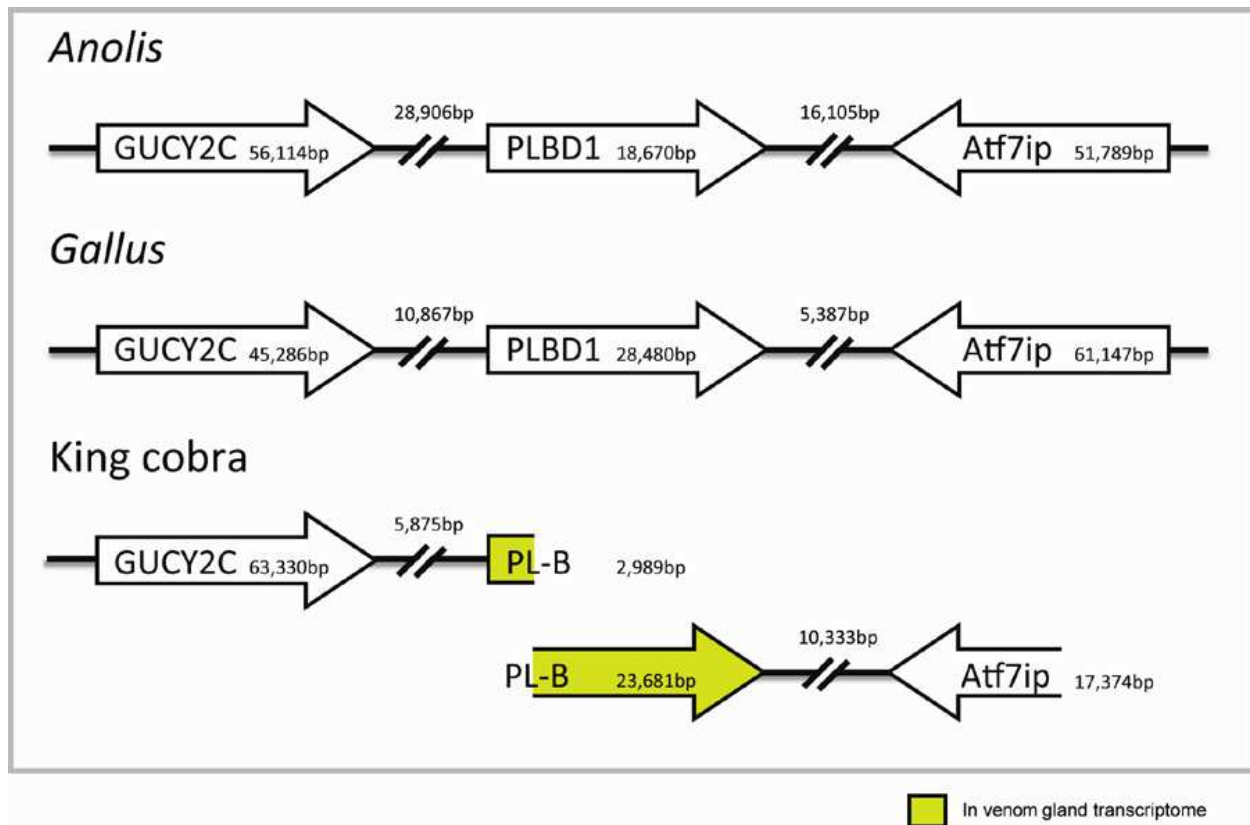
Figure 28 | Scheme of the genomic synteny of the PL-B genes in the Anolis, Gallus and king cobra. Anolis, Gallus genomic sequences from www.ensembl.org. Mapping of the transcriptome reads onto one scaffolds containing the PL-B gene shows that this gene is expressed in the venom gland (data not shown). This synteny is consistent with the scenario of recruitment of the existing PLBD1 into the venom gland during snake evolution. The alignment of the PL-B gene with reference sequences showing that our identified genes belong to the PL-B gene family (data not shown).

74

It has been shown, in the case of factor X toxin in the rough-scaled snake (*Tropidechis carinatus*), that a specific insertion in the promoter region of the toxin was responsible for the selective recruitment to the venom gland[169]. We have scanned all our scaffolds for this sequence but could not find anything similar. This suggests that that the specific insertion is not a universal feature of toxin gene recruitment, and that several distinct mechanisms are responsible for the origin and recruitment of venom proteins.

Conclusion and discussion

Comparative genomics has revealed flexible mechanisms of mutation and recruitment of venom genes. We found evidence of tandem duplication of genes encoding physiological L-amino acid oxidase, cysteine-rich secretory proteins and metalloproteinases, followed by recruitment through selective expression in the venom gland. By contrast, nerve growth factor toxins appear to have evolved by duplication and dual recruitment, while hyaluronidase and phospholipase B evolved by recruitment of existing physiological genes without further duplication. We also identify 21 different three-finger toxin (3FTX) genes in the genome, suggesting a massive expansion of this family. We find a significant variation in the expression levels of these 3FTX genes in the venom. Our data therefore shows that venom proteins originate and evolve through multiple distinct mechanisms.

The king cobra genome is an important resource for studying molecular evolution. The powerful combination of genomics with transcriptomics here used lead to the identification of toxin genes previously unknown in the king cobra. Because of the massive functional diversity known to exist not only between toxins but also their isoforms[176], functional studies of these new toxins could prove to be of great interest.

Methods summary

king cobra tissue acquisition and processing. All animal procedure complied with local ethical approval. Genome sequencing was done on a blood sample obtained from an adult male king Cobra from a captive specimen that originated from Bali, Indonesia. Blood was obtained by caudal puncture and frozen in liquid nitrogen. The venom gland and other tissue samples were dissected from a freshly euthanized second adult male specimen and stored in RNAlater.

Sequencing and assembly. We used a whole-genome shotgun sequencing strategy and Illumina Genome Analyser sequencing technology. Two paired-end and four mate pair libraries were constructed with insert sizes of up to 15K nucleotides. In total, we generated 41.2 Gbp (approximately 28x genome coverage) of sequence data for contig building, 21.1 Gbp for scaffolding, and 1.7 Gbp for the transcriptome. We built contigs from the short reads using the CLC bio *de novo* assembler (CLC bio, Aarhus, Denmark) and oriented these contigs using SSPACE. A more extensive methods section is included in the supplementary information.

Annotation and gene prediction. Gene prediction was carried out automatically using Augustus software[178], using venom gland transcripts as hints. Further extensive manual annotation was performed to establish the intron-exon boundaries.

Author contributions F.J.V., study concept and design, tissue preparation, flow cytometry, sequence analysis, writing of manuscript. C.V.H., genome assembly and analysis, transcriptome analysis, preparation of figures, writing of manuscript. R.M.K., analysis of venom genes, comparative genomics. H.M.IJ.K, sequence analysis, comparative genomics, drawing of figures. H.P.S,, study concept and design, genome and transcriptome sequencing, sequence analysis. H.J.J. genome and transcriptome

sequencing. S.A.H. sequence analysis, comparative genomics, drawing of figures. P.A. study design and financing. G.E.E.J.M.vd.T, sequencing facilities, M.B. and W.P., assembly, . R.P.H.D., sequence analysis M.K.R., project leader, study concept and design, sequence analysis, preparation of figures and manuscript.