Massive evolutionary expansion of venom genes in the king cobra

One sentence summary:

Sequencing and data mining of the king cobra genome and transcriptomes reveals an astonishing expansion of venom genes by duplication and other mechanisms.

Snake venom has evolved into a lethal cocktail of active compounds. These act synergistically to disrupt vital functions in the person or animal bitten. Virtually nothing is known at the genomic level about how the venom gland came to express such a wide array of active molecules. We have sequenced the king cobra (*Ophiophagus hannah*) genome and deep-sequenced its venom gland transcriptome. We find an astonishing diversity of mechanisms of snake toxin radiation, including repeated gene duplication leading to increased transcript abundance. We also show for the first time how harmless ancestral genes have become recruited to the venom gland. This first snake genome, and its comparison with genomes of ancestors, could help unravel the molecular basis of the evolution of new gene function.

Freek J. Vonk[1,2,*], Christiaan V. Henkel[3*], R. Manjunatha Kini[4*], Harald M. IJ. Kerkkamp[1], Herman P. Spaink[1], Hans J. Jansen[3], S. Asad Hyder[1], Pim Arntzen[2], Guido E.E.J.M. van den Thillart[1,3], Marten Boetzer[5], Walter Pirovano[5], Ron P.H. Dirks[3] & Michael K. Richardson[1]

*These authors contributed equally to this work.

¶Corresponding Author

1 Leiden University, Institute of Biology, Sylvius Laboratory, Sylviusweg 72, 2333 BE, Leiden, the Netherlands.

2 Netherlands Centre for Biodiversity Naturalis, P.O. Box 9517, 2300 RA Leiden, the Netherlands.

3 ZF-screens B.V., Niels Bohrweg 11, 2333 CA Leiden, the Netherlands.

4 Protein Science Laboratory, Department of Biological Sciences, National University of Singapore, Science Drive 4, Singapore 117543.

5 BaseClear B.V., Einsteinweg 5, 2333 CC Leiden, the Netherlands.

Snake venom is a complex mixture of proteins and peptides evolved to immobilize prey and deter enemies(*1*). It is produced in a post-orbital venom gland which may have evolved from an ancestral gland in the posterior part of the mouth(*2*). One hypothesis of snake venom evolution envisages the duplication of normal physiological genes, followed by recruitment and expression in the venom gland(*3-6*). However, the identification of duplicates in snakes has been impossible due to the absence of a snake genome. Furthermore, a recent analysis of the genome of a venomous mammal, the platypus, found that gene duplication accounted for only a minor part of venom evolution(*7*).

To examine these issues, 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 (**SOI Methods**). Haploid genome size was estimated using flow cytometry to be around 1.36-1.59 Gbp (**SOI Fig 1a**). Our assembled draft has an N50 contig size of 3,982 bp, and an N50 scaffold size of 226 Kbp. The contigs sum to 1.45 Gbp, and the scaffolds (which contain gaps) to 1.66 Gbp.

Mitochondrial genome phylogeny confirms that the male specimen we used for genome sequencing clusters in the *Ophiophagus* group with other king cobras (**SOI Fig 2b**). Using Augustus gene prediction(*8*), and our transcriptome data (**Figure 1**), we estimate that the king cobra has approximately 22,183 protein-coding genes. 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(*9-11*).

We identified 17 different toxin families in the venom gland transcriptome by blasting against reference sequences (from www.ncbi.nlm.nih.gov) and annotated nine of them in the genome (**Figure 1**). These 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.

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(*12, 13*). We find evidence for massive expansion in the genome in both these families. We found seven different exons-2 that belong to PLA2 (**SOI Fig 2**). These genomic sequences do not contain premature stop codons or frameshifts (**SOI Fig 2**) indicating that they do not contain pseudogenes. 3FTXs are three-exon genes, of which the second exon is most readily identified. We found 21 of these exons-2 in the genome (**Figure 2**). 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 expression levels differ by five orders of magnitude (**Figure 2**). One non-expressed isoform (isoform 19) contains a premature stop codon and may be part of a pseudogene (**SOI Fig 3**). 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 (**Figure 2**).

There is a substantial difference in expression levels of each of the 3FTX isoform (**Figure 2**). 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(*14*). In general, we find that a high genomic copy number is associated with a high relative expression value (**Figure 2**). All highly expressed 3FTX genes share sequence similarities (**SOI Fig 3**).

Reptile venom CRISPs act as regulators of several types of ion channels(*15*). We find three CRISP genes in tandem in the king cobra genome (**Figure 3**) only two which are represented in our venom gland transcriptome (**SOI Fig 4a**). Together with our comparative genomic data (**Figure 3**) 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.

Venom metalloproteinases belong to the ADAM family and target various stages of blood coagulation and platelet aggregation and are responsible for hemorrhage(*16*). We also find three ADAM genes in tandem (**SOI Fig 5a**), only one of which was expressed in the venom gland transcriptome (**SOI Fig 5b-d**). There are additional metalloproteinase genes on different scaffolds.

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(*17*). We find two LAAO genes on two different scaffolds (**Figure 4**a). Based on the mapping of venom gland transcriptome reads (**SOI Fig 6**), only one LAAO gene appears to be expressed in the venom gland; the other is presumably the 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.

The physiological role of venom NGF is not clear(*18*). 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 (**SOI Fig 7**). 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(*19*). 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 (**Figure 4b**). Only the gene corresponding to HYALP1 is expressed in the venom gland (**SOI Fig 8**), which is interesting because in the mouse this gene appears to be inactive(*20*). 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.

Recently, PL-B was also found to be expressed in the venom gland(*21*) but its role in toxicity is yet unclear. We could only find one PL-B gene (**SOI Fig 9**). 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(*22*) that both the neuronal and the venom enzymes are encoded by the same gene, although alternatively spliced (**SOI Fig 10**).

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(*23*). 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.

The king cobra genome indicates that a whole array of mechanism of molecular evolution have been mobilised in venom evolution. We believe that this previously unknown diversity of mechanisms is a reflection of the multiple selective pressures on venom composition. There is evidence that not only the enemies of a snake(*24*) but also its range of prey species(*25*) can influence venom composition. Other possible selective pressures on venom composition include the need for dynamic change of venom composition over time, in order to combat the development of resistance in the opponents; and the targeting of multiple pharmacological pathways with a cocktail of venoms, providing resistance and an increase in potency.

More generally, the results show that mechanisms of molecular evolution in a given system will depend on phylogeny and selection pressures. For our results here, from a venomous reptile, are in contrast to findings from the duck-billed platypus (*Ornithorhynchus anatinus*), a venomous mammal. In that species, duplication does not appear to have been a dominant mechanisms of venom evolution, and the difference could be related to the different function of venom: the male platypus may only use its venom in competition with other males(*7*). We are currently comparing the king cobra sequences with those from other snakes in order to examine these fundamental issues in more detail. We believe that it could help unravel the molecular basis of the evolution of new gene functions.

ACKNOWLEDGEMENTS

We thank Austin Hughes, thank Daniëlle de Wijze and Yuki Minegishi for discussions. This was funded by The Netherlands Centre for Biodiversity Naturalis and the Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science. Jeroen Admiraal helped with illustrations.

REFERENCES

1. F. J. Vonk *et al.*, *Bioessays* **33**, 269 (2011).
2. F. J. Vonk *et al.*, *Nature* **454**, 630 (2008).
3. B. G. Fry, *Genome Research* **15**, 403 (2005).
4. S. Kwong, A. E. Woods, P. J. Mirtschin, R. Ge, R. M. Kini, *Thrombosis. and haemostasis* **102**, 469 (2009).
5. T. N. Minh Le, M. A. Reza, S. Swarup, R. M. Kini, *Thrombosis. and haemostasis* **93**, 420 (2005).
6. M. A. Reza, S. Swarup, R. M. Kini, *Pathophysiology. of haemostasis and thrombosis.* **34**, 205 (2005).
7. E. S. Wong, A. T. Papenfuss, C. M. Whittington, W. C. Warren, K. Belov, *Molecular. biology. and evolution* (2011).
8. M. Stanke, O. Schoffmann, B. Morgenstern, S. Waack, *BMC. Bioinformatics.* **7**, 62 (2006).
9. W. C. Warren *et al.*, *Nature* **453**, 175 (2008).
10. R. Li *et al.*, *Nature* **463**, 311 (2010).
11. J. W. Wallis *et al.*, *Nature* **432**, 761 (2004).
12. R. M. Kini, R. Doley, *Toxicon* **56**, 855 (2010).
13. *Venom Phospholipase A2 Enzymes: Structure, Function and Mechanism* (**John Wiley & Sons**, **Chichester, England**, 1997), p. -511.
14. N. R. Casewell, R. A. Harrison, W. Wuster, S. C. Wagstaff, *BMC. Genomics* **10**, 564 (2009).
15. G. M. Gibbs, M. K. O'Bryan, *Soc Reprod Fertil. Suppl* **65**, 261 (2007).
16. A. M. Moura-da-Silva, D. Butera, I. Tanjoni, *Current. pharmaceutical. design.* **13**, 2893 (2007).
17. X. Y. Du, K. J. Clemetson, *Toxicon* **40**, 659 (2002).

18. T. Kostiza, J. Meier, *Toxicon* **34**, 787 (1996).

19. K. Kemparaju, K. S. Girish, *Cell biochemistry and function.* **24**, 7 (2006).

20. S. Reitinger *et al.*, *Protein expression and purification.* **57**, 226 (2008).

21. S. T. Chatrath *et al.*, *Journal of proteome research.* **10**, 739 (2011).

22. X. Cousin, S. Bon, J. Massoulie, C. Bon, *The Journal of biological. chemistry.* **273**, 9812 (1998).

23. M. A. Reza, S. Swarup, R. M. Kini, *J Thromb Haemost* **5**, 117 (2007).

24. S. A. Jansa, R. S. Voss, *PLoS. ONE.* **6**, e20997 (2011).

25. S. Pahari, D. Bickford, B. G. Fry, R. M. Kini, *BMC evolutionary. biology* **7**, 175 (2007).

26. Z. J. Jiang *et al.*, *BMC evolutionary. biology* **7**, 123 (2007).

27. N. Chen, S. Zhao, *Mitochondrial. DNA* **20**, 69 (2009).

28. T. A. Castoe *et al.*, *Cytogenet. Genome Res* **127**, 112 (2009).

29. J. Yan, H. Li, K. Zhou, *BMC Genomics* **9**, 569 (2008).

FIGURES

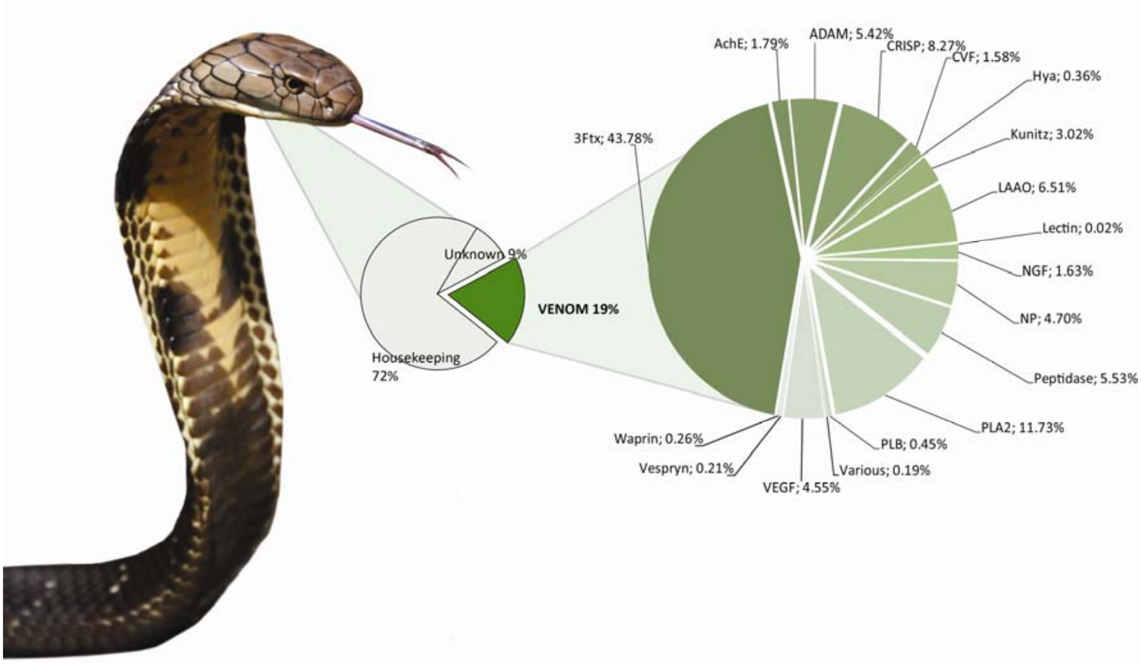


Figure 1. Relative abundance of the venom toxins in the transcriptome. The percentages are calculated based on the expression value of the transcripts sequenced from the venom gland transcriptome. The most abundant family is the three-finger toxins (43.78% of all toxin transcripts identified), represented in the genome by at least 21 different isoforms (see also **Figure 2**).

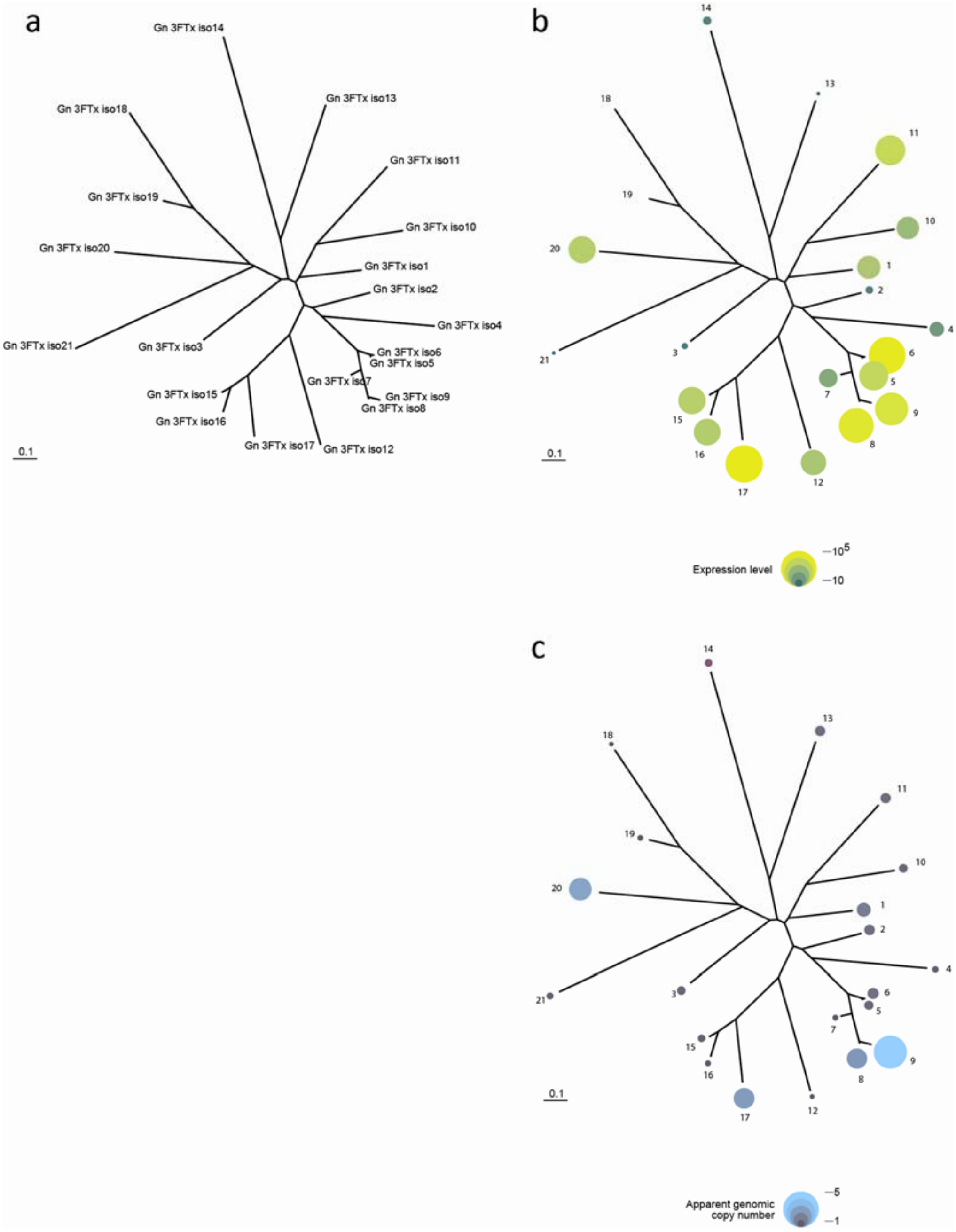


Figure 2. Unrooted phylogenetic tree constructed from all different exon-2 sequences of the three-finger toxin genes. Isoform 19 contains a premature stop codon, thus most likely is a pseudogene. Green circles indicate relative expression levels (on a logarithmic scale), blue circles apparent genomic copy numbers, both based on local coverage by venom gland

transcriptomic sequencing reads or genomic sequencing reads, respectively. **a**) with gene labels; **b**) the same with transcript abundance in the venom gland transcriptome; **c**) the same showing number of copies in genome.

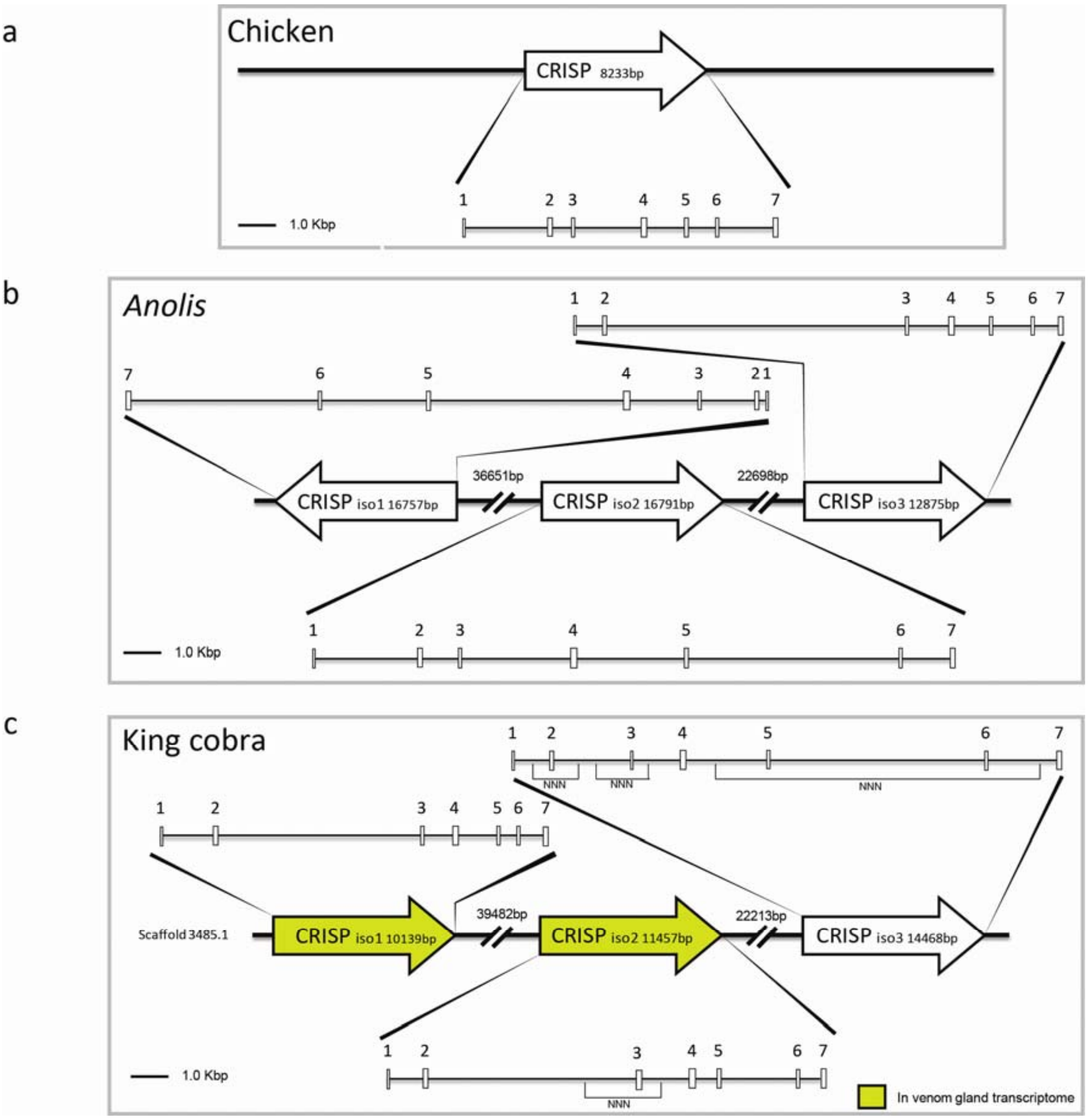


Figure 3. Comparative genomic architecture of the CRISP genes. **a,** chicken (*Gallus gallu*s); **b,** anole lizard (*Anolis carolinensis*); and **c,** King cobra (*Ophiophagus hannah*). Chick and *Anolis* sequences are from www.ensembl.org. The exploded views show scale diagrams of the exons and introns. 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.

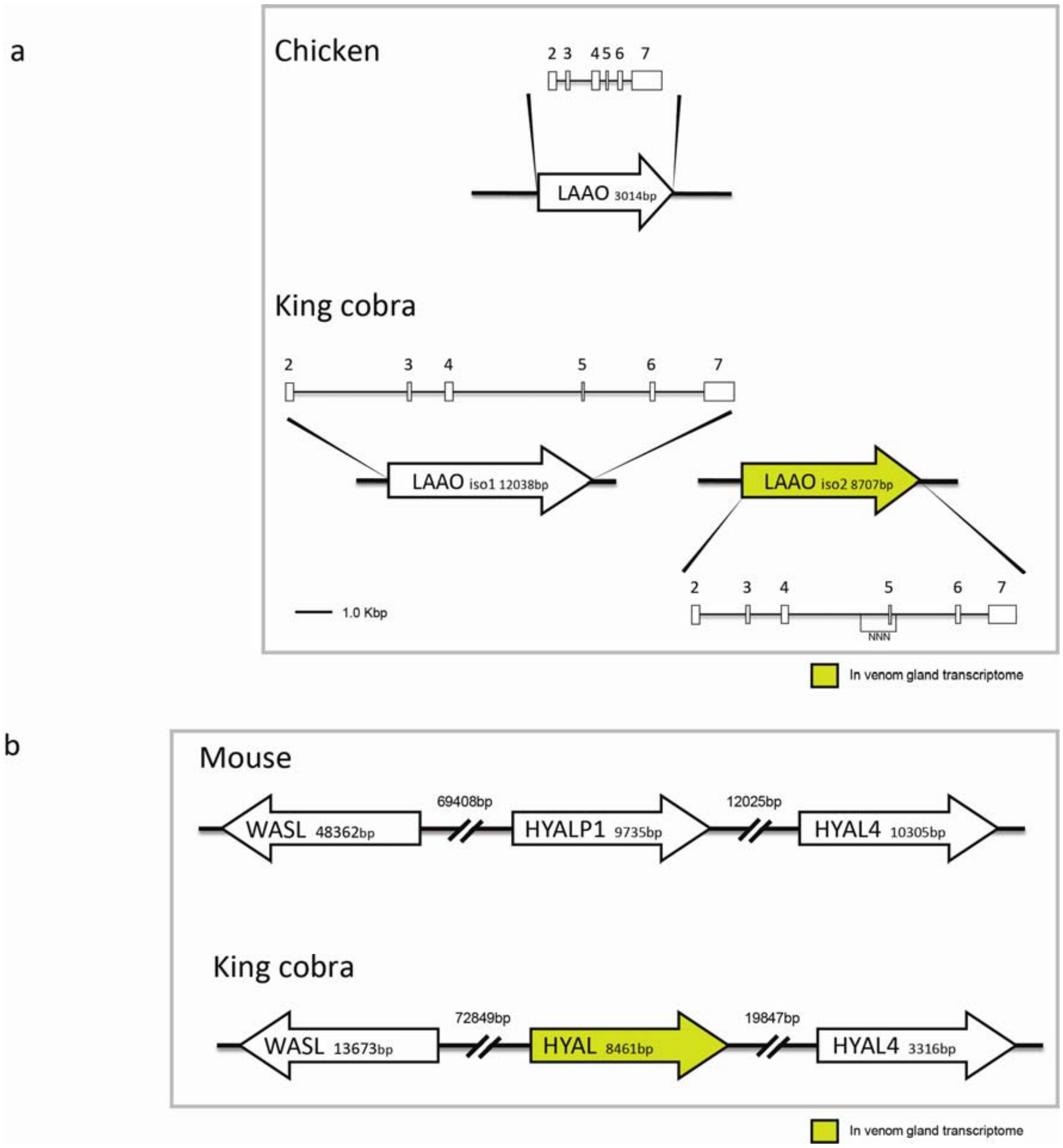


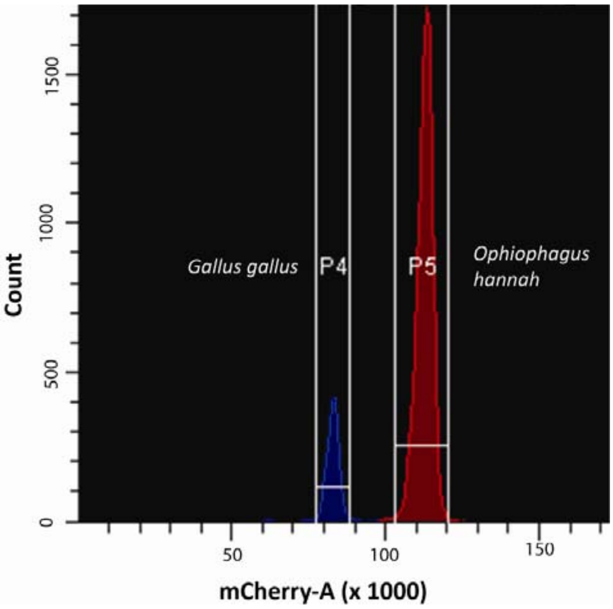
Figure 4. a, Genomic architecture of l-amino acid oxidase (LAAO) genes in the chicken and king cobra. **b,** scheme of the genomic context of the hyaluronidase genes in the mouse (*Mus musculus*) and king cobra. Mouse genomic sequences from www.ensembl.org. Scale bar refers to the exploded views. NNN, unresolved sequence.

SUPPORTING ONLINE MATERIAL

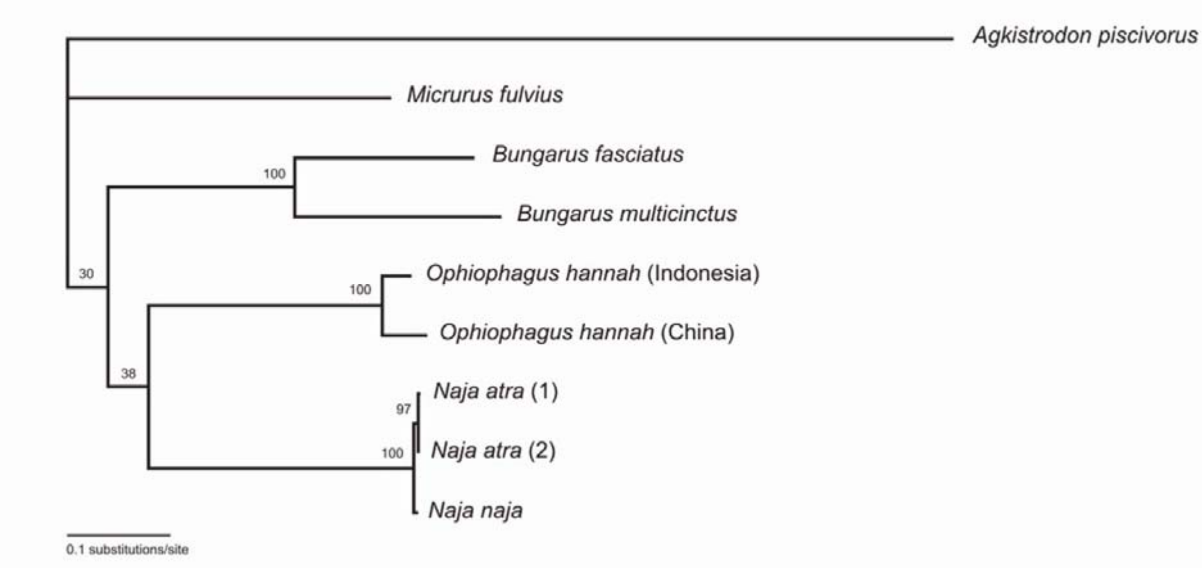
SOI Figure 1

a flow cytometry; **b**, mtDNA phylogeny of king cobra.

a

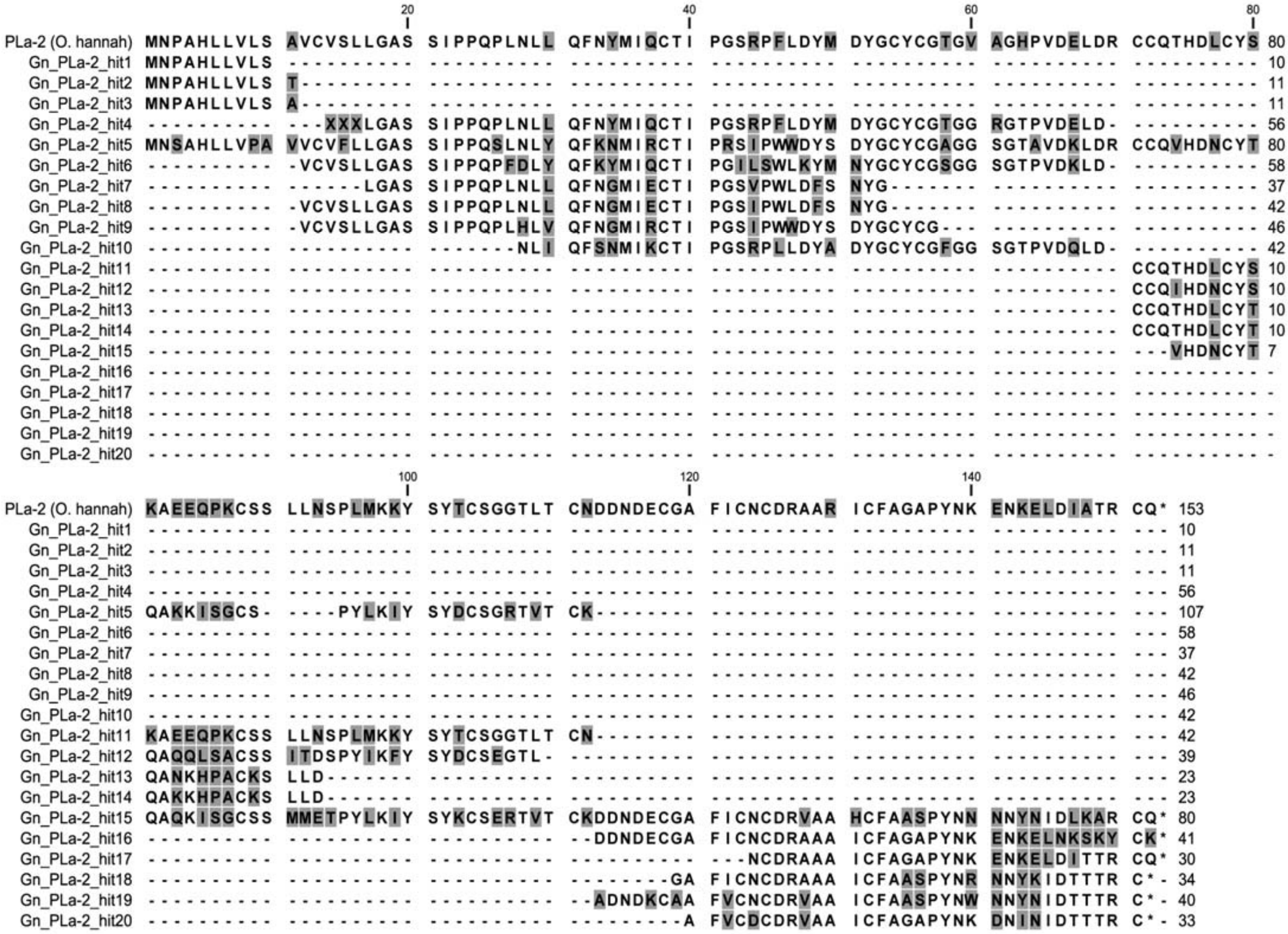


b



SOI Figure 2

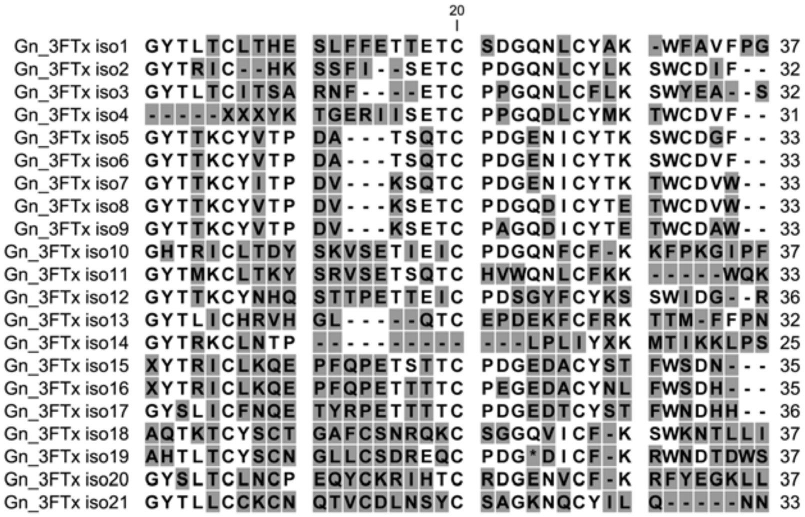
Alignment of multiple PLA2 genomic hits.



```
                                      20                    40                    60                    80
                                       |                     |                     |                     |
PLa-2 (O. hannah)  MNPAHLLVLS AVCVSLLGAS SIPPQPLNLL QFNYMIQCTI PGSRPFLDYM DYGCYCGTGV AGHPVDELDR CCQTHDLCYS 80
Gn_PLa-2_hit1      MNPAHLLVLS ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10
Gn_PLa-2_hit2      MNPAHLLVLS T--------- ---------- ---------- ---------- ---------- ---------- ---------- 11
Gn_PLa-2_hit3      MNPAHLLVLS A--------- ---------- ---------- ---------- ---------- ---------- ---------- 11
Gn_PLa-2_hit4      ---------- ---XXXLGAS SIPPQPLNLL QFNYMIQCTI PGSRPFLDYM DYGCYCGTGG RGTPVDELD- ---------- 56
Gn_PLa-2_hit5      MNSAHLLVPA VVCVFLLGAS SIPPQSLNLY QFKNMIRCTI PRSIPWWDYS DYGCYCGAGG SGTAVDKLDR CCQVHDNCYT 80
Gn_PLa-2_hit6      ---------- -VCVSLLGAS SIPPQPFDLY QFKYMIQCTI PGILSWLKYM NYGCYCGSGG SGTPVDKLD- ---------- 58
Gn_PLa-2_hit7      ---------- ------LGAS SIPPQPLNLL QFNGMIECTI PGSVPWLDFS NYG------- ---------- ---------- 37
Gn_PLa-2_hit8      ---------- -VCVSLLGAS SIPPQPLNLL QFNGMIECTI PGSIPWLDFS NYG------- ---------- ---------- 42
Gn_PLa-2_hit9      ---------- -VCVSLLGAS SIPPQPLHLV QFNGMIRCTI PGSIPWWDYS DYGCYCG--- ---------- ---------- 46
Gn_PLa-2_hit10     ---------- ---------- -------NLI QFSNMIKCTI PGSRPLLDYA DYGCYCGFGG SGTPVDQLD- ---------- 42
Gn_PLa-2_hit11     ---------- ---------- ---------- ---------- ---------- ---------- ---------- CCQTHDLCYS 10
Gn_PLa-2_hit12     ---------- ---------- ---------- ---------- ---------- ---------- ---------- CCQIHDNCYS 10
Gn_PLa-2_hit13     ---------- ---------- ---------- ---------- ---------- ---------- ---------- CCQTHDLCYT 10
Gn_PLa-2_hit14     ---------- ---------- ---------- ---------- ---------- ---------- ---------- CCQTHDLCYT 10
Gn_PLa-2_hit15     ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---VHDNCYT 7
Gn_PLa-2_hit16     ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -
Gn_PLa-2_hit17     ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -
Gn_PLa-2_hit18     ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -
Gn_PLa-2_hit19     ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -
Gn_PLa-2_hit20     ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -

                                     100                   120                   140
                                       |                     |                     |
PLa-2 (O. hannah)  KAEEQPKCSS LLNSPLMKKY SYTCSGGTLT CNDDNDECGA FICNCDRAAR ICFAGAPYNK ENKELDIATR CQ* 153
Gn_PLa-2_hit1      ---------- ---------- ---------- ---------- ---------- ---------- ---------- --- 10
Gn_PLa-2_hit2      ---------- ---------- ---------- ---------- ---------- ---------- ---------- --- 11
Gn_PLa-2_hit3      ---------- ---------- ---------- ---------- ---------- ---------- ---------- --- 11
Gn_PLa-2_hit4      ---------- ---------- ---------- ---------- ---------- ---------- ---------- --- 56
Gn_PLa-2_hit5      QAKKISGCS- ----PYLKIY SYDCSGRTVT CK-------- ---------- ---------- ---------- --- 107
Gn_PLa-2_hit6      ---------- ---------- ---------- ---------- ---------- ---------- ---------- --- 58
Gn_PLa-2_hit7      ---------- ---------- ---------- ---------- ---------- ---------- ---------- --- 37
Gn_PLa-2_hit8      ---------- ---------- ---------- ---------- ---------- ---------- ---------- --- 42
Gn_PLa-2_hit9      ---------- ---------- ---------- ---------- ---------- ---------- ---------- --- 46
Gn_PLa-2_hit10     ---------- ---------- ---------- ---------- ---------- ---------- ---------- --- 42
Gn_PLa-2_hit11     KAEEQPKCSS LLNSPLMKKY SYTCSGGTLT CN-------- ---------- ---------- ---------- --- 42
Gn_PLa-2_hit12     QAQQLSACSS ITDSPYIKFY SYDCSEGTL- ---------- ---------- ---------- ---------- --- 39
Gn_PLa-2_hit13     QANKHPACKS LLD------- ---------- ---------- ---------- ---------- ---------- --- 23
Gn_PLa-2_hit14     QAKKHPACKS LLD------- ---------- ---------- ---------- ---------- ---------- --- 23
Gn_PLa-2_hit15     QAQKISGCSS MMETPYLKIY SYKCSERTVT CKDDNDECGA FICNCDRVAA HCFAASPYNN NNYNIDLKAR CQ* 80
Gn_PLa-2_hit16     ---------- ---------- ---------- --DDNDECGA FICNCDRAAA ICFAGAPYNK ENKELNKSKY CK* 41
Gn_PLa-2_hit17     ---------- ---------- ---------- ---------- ---NCDRAAA ICFAGAPYNK ENKELDITTR CQ* 30
Gn_PLa-2_hit18     ---------- ---------- ---------- --------GA FICNCDRAAA ICFAASPYNR NNYKIDTTTR C*- 34
Gn_PLa-2_hit19     ---------- ---------- ---------- --ADNDKCAA FVCNCDRVAA ICFAASPYNW NNYNIDTTTR C*- 40
Gn_PLa-2_hit20     ---------- ---------- ---------- ---------A FVCDCDRVAA ICFAGAPYNK DNINIDTTTR C*- 33
```

SOI Figure 3

Alignment of multiple 3FTx exon2 isoforms.

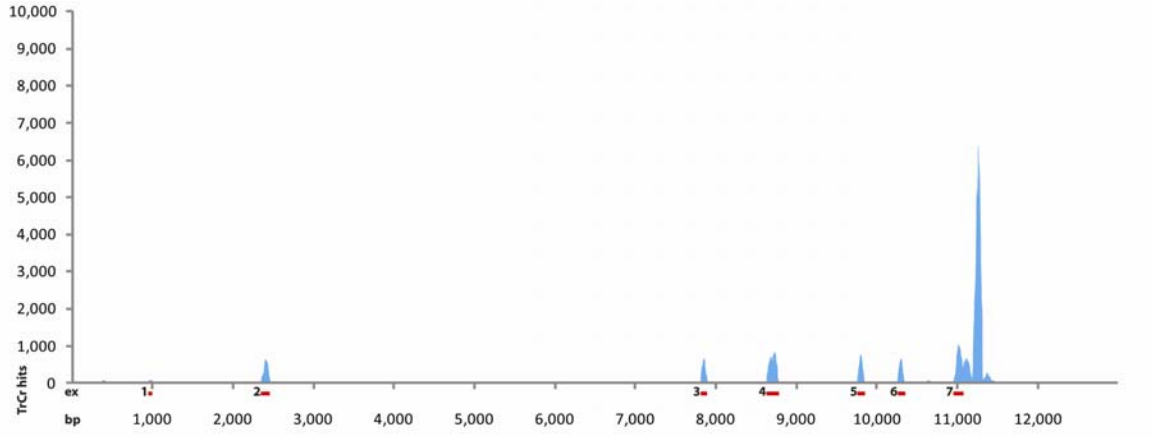


```
                                        20
                                         |
Gn_3FTx iso1   GYTLTCLTHE  SLFFETTETC  SDGQNLCYAK  -WFAVFPG 37
Gn_3FTx iso2   GYTRIC--HK  SSFI--SETC  PDGQNLCYLK  SWCDIF-- 32
Gn_3FTx iso3   GYTLTCITSA  RNF----ETC  PPGQNLCFLK  SWYEA--S 32
Gn_3FTx iso4   -----XXXYK  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
```

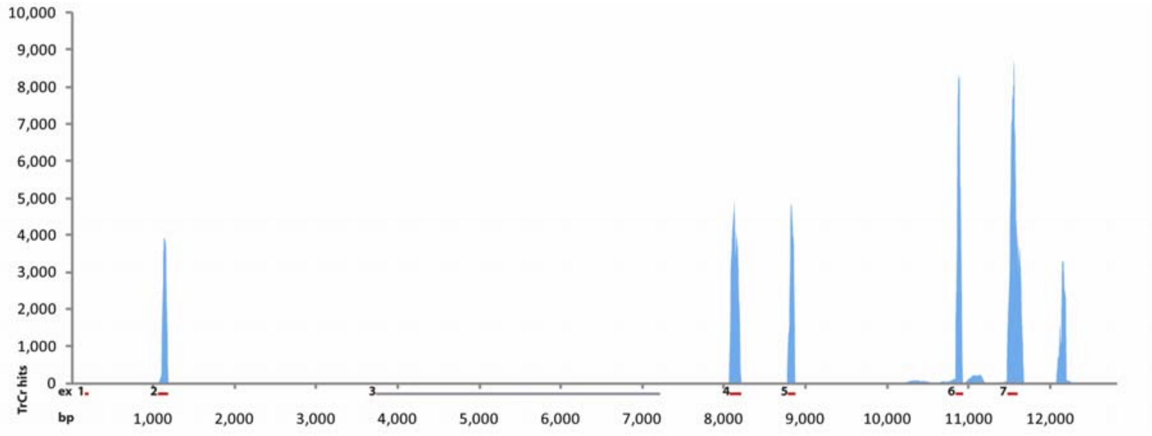
SOI Figure 4

a-c The scaffold containing three CRISP genes with different isoform transcripts (see main text **Figure 3c** for further details) 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.

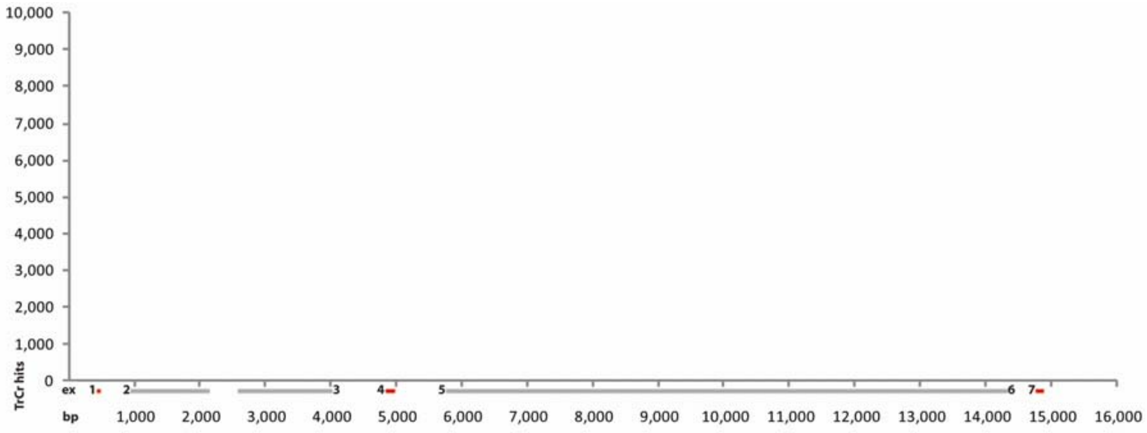
a



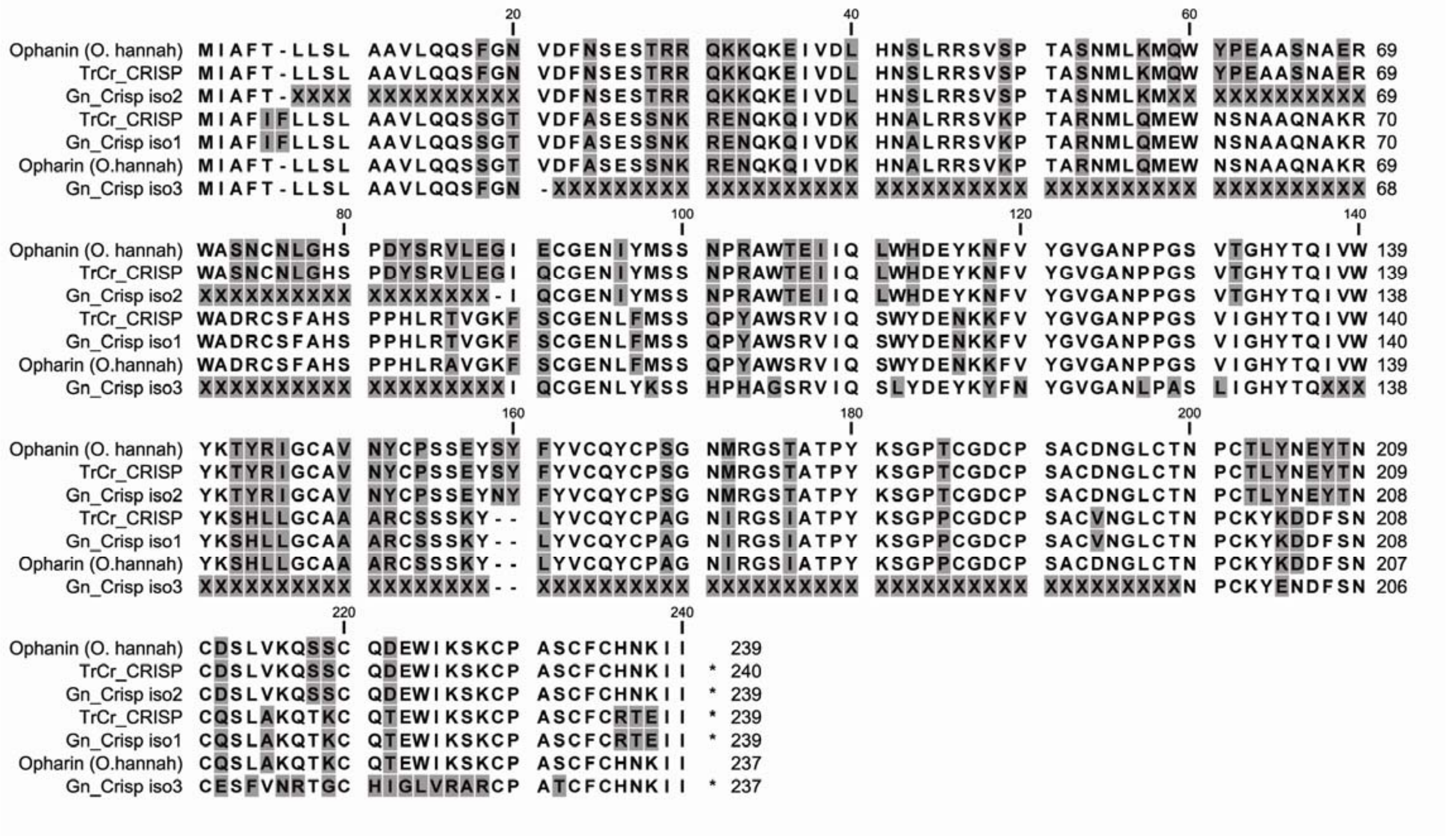
b



c



d

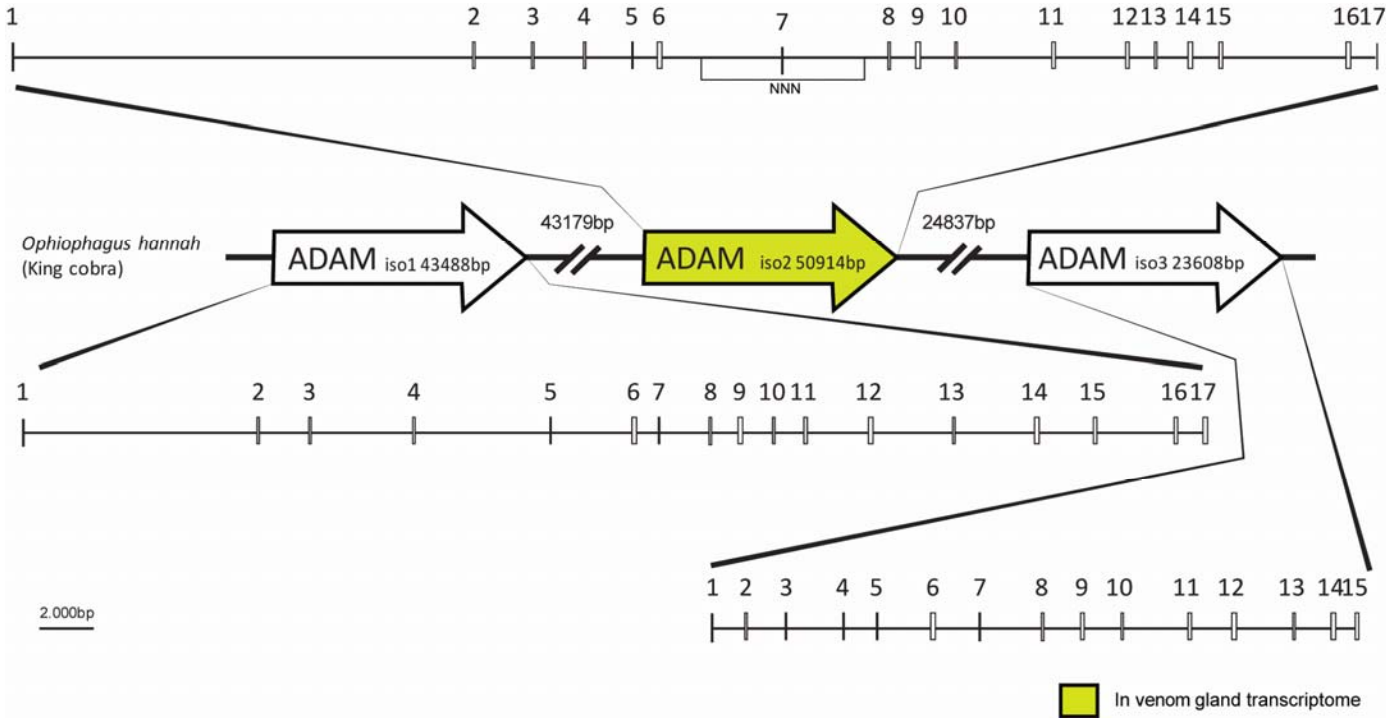


```
                                                20                             40                             60
                                                 |                              |                              |
Ophanin (O. hannah)  MIAFT-LLSL AAVLQQSFGN VDFNSESTRR QKKQKEIVDL HNSLRRSVSP TASNMLKMQW YPEAASNAER 69
         TrCr_CRISP  MIAFT-LLSL AAVLQQSFGN VDFNSESTRR QKKQKEIVDL HNSLRRSVSP TASNMLKMQW YPEAASNAER 69
       Gn_Crisp iso2  MIAFT-XXXX XXXXXXXXXX VDFNSESTRR QKKQKEIVDL HNSLRRSVSP TASNMLKMXX XXXXXXXXXX 69
         TrCr_CRISP  MIAFIFLLSL AAVLQQSSGT VDFASESSNK RENQKQIVDK HNALRRSVKP TARNMLQMEW NSNAAQNAKR 70
       Gn_Crisp iso1  MIAFIFLLSL AAVLQQSSGT VDFASESSNK RENQKQIVDK HNALRRSVKP TARNMLQMEW NSNAAQNAKR 70
 Opharin (O.hannah)  MIAFT-LLSL AAVLQQSSGT VDFASESSNK RENQKQIVDK HNALRRSVKP TARNMLQMEW NSNAAQNAKR 69
       Gn_Crisp iso3  MIAFT-LLSL AAVLQQSFGN -XXXXXXXXX XXXXXXXXXX XXXXXXXXXX XXXXXXXXXX XXXXXXXXXX 68

                             80                            100                            120                            140
                              |                              |                              |                              |
Ophanin (O. hannah)  WASNCNLGHS PDYSRVLEGI ECGENIYMSS NPRAWTEIIQ LWHDEYKNFV YGVGANPPGS VTGHYTQIVW 139
         TrCr_CRISP  WASNCNLGHS PDYSRVLEGI QCGENIYMSS NPRAWTEIIQ LWHDEYKNFV YGVGANPPGS VTGHYTQIVW 139
       Gn_Crisp iso2  XXXXXXXXXX XXXXXXXX-I QCGENIYMSS NPRAWTEIIQ LWHDEYKNFV YGVGANPPGS VTGHYTQIVW 138
         TrCr_CRISP  WADRCSFAHS PPHLRTVGKF SCGENLFMSS QPYAWSRVIQ SWYDENKKFV YGVGANPPGS VIGHYTQIVW 140
       Gn_Crisp iso1  WADRCSFAHS PPHLRTVGKF SCGENLFMSS QPYAWSRVIQ SWYDENKKFV YGVGANPPGS VIGHYTQIVW 140
 Opharin (O.hannah)  WADRCSFAHS PPHLRAVGKF SCGENLFMSS QPYAWSRVIQ SWYDENKKFV YGVGANPPGS VIGHYTQIVW 139
       Gn_Crisp iso3  XXXXXXXXXX XXXXXXXXXI QCGENLYKSS HPHAGSRVIQ SLYDEYKYFN YGVGANLPAS LIGHYTQXXX 138

                                                160                            180                            200
                                                 |                              |                              |
Ophanin (O. hannah)  YKTYRIGCAV NYCPSSEYSY FYVCQYCPSG NMRGSTATPY KSGPTCGDCP SACDNGLCTN PCTLYNEYTN 209
         TrCr_CRISP  YKTYRIGCAV NYCPSSEYSY FYVCQYCPSG NMRGSTATPY KSGPTCGDCP SACDNGLCTN PCTLYNEYTN 209
       Gn_Crisp iso2  YKTYRIGCAV NYCPSSEYNY FYVCQYCPSG NMRGSTATPY KSGPTCGDCP SACDNGLCTN PCTLYNEYTN 208
         TrCr_CRISP  YKSHLLGCAA ARCSSSKY-- LYVCQYCPAG NIRGSIATPY KSGPPCGDCP SACVNGLCTN PCKYKDDFSN 208
       Gn_Crisp iso1  YKSHLLGCAA ARCSSSKY-- LYVCQYCPAG NIRGSIATPY KSGPPCGDCP SACVNGLCTN PCKYKDDFSN 208
 Opharin (O.hannah)  YKSHLLGCAA ARCSSSKY-- LYVCQYCPAG NIRGSIATPY KSGPPCGDCP SACDNGLCTN PCKYKDDFSN 207
       Gn_Crisp iso3  XXXXXXXXXX XXXXXXXX-- XXXXXXXXXX XXXXXXXXXX XXXXXXXXXX XXXXXXXXXN PCKYENDFSN 206

                             220                            240
                              |                              |
Ophanin (O. hannah)  CDSLVKQSSC QDEWIKSKCP ASCFCHNKII   239
         TrCr_CRISP  CDSLVKQSSC QDEWIKSKCP ASCFCHNKII * 240
       Gn_Crisp iso2  CDSLVKQSSC QDEWIKSKCP ASCFCHNKII * 239
         TrCr_CRISP  CQSLAKQTKC QTEWIKSKCP ASCFCRTEII * 239
       Gn_Crisp iso1  CQSLAKQTKC QTEWIKSKCP ASCFCRTEII * 239
 Opharin (O.hannah)  CQSLAKQTKC QTEWIKSKCP ASCFCHNKII   237
       Gn_Crisp iso3  CESFVNRTGC HIGLVRARCP ATCFCHNKII * 237
```

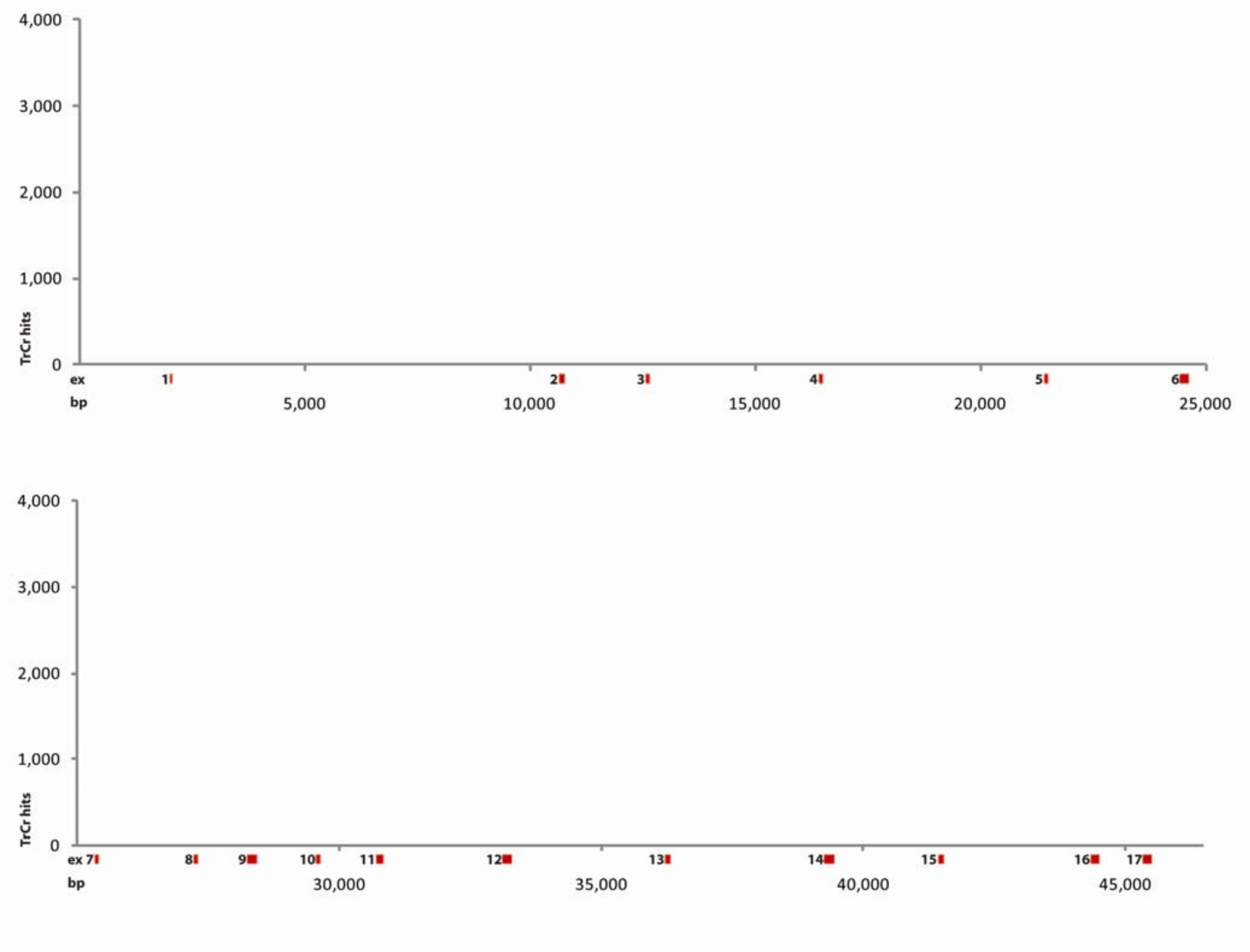
SOI Figure 5

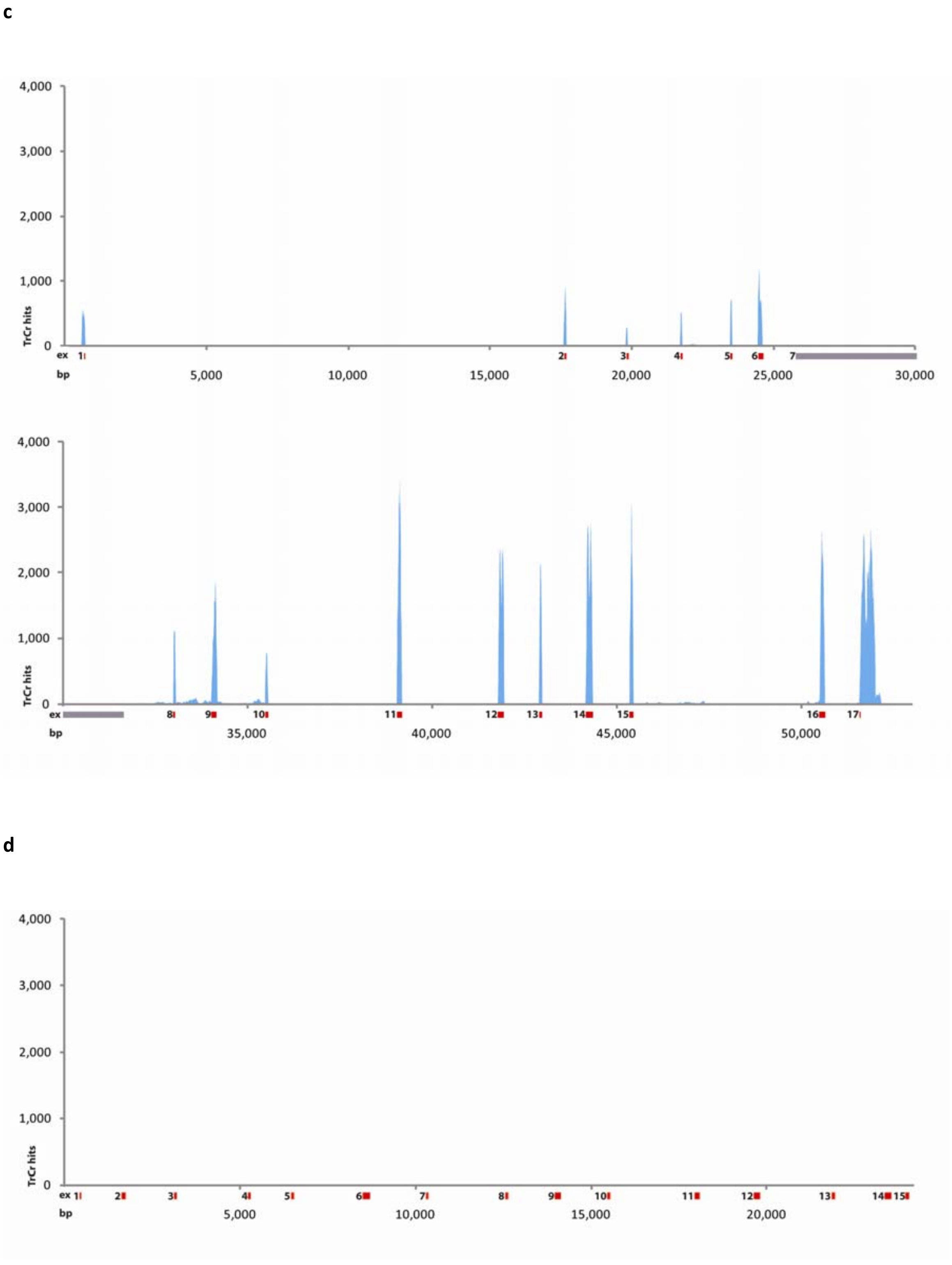
a) the scaffold containing three ADAM genes; **b**) isoform 1; **c**) isoform 2; **d**) isoform 3. As can be seen, only isoform 2 is expressed in the venom gland; **e**) amino acid alignments of these three metalloproteinase genes with the single transcriptome sequence shows that one gene is identical 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* isoform 3 appears to be expressed, since *N. atra* metalloproteinase is more similar to isoform 3 than isoform 2.

a

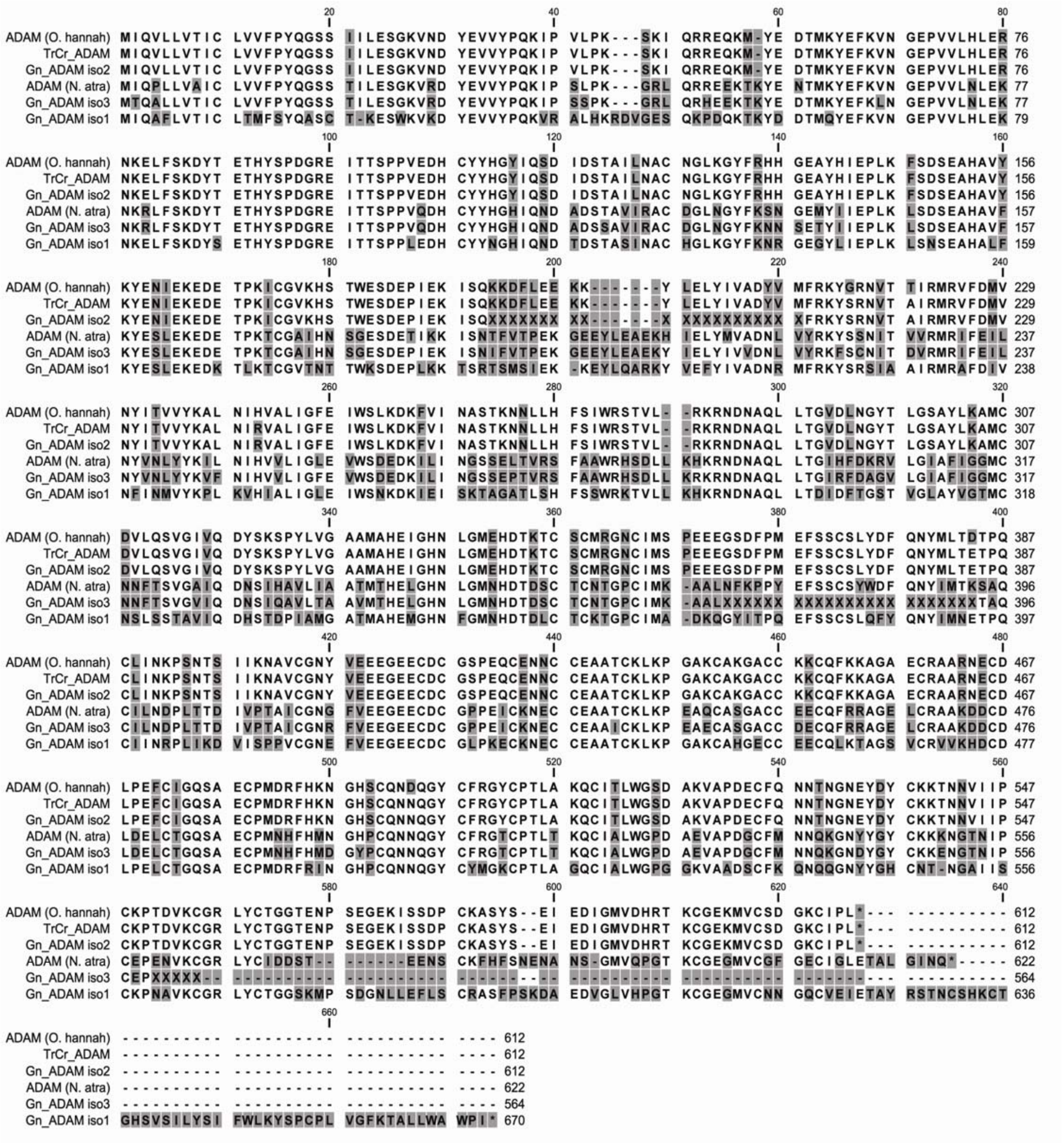


b


4,000
3,000
2,000
1,000
0
TrCr hits
ex
1
2
3
4
5
6
bp
5,000
10,000
15,000
20,000
25,000
4,000
3,000
2,000
1,000
0
TrCr hits
ex 7
8
9
10
11
12
13
14
15
16
17
bp
30,000
35,000
40,000
45,000


c
4,000
3,000
2,000
1,000
0
TrCr hits
ex 1 2 3 4 5 6 7
bp 5,000 10,000 15,000 20,000 25,000 30,000
4,000
3,000
2,000
1,000
0
TrCr hits
ex 8 9 10 11 12 13 14 15 16 17
bp 35,000 40,000 45,000 50,000
d
4,000
3,000
2,000
1,000
0
TrCr hits
ex 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
bp 5,000 10,000 15,000 20,000

e

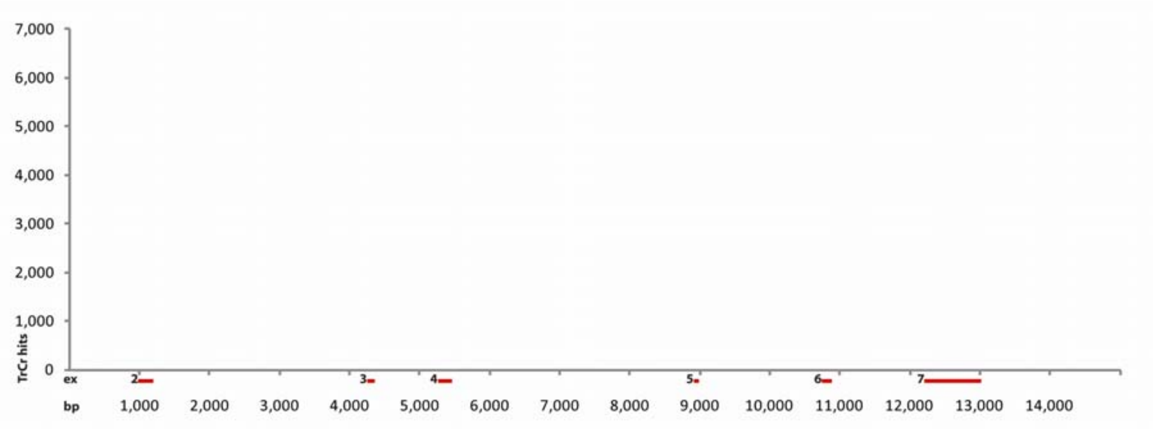


```
                                     20         40         60         80
                                      |          |          |          |
ADAM (O. hannah) MIQVLLVTIC LVVFPYQGSS IILESGKVND YEVVYPQKIP VLPK---SKI QRREQKM-YE DTMKYEFKVN GEPVVLHLER 76
       TrCr_ADAM MIQVLLVTIC LVVFPYQGSS IILESGKVND YEVVYPQKIP VLPK---SKI QRREQKM-YE DTMKYEFKVN GEPVVLHLER 76
    Gn_ADAM iso2 MIQVLLVTIC LVVFPYQGSS IILESGKVND YEVVYPQKIP VLPK---SKI QRREQKM-YE DTMKYEFKVN GEPVVLHLER 76
  ADAM (N. atra) MIQPLLVAIC LVVFPYQGSS TILESGKVRD YEVVYPQKIP SLPK---GRL QRREEKTKYE NTMKYEFKVN GEPVVLNLEK 77
    Gn_ADAM iso3 MTQALLVTIC LVVFPYQGSS TILESGKVRD YEVVYPQKIP SSPK---GRL QRHEEKTKYE DTMKYEFKLN GEPVVLNLEK 77
    Gn_ADAM iso1 MIQAFLVTIC LTMFSYQASC T-KESWKVKD YEVVYPQKVR ALHKRDVGES QKPDQKTKYD DTMQYEFKVN GEPVVLHLEK 79
                                    100        120        140        160
                                      |          |          |          |
ADAM (O. hannah) NKELFSKDYT ETHYSPDGRE ITTSPPVEDH CYYHGYIQSD IDSTAILNAC NGLKGYFRHH GEAYHIEPLK FSDSEAHAVY 156
       TrCr_ADAM NKELFSKDYT ETHYSPDGRE ITTSPPVEDH CYYHGYIQSD IDSTAILNAC NGLKGYFRHH GEAYHIEPLK FSDSEAHAVY 156
    Gn_ADAM iso2 NKELFSKDYT ETHYSPDGRE ITTSPPVEDH CYYHGYIQSD IDSTAILNAC NGLKGYFRHH GEAYHIEPLK FSDSEAHAVY 156
  ADAM (N. atra) NKRLFSKDYT ETHYSPDGRE ITTSPPVQDH CYYHGHIQND ADSTAVIRAC DGLNGYFKSN GEMYIIEPLK LSDSEAHAVF 157
    Gn_ADAM iso3 NKRLFSKDYT ETHYSPDGRE ITTSPPVQDH CYYHGHIQND ADSSAVIRAC DGLNGYFKNN SETYIIEPLK LSDSEAHAVF 157
    Gn_ADAM iso1 NKELFSKDYS ETHYSPDGRE ITTSPPLEDH CYYNGHIQND TDSTASINAC HGLKGYFKNR GEGYLIEPLK LSNSEAHALF 159
                                    180        200        220        240
                                      |          |          |          |
ADAM (O. hannah) KYENIEKEDE TPKICGVKHS TWESDEPIEK ISQKKDFLEE KK-------Y LELYIVADYV MFRKYGRNVT TIRMRVFDMV 229
       TrCr_ADAM KYENIEKEDE TPKICGVKHS TWESDEPIEK ISQKKDFLEE KK-------Y LELYIVADYV MFRKYSRNVT AIRMRVFDMV 229
    Gn_ADAM iso2 KYENIEKEDE TPKICGVKHS TWESDEPIEK ISQXXXXXXX XX-------X XXXXXXXXXX XFRKYSRNVT AIRMRVFDMV 229
  ADAM (N. atra) KYESLEKEDE TPKTCGAIHN SGESDETIKK ISNTFVTPEK GEEYLEAEKH IELYMVADNL VYRKYSSNIT VVRMRIFEIL 237
    Gn_ADAM iso3 KYESLEKEDE TPKTCGAIHN SGESDEPIEK ISNIFVTPEK GEEYLEAEKY IELYIVVDNL VYRKFSCNIT DVRMRIFEIL 237
    Gn_ADAM iso1 KYESLEKEDK TLKTCGVTNT TWKSDEPLKK TSRTSMSIEK -KEYLQARKY VEFYIVADNR MFRKYSRSIA AIRMRAFDIV 238
                                    260        280        300        320
                                      |          |          |          |
ADAM (O. hannah) NYITVVYKAL NIHVALIGFE IWSLKDKFVI NASTKNNLLH FSIWRSTVL- -RKRNDNAQL LTGVDLNGYT LGSAYLKAMC 307
       TrCr_ADAM NYITVVYKAL NIRVALIGFE IWSLKDKFVI NASTKNNLLH FSIWRSTVL- -RKRNDNAQL LTGVDLNGYT LGSAYLKAMC 307
    Gn_ADAM iso2 NYITVVYKAL NIRVALIGFE IWSLKDKFVI NASTKNNLLH FSIWRSTVL- -RKRNDNAQL LTGVDLNGYT LGSAYLKAMC 307
  ADAM (N. atra) NYVNLYYKIL NIHVVLIGLE VWSDEDKILI NGSSELTVRS FAAWRHSDLL KHKRNDNAQL LTGIHFDKRV LGIAFIGGMC 317
    Gn_ADAM iso3 NYVNLYYKVF NIHVVLIGFE VWSDEDKILI NGSSEPTVRS FAAWRHSDLL KRKRNDNAQL LTGIRFDAGV LGIAFIGGMC 317
    Gn_ADAM iso1 NFINMVYKPL KVHIALIGLE IWSNKDKIEI SKTAGATLSH FSSWRKTVLL KHKRNDNAQL LTDIDFTGST VGLAYVGTMC 318
                                    340        360        380        400
                                      |          |          |          |
ADAM (O. hannah) DVLQSVGIVQ DYSKSPYLVG AAMAHEIGHN LGMEHDTKTC SCMRGNCIMS PEEEGSDFPM EFSSCSLYDF QNYMLTDTPQ 387
       TrCr_ADAM DVLQSVGIVQ DYSKSPYLVG AAMAHEIGHN LGMEHDTKTC SCMRGNCIMS PEEEGSDFPM EFSSCSLYDF QNYMLTETPQ 387
    Gn_ADAM iso2 DVLQSVGIVQ DYSKSPYLVG AAMAHEIGHN LGMEHDTKTC SCMRGNCIMS PEEEGSDFPM EFSSCSLYDF QNYMLTETPQ 387
  ADAM (N. atra) NNFTSVGAIQ DNSIHAVLIA ATMTHELGHN LGMNHDTDSC TCNTGPCIMK -AALNFKPPY EFSSCSYWDF QNYIMTKSAQ 396
    Gn_ADAM iso3 NNFTSVGVIQ DNSIQAVLTA AVMTHELGHN LGMNHDTDSC TCNTGPCIMK -AALXXXXXX XXXXXXXXXX XXXXXXXTAQ 396
    Gn_ADAM iso1 NSLSSTAVIQ DHSTDPIAMG ATMAHEMGHN FGMNHDTDLC TCKTGPCIMA -DKQGYITPQ EFSSCSLQFY QNYIMNETPQ 397
                                    420        440        460        480
                                      |          |          |          |
ADAM (O. hannah) CLINKPSNTS IIKNAVCGNY VEEEGEECDC GSPEQCENNC CEAATCKLKP GAKCAKGACC KKCQFKKAGA ECRAARNECD 467
       TrCr_ADAM CLINKPSNTS IIKNAVCGNY VEEEGEECDC GSPEQCENNC CEAATCKLKP GAKCAKGACC KKCQFKKAGA ECRAARNECD 467
    Gn_ADAM iso2 CLINKPSNTS IIKNAVCGNY VEEEGEECDC GSPEQCENNC CEAATCKLKP GAKCAKGACC KKCQFKKAGA ECRAARNECD 467
  ADAM (N. atra) CILNDPLTTD IVPTAICGNG FVEEGEECDC GPPEICKNEC CEAATCKLKP EAQCASGACC EECQFRRAGE LCRAAKDDCD 476
    Gn_ADAM iso3 CILNDPLTTD IVPTAICGNR FVEEGEECDC GPPEICKNEC CEAAICKLKP EAECASGACC DECQFRRAGE LCRAAKDDCD 476
    Gn_ADAM iso1 CIINRPLIKD VISPPVCGNE FVEEGEECDC GLPKECKNEC CEAATCKLKP GAKCAHGECC EECQLKTAGS VCRVVKHDCD 477
                                    500        520        540        560
                                      |          |          |          |
ADAM (O. hannah) LPEFCIGQSA ECPMDRFHKN GHSCQNDQGY CFRGYCPTLA KQCITLWGSD AKVAPDECFQ NNTNGNEYDY CKKTNNVIIP 547
       TrCr_ADAM LPEFCIGQSA ECPMDRFHKN GHSCQNNQGY CFRGYCPTLA KQCITLWGSD AKVAPDECFQ NNTNGNEYDY CKKTNNVIIP 547
    Gn_ADAM iso2 LPEFCIGQSA ECPMDRFHKN GHSCQNNQGY CFRGYCPTLA KQCITLWGSD AKVAPDECFQ NNTNGNEYDY CKKTNNVIIP 547
  ADAM (N. atra) LDELCTGQSA ECPMNHFHMN GHPCQNNQGY CFRGTCPTLT KQCIALWGPD AEVAPDGCFM NNQKGNYYGY CKKKNGTNIP 556
    Gn_ADAM iso3 LDELCTGQSA ECPMNHFHMD GYPCQNNQGY CFRGTCPTLT KQCIALWGPD AEVAPDGCFM NNQKGNDYGY CKKENGTNIP 556
    Gn_ADAM iso1 LPELCTGQSA ECPMDRFRIN GHPCQNNQGY CYMGKCPTLA GQCIALWGPG GKVAADSCFK QNQQGNYYGH CNT-NGAIIS 556
                                    580        600        620        640
                                      |          |          |          |
ADAM (O. hannah) CKPTDVKCGR LYCTGGTENP SEGEKISSDP CKASYS--EI EDIGMVDHRT KCGEKMVCSD GKCIPL*--- ---------- 612
       TrCr_ADAM CKPTDVKCGR LYCTGGTENP SEGEKISSDP CKASYS--EI EDIGMVDHRT KCGEKMVCSD GKCIPL*--- ---------- 612
    Gn_ADAM iso2 CKPTDVKCGR LYCTGGTENP SEGEKISSDP CKASYS--EI EDIGMVDHRT KCGEKMVCSD GKCIPL*--- ---------- 612
  ADAM (N. atra) CEPENVKCGR LYCIDDST-- ------EENS CKFHFSNENA NS-GMVQPGT KCGEGMVCGF GECIGLETAL GINQ*----- 622
    Gn_ADAM iso3 CEPXXXXX-- ---------- ---------- ---------- ---------- ---------- ---------- ---------- 564
    Gn_ADAM iso1 CKPNAVKCGR LYCTGGSKMP SDGNLLEFLS CRASFPSKDA EDVGLVHPGT KCGEGMVCNN GQCVEIETAY RSTNCSHKCT 636
                                    660
                                      |
ADAM (O. hannah) ---------- ---------- ---------- ---- 612
       TrCr_ADAM ---------- ---------- ---------- ---- 612
    Gn_ADAM iso2 ---------- ---------- ---------- ---- 612
  ADAM (N. atra) ---------- ---------- ---------- ---- 622
    Gn_ADAM iso3 ---------- ---------- ---------- ---- 564
    Gn_ADAM iso1 GHSVSILYSI FWLKYSPCPL VGFKTALLWA WPI* 670
```

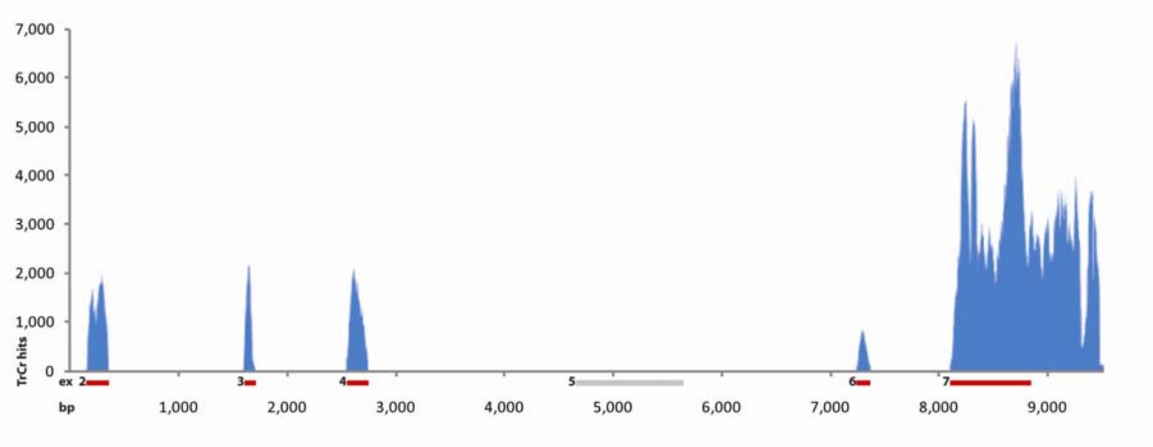
SOI Figure 6

Mapping of the transcriptome reads onto the two scaffolds containing two L-amino acid oxidase (LAAO) genes shows that only one of these genes is expressed in the venom gland. **a**) isoform 1; **b**) isoform 2; **c**) alignment of the two LAAO genes with reference sequences showing that our identified genes belong to the LAAO gene family. In *O. hannah* isoform 2 is expressed in the venom gland, while in *N. atra* isoform 1 appears to be expressed, since *N. atra* metalloproteinase is more similar to isoform 1 than isoform 2. Also see **Figure 4a** in the main text for further details.

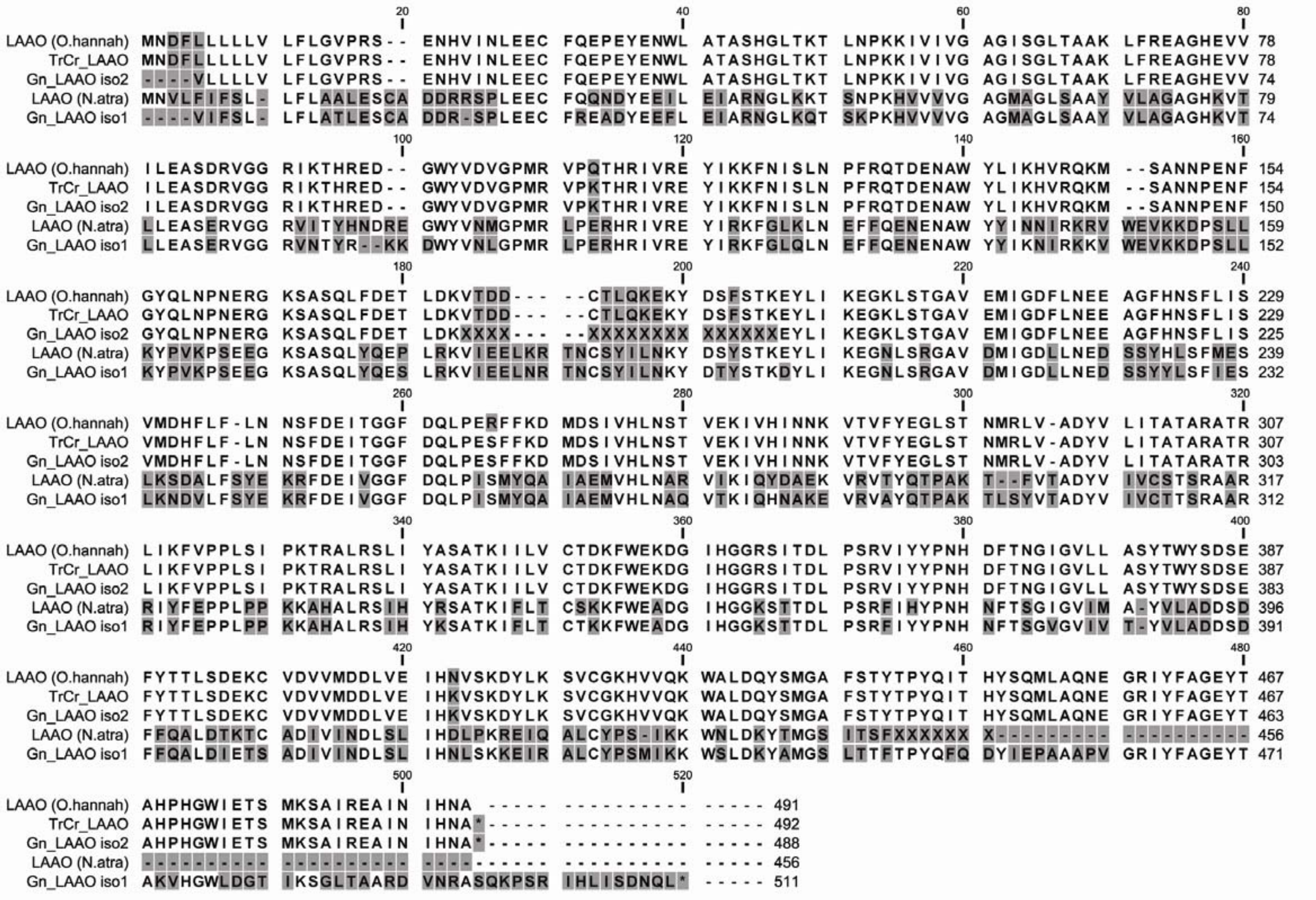
a



b



c

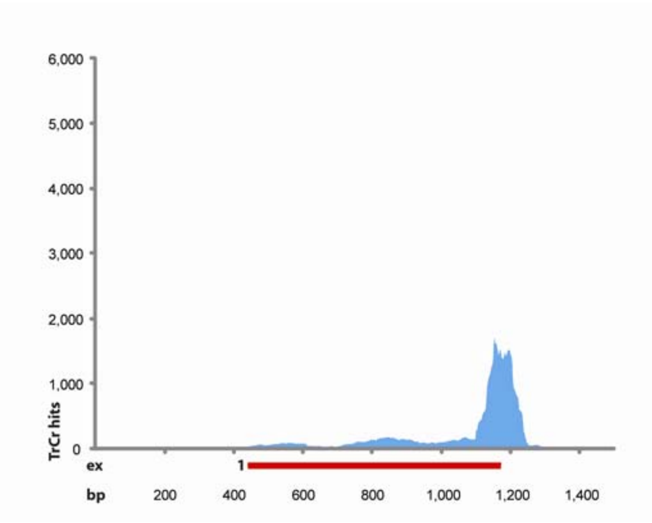


```
                                       20                                     40                                     60                                     80
                                        |                                      |                                      |                                      |
LAAO (O.hannah)  MNDFLLLLLV  LFLGVPRS--  ENHVINLEEC  FQEPEYENWL  ATASHGLTKT  LNPKKIVIVG  AGISGLTAAK  LFREAGHEVV  78
     TrCr_LAAO  MNDFLLLLLV  LFLGVPRS--  ENHVINLEEC  FQEPEYENWL  ATASHGLTKT  LNPKKIVIVG  AGISGLTAAK  LFREAGHEVV  78
  Gn_LAAO iso2  ----VLLLLV  LFLGVPRS--  ENHVINLEEC  FQEPEYENWL  ATASHGLTKT  LNPKKIVIVG  AGISGLTAAK  LFREAGHEVV  74
 LAAO (N.atra)  MNVLFIFSL-  LFLAALESCA  DDRRSPLEEC  FQQNDYEEIL  EIARNGLKKT  SNPKHVVVVG  AGMAGLSAAY  VLAGAGHKVT  79
  Gn_LAAO iso1  ----VIFSL-  LFLATLESCA  DDR-SPLEEC  FREADYEEFL  EIARNGLKQT  SKPKHVVVVG  AGMAGLSAAY  VLAGAGHKVT  74

                                      100                                    120                                    140                                    160
                                        |                                      |                                      |                                      |
LAAO (O.hannah)  ILEASDRVGG  RIKTHRED--  GWYVDVGPMR  VPQTHRIVRE  YIKKFNISLN  PFRQTDENAW  YLIKHVRQKM  --SANNPENF  154
     TrCr_LAAO  ILEASDRVGG  RIKTHRED--  GWYVDVGPMR  VPKTHRIVRE  YIKKFNISLN  PFRQTDENAW  YLIKHVRQKM  --SANNPENF  154
  Gn_LAAO iso2  ILEASDRVGG  RIKTHRED--  GWYVDVGPMR  VPKTHRIVRE  YIKKFNISLN  PFRQTDENAW  YLIKHVRQKM  --SANNPENF  150
 LAAO (N.atra)  LLEASERVGG  RVITYHNDRE  GWYVNMGPMR  LPERHRIVRE  YIRKFGLKLN  EFFQENENAW  YYINNIRKRV  WEVKKDPSLL  159
  Gn_LAAO iso1  LLEASERVGG  RVNTYR--KK  DWYVNLGPMR  LPERHRIVRE  YIRKFGLQLN  EFFQENENAW  YYIKNIRKKV  WEVKKDPSLL  152

                                      180                                    200                                    220                                    240
                                        |                                      |                                      |                                      |
LAAO (O.hannah)  GYQLNPNERG  KSASQLFDET  LDKVTDD---  --CTLQKEKY  DSFSTKEYLI  KEGKLSTGAV  EMIGDFLNEE  AGFHNSFLIS  229
     TrCr_LAAO  GYQLNPNERG  KSASQLFDET  LDKVTDD---  --CTLQKEKY  DSFSTKEYLI  KEGKLSTGAV  EMIGDFLNEE  AGFHNSFLIS  229
  Gn_LAAO iso2  GYQLNPNERG  KSASQLFDET  LDKXXXX---  --XXXXXXXX  XXXXXXEYLI  KEGKLSTGAV  EMIGDFLNEE  AGFHNSFLIS  225
 LAAO (N.atra)  KYPVKPSEEG  KSASQLYQEP  LRKVIEELKR  TNCSYILNKY  DSYSTKEYLI  KEGNLSRGAV  DMIGDLLNED  SSYHLSFMES  239
  Gn_LAAO iso1  KYPVKPSEEG  KSASQLYQES  LRKVIEELNR  TNCSYILNKY  DTYSTKDYLI  KEGNLSRGAV  DMIGDLLNED  SSYYLSFIES  232

                                      260                                    280                                    300                                    320
                                        |                                      |                                      |                                      |
LAAO (O.hannah)  VMDHFLF-LN  NSFDEITGGF  DQLPERFFKD  MDSIVHLNST  VEKIVHINNK  VTVFYEGLST  NMRLV-ADYV  LITATARATR  307
     TrCr_LAAO  VMDHFLF-LN  NSFDEITGGF  DQLPESFFKD  MDSIVHLNST  VEKIVHINNK  VTVFYEGLST  NMRLV-ADYV  LITATARATR  307
  Gn_LAAO iso2  VMDHFLF-LN  NSFDEITGGF  DQLPESFFKD  MDSIVHLNST  VEKIVHINNK  VTVFYEGLST  NMRLV-ADYV  LITATARATR  303
 LAAO (N.atra)  LKSDALFSYE  KRFDEIVGGF  DQLPISMYQA  IAEMVHLNAR  VIKIQYDAEK  VRVTYQTPAK  T--FVTADYV  IVCSTSRAAR  317
  Gn_LAAO iso1  LKNDVLFSYE  KRFDEIVGGF  DQLPISMYQA  IAEMVHLNAQ  VTKIQHNAKE  VRVAYQTPAK  TLSYVTADYV  IVCTTSRAAR  312

                                      340                                    360                                    380                                    400
                                        |                                      |                                      |                                      |
LAAO (O.hannah)  LIKFVPPLSI  PKTRALRSLI  YASATKIILV  CTDKFWEKDG  IHGGRSITDL  PSRVIYYPNH  DFTNGIGVLL  ASYTWYSDSE  387
     TrCr_LAAO  LIKFVPPLSI  PKTRALRSLI  YASATKIILV  CTDKFWEKDG  IHGGRSITDL  PSRVIYYPNH  DFTNGIGVLL  ASYTWYSDSE  387
  Gn_LAAO iso2  LIKFVPPLSI  PKTRALRSLI  YASATKIILV  CTDKFWEKDG  IHGGRSITDL  PSRVIYYPNH  DFTNGIGVLL  ASYTWYSDSE  383
 LAAO (N.atra)  RIYFEPPLPP  KKAHALRSIH  YRSATKIFLT  CSKKFWEADG  IHGGKSTTDL  PSRFIHYPNH  NFTSGIGVIM  A-YVLADDSD  396
  Gn_LAAO iso1  RIYFEPPLPP  KKAHALRSIH  YKSATKIFLT  CTKKFWEADG  IHGGKSTTDL  PSRFIYYPNH  NFTSGVGVIV  T-YVLADDSD  391

                                      420                                    440                                    460                                    480
                                        |                                      |                                      |                                      |
LAAO (O.hannah)  FYTTLSDEKC  VDVVMDDLVE  IHNVSKDYLK  SVCGKHVVQK  WALDQYSMGA  FSTYTPYQIT  HYSQMLAQNE  GRIYFAGEYT  467
     TrCr_LAAO  FYTTLSDEKC  VDVVMDDLVE  IHKVSKDYLK  SVCGKHVVQK  WALDQYSMGA  FSTYTPYQIT  HYSQMLAQNE  GRIYFAGEYT  467
  Gn_LAAO iso2  FYTTLSDEKC  VDVVMDDLVE  IHKVSKDYLK  SVCGKHVVQK  WALDQYSMGA  FSTYTPYQIT  HYSQMLAQNE  GRIYFAGEYT  463
 LAAO (N.atra)  FFQALDTKTC  ADIVINDLSL  IHDLPKREIQ  ALCYPS-IKK  WNLDKYTMGS  ITSFXXXXXX  X---------  ----------  456
  Gn_LAAO iso1  FFQALDIETS  ADIVINDLSL  IHNLSKKEIR  ALCYPSMIKK  WSLDKYAMGS  LTTFTPYQFQ  DYIEPAAAPV  GRIYFAGEYT  471

                                      500                                    520
                                        |                                      |
LAAO (O.hannah)  AHPHGWIETS  MKSAIREAIN  IHNA------  ----------  -----  491
     TrCr_LAAO  AHPHGWIETS  MKSAIREAIN  IHNA*-----  ----------  -----  492
  Gn_LAAO iso2  AHPHGWIETS  MKSAIREAIN  IHNA*-----  ----------  -----  488
 LAAO (N.atra)  ----------  ----------  ----------  ----------  -----  456
  Gn_LAAO iso1  AKVHGWLDGT  IKSGLTAARD  VNRASQKPSR  IHLISDNQL*  -----  511
```

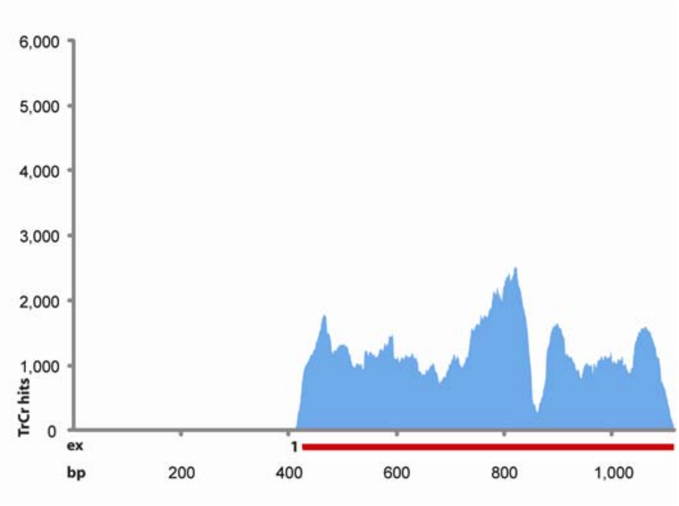
SOI Figure 7

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.

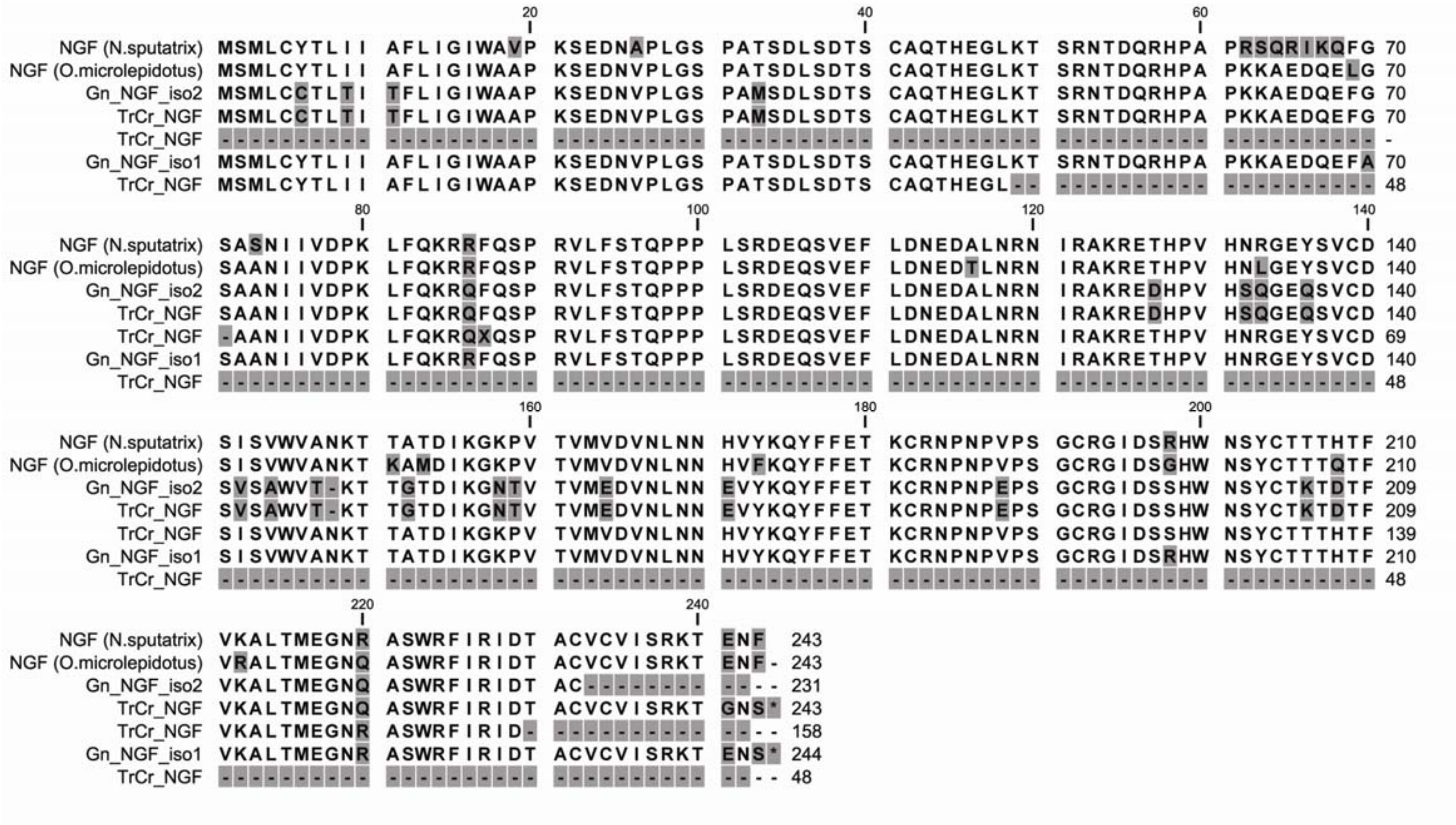
a



b



c

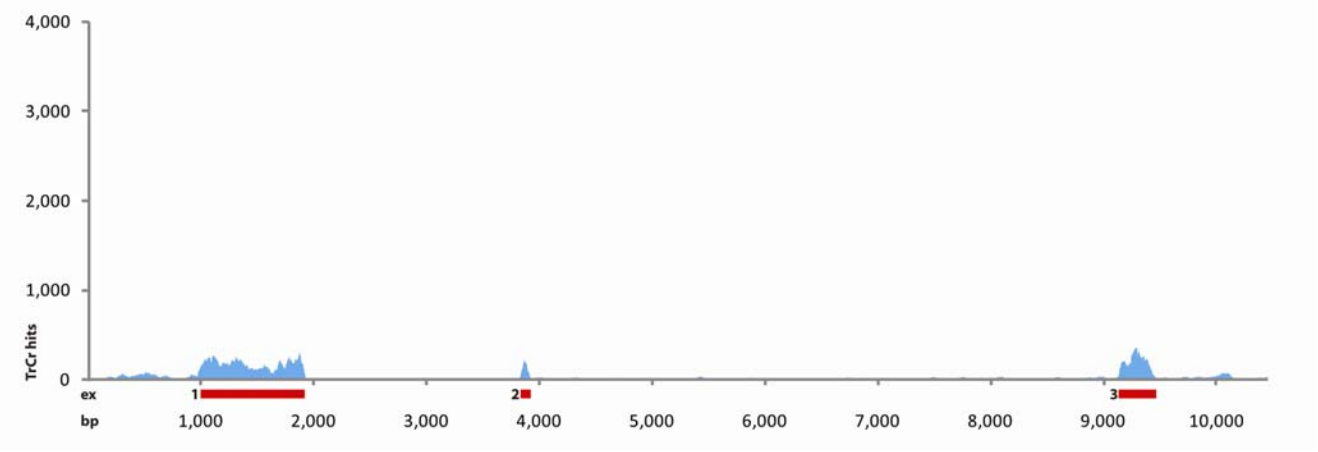


```
                                           20                                    40                                    60
                                           |                                     |                                     |
       NGF (N.sputatrix) MSMLCYTLII AFLIGIWAVP KSEDNAPLGS PATSDLSDTS CAQTHEGLKT SRNTDQRHPA PRSQRIKQFG 70
NGF (O.microlepidotus)   MSMLCYTLII AFLIGIWAAP KSEDNVPLGS PATSDLSDTS CAQTHEGLKT SRNTDQRHPA PKKAEDQELG 70
            Gn_NGF_iso2  MSMLCCTLTI TFLIGIWAAP KSEDNVPLGS PAMSDLSDTS CAQTHEGLKT SRNTDQRHPA PKKAEDQEFG 70
              TrCr_NGF   MSMLCCTLTI TFLIGIWAAP KSEDNVPLGS PAMSDLSDTS CAQTHEGLKT SRNTDQRHPA PKKAEDQEFG 70
              TrCr_NGF   ---------- ---------- ---------- ---------- ---------- ---------- ---------- -
            Gn_NGF_iso1  MSMLCYTLII AFLIGIWAAP KSEDNVPLGS PATSDLSDTS CAQTHEGLKT SRNTDQRHPA PKKAEDQEFA 70
              TrCr_NGF   MSMLCYTLII AFLIGIWAAP KSEDNVPLGS PATSDLSDTS CAQTHEGL-- ---------- ---------- 48
                                  80                   100                   120                   140
                                  |                     |                     |                     |
       NGF (N.sputatrix) SASNIIVDPK LFQKRRFQSP RVLFSTQPPP LSRDEQSVEF LDNEDALNRN IRAKRETHPV HNRGEYSVCD 140
NGF (O.microlepidotus)   SAANIIVDPK LFQKRRFQSP RVLFSTQPPP LSRDEQSVEF LDNEDTLNRN IRAKRETHPV HNLGEYSVCD 140
            Gn_NGF_iso2  SAANIIVDPK LFQKRQFQSP RVLFSTQPPP LSRDEQSVEF LDNEDALNRN IRAKREDHPV HSQGEQSVCD 140
              TrCr_NGF   SAANIIVDPK LFQKRQFQSP RVLFSTQPPP LSRDEQSVEF LDNEDALNRN IRAKREDHPV HSQGEQSVCD 140
              TrCr_NGF   -AANIIVDPK LFQKRQXQSP RVLFSTQPPP LSRDEQSVEF LDNEDALNRN IRAKRETHPV HNRGEYSVCD 69
            Gn_NGF_iso1  SAANIIVDPK LFQKRRFQSP RVLFSTQPPP LSRDEQSVEF LDNEDALNRN IRAKRETHPV HNRGEYSVCD 140
              TrCr_NGF   ---------- ---------- ---------- ---------- ---------- ---------- ---------- 48
                                             160                   180                   200
                                             |                     |                     |
       NGF (N.sputatrix) SISVWVANKT TATDIKGKPV TVMVDVNLNN HVYKQYFFET KCRNPNPVPS GCRGIDSRHW NSYCTTTHTF 210
NGF (O.microlepidotus)   SISVWVANKT KAMDIKGKPV TVMVDVNLNN HVFKQYFFET KCRNPNPVPS GCRGIDSGHW NSYCTTTQTF 210
            Gn_NGF_iso2  SVSAWVT-KT TGTDIKGNTV TVMEDVNLNN EVYKQYFFET KCRNPNPEPS GCRGIDSSHW NSYCTKTDTF 209
              TrCr_NGF   SVSAWVT-KT TGTDIKGNTV TVMEDVNLNN EVYKQYFFET KCRNPNPEPS GCRGIDSSHW NSYCTKTDTF 209
              TrCr_NGF   SISVWVANKT TATDIKGKPV TVMVDVNLNN HVYKQYFFET KCRNPNPVPS GCRGIDSSHW NSYCTTTHTF 139
            Gn_NGF_iso1  SISVWVANKT TATDIKGKPV TVMVDVNLNN HVYKQYFFET KCRNPNPVPS GCRGIDSRHW NSYCTTTHTF 210
              TrCr_NGF   ---------- ---------- ---------- ---------- ---------- ---------- ---------- 48
                                220                   240
                                |                     |
       NGF (N.sputatrix) VKALTMEGNR ASWRFIRIDT ACVCVISRKT ENF  243
NGF (O.microlepidotus)   VRALTMEGNQ ASWRFIRIDT ACVCVISRKT ENF- 243
            Gn_NGF_iso2  VKALTMEGNQ ASWRFIRIDT AC-------- ---- 231
              TrCr_NGF   VKALTMEGNQ ASWRFIRIDT ACVCVISRKT GNS* 243
              TrCr_NGF   VKALTMEGNR ASWRFIRID- ---------- ---- 158
            Gn_NGF_iso1  VKALTMEGNR ASWRFIRIDT ACVCVISRKT ENS* 244
              TrCr_NGF   ---------- ---------- ---------- ---- 48
```

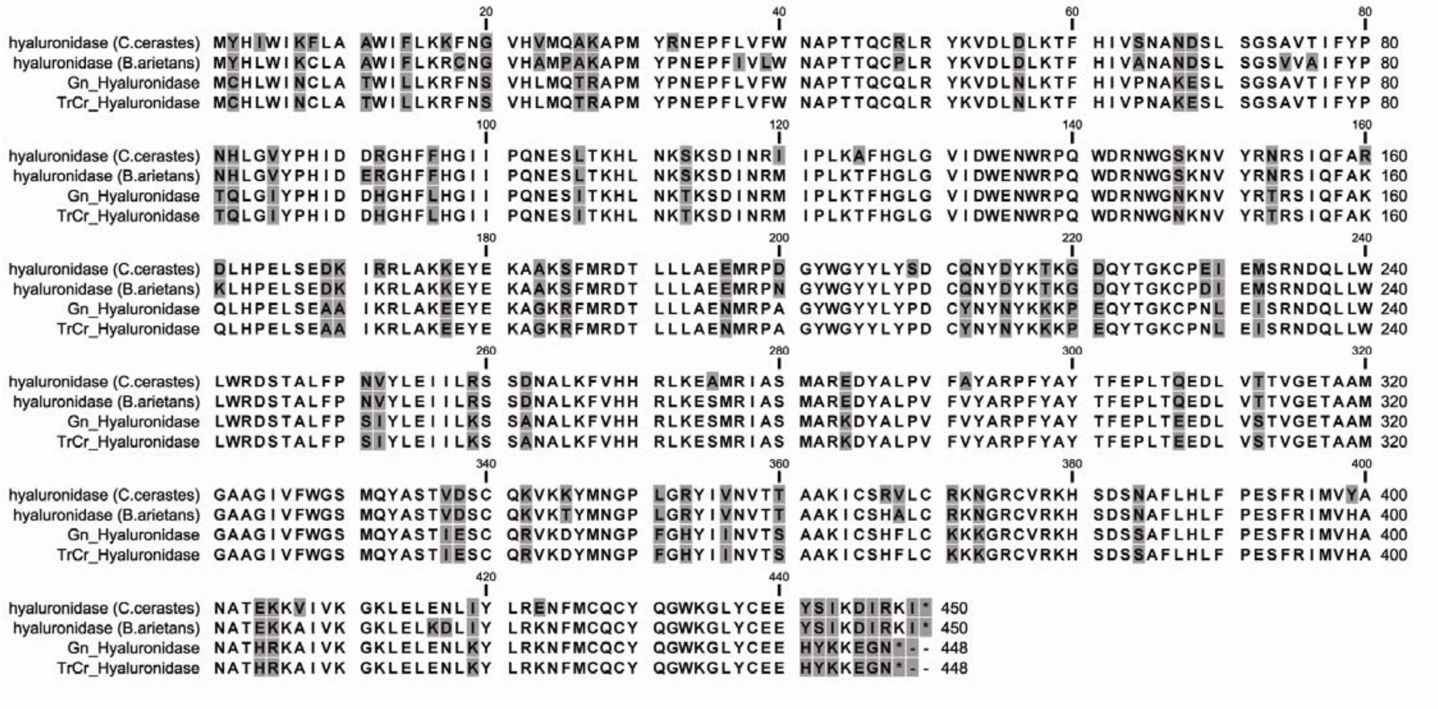
SOI Figure 8

a) mapping of the transcriptome reads onto the scaffolds containing the HYA gene shows that this gene is expressed in the venom gland; **b**) alignment of the HYA gene with reference sequences showing that our identified genes belong to the HYA gene family.

a



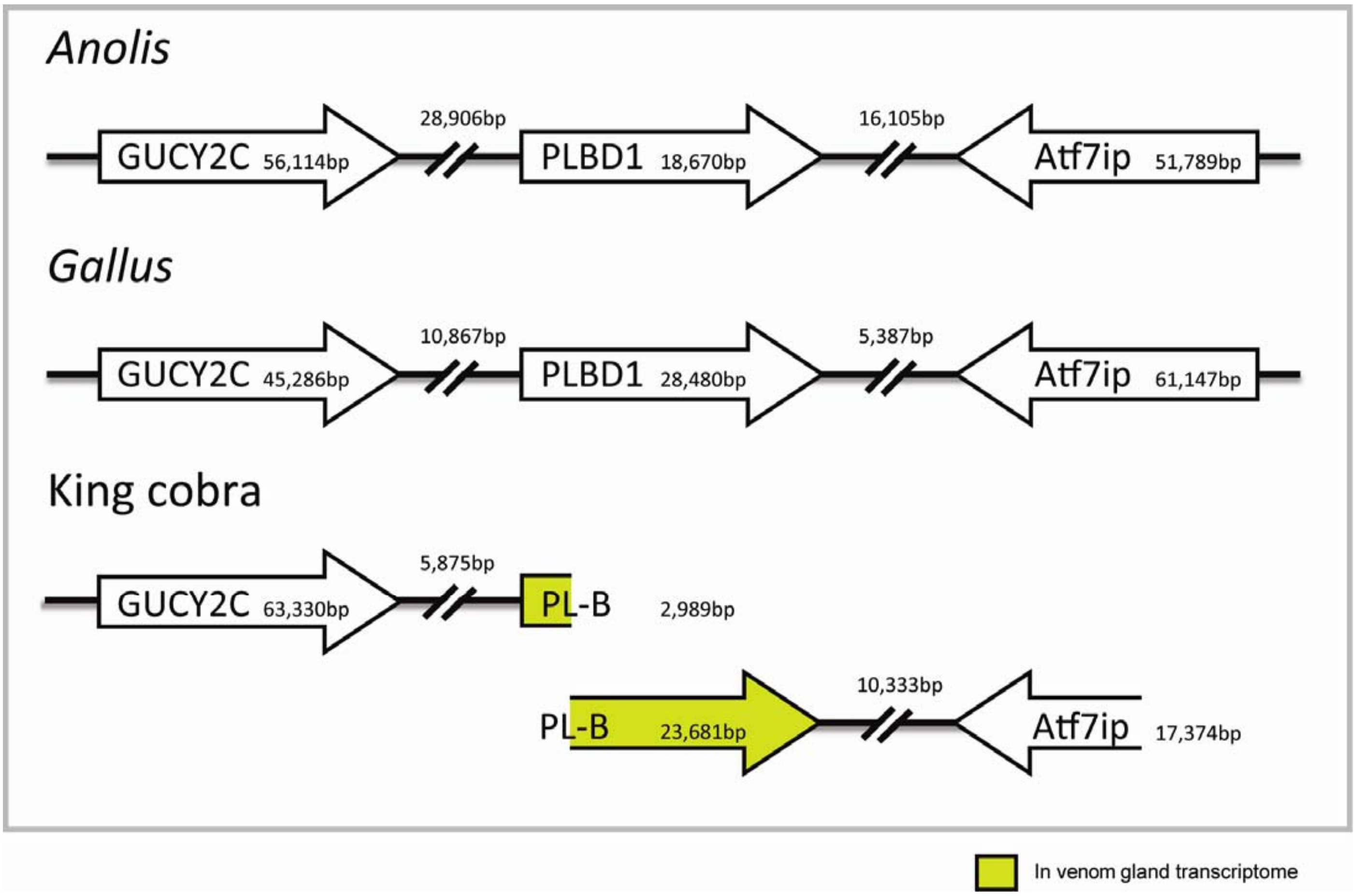
b



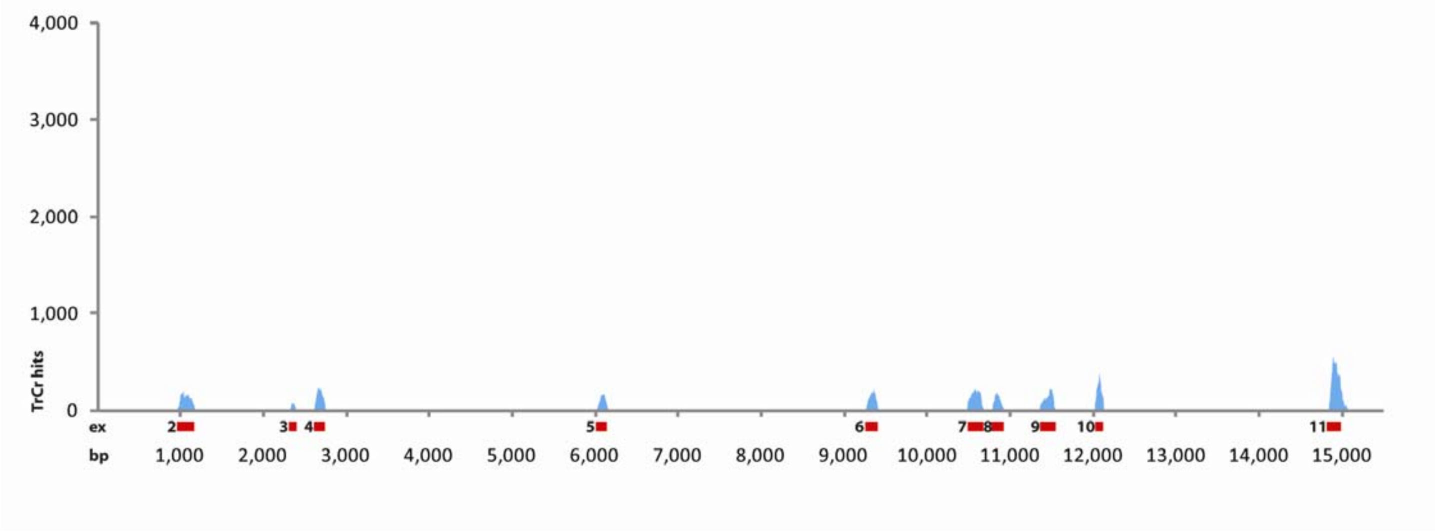
```
                                             20                                           40                                           60                                           80
hyaluronidase (C.cerastes)  MYHIWIKFLA  AWIFLKKFNG  VHVMQAKAPM  YRNEPFLVFW  NAPTTQCRLR  YKVDLDLKTF  HIVSNANDSL  SGSAVTIFYP  80
hyaluronidase (B.arietans)  MYHLWIKCLA  AWIFLKRCNG  VHAMPAKAPM  YPNEPFIVLW  NAPTTQCPLR  YKVDLDLKTF  HIVANANDSL  SGSVVAIFYP  80
        Gn_Hyaluronidase  MCHLWINCLA  TWILLKRFNS  VHLMQTRAPM  YPNEPFLVFW  NAPTTQCQLR  YKVDLNLKTF  HIVPNAKESL  SGSAVTIFYP  80
      TrCr_Hyaluronidase  MCHLWINCLA  TWILLKRFNS  VHLMQTRAPM  YPNEPFLVFW  NAPTTQCQLR  YKVDLNLKTF  HIVPNAKESL  SGSAVTIFYP  80
                                            100                                          120                                          140                                          160
hyaluronidase (C.cerastes)  NHLGVYPHID  DRGHFFHGII  PQNESLTKHL  NKSKSDINRI  IPLKAFHGLG  VIDWENWRPQ  WDRNWGSKNV  YRNRSIQFAR  160
hyaluronidase (B.arietans)  NHLGVYPHID  ERGHFFHGII  PQNESLTKHL  NKSKSDINRM  IPLKTFHGLG  VIDWENWRPQ  WDRNWGSKNV  YRNRSIQFAK  160
        Gn_Hyaluronidase  TQLGIYPHID  DHGHFLHGII  PQNESITKHL  NKTKSDINRM  IPLKTFHGLG  VIDWENWRPQ  WDRNWGNKNV  YRTRSIQFAK  160
      TrCr_Hyaluronidase  TQLGIYPHID  DHGHFLHGII  PQNESITKHL  NKTKSDINRM  IPLKTFHGLG  VIDWENWRPQ  WDRNWGNKNV  YRTRSIQFAK  160
                                            180                                          200                                          220                                          240
hyaluronidase (C.cerastes)  DLHPELSEDK  IRRLAKKEYE  KAAKSFMRDT  LLLAEEMRPD  GYWGYYLYSD  CQNYDYKTKG  DQYTGKCPEI  EMSRNDQLLW  240
hyaluronidase (B.arietans)  KLHPELSEDK  IKRLAKKEYE  KAAKSFMRDT  LLLAEEMRPN  GYWGYYLYPD  CQNYDYKTKG  DQYTGKCPDI  EMSRNDQLLW  240
        Gn_Hyaluronidase  QLHPELSEAA  IKRLAKEEYE  KAGKRFMRDT  LLLAENMRPA  GYWGYYLYPD  CYNYNYKKKP  EQYTGKCPNL  EISRNDQLLW  240
      TrCr_Hyaluronidase  QLHPELSEAA  IKRLAKEEYE  KAGKRFMRDT  LLLAENMRPA  GYWGYYLYPD  CYNYNYKKKP  EQYTGKCPNL  EISRNDQLLW  240
                                            260                                          280                                          300                                          320
hyaluronidase (C.cerastes)  LWRDSTALFP  NVYLEIILRS  SDNALKFVHH  RLKEAMRIAS  MAREDYALPV  FAYARPFYAY  TFEPLTQEDL  VTTVGETAAM  320
hyaluronidase (B.arietans)  LWRDSTALFP  NVYLEIILRS  SDNALKFVHH  RLKESMRIAS  MAREDYALPV  FVYARPFYAY  TFEPLTQEDL  VTTVGETAAM  320
        Gn_Hyaluronidase  LWRDSTALFP  SIYLEIILKS  SANALKFVHH  RLKESMRIAS  MARKDYALPV  FVYARPFYAY  TFEPLTEEDL  VSTVGETAAM  320
      TrCr_Hyaluronidase  LWRDSTALFP  SIYLEIILKS  SANALKFVHH  RLKESMRIAS  MARKDYALPV  FVYARPFYAY  TFEPLTEEDL  VSTVGETAAM  320
                                            340                                          360                                          380                                          400
hyaluronidase (C.cerastes)  GAAGIVFWGS  MQYASTVDSC  QKVKKYMNGP  LGRYIVNVTT  AAKICSRVLC  RKNGRCVRKH  SDSNAFLHLF  PESFRIMVYA  400
hyaluronidase (B.arietans)  GAAGIVFWGS  MQYASTVDSC  QKVKTYMNGP  LGRYIVNVTT  AAKICSHALC  RKNGRCVRKH  SDSNAFLHLF  PESFRIMVHA  400
        Gn_Hyaluronidase  GAAGIVFWGS  MQYASTIESC  QRVKDYMNGP  FGHYIINVTS  AAKICSHFLC  KKKGRCVRKH  SDSSAFLHLF  PESFRIMVHA  400
      TrCr_Hyaluronidase  GAAGIVFWGS  MQYASTIESC  QRVKDYMNGP  FGHYIINVTS  AAKICSHFLC  KKKGRCVRKH  SDSSAFLHLF  PESFRIMVHA  400
                                            420                                          440
hyaluronidase (C.cerastes)  NATEKKVIVK  GKLELENLIY  LRENFMCQCY  QGWKGLYCEE  YSIKDIRKI*  450
hyaluronidase (B.arietans)  NATEKKAIVK  GKLELKDLIY  LRKNFMCQCY  QGWKGLYCEE  YSIKDIRKI*  450
        Gn_Hyaluronidase  NATHRKAIVK  GKLELENLKY  LRKNFMCQCY  QGWKGLYCEE  HYKKEGN*--  448
      TrCr_Hyaluronidase  NATHRKAIVK  GKLELENLKY  LRKNFMCQCY  QGWKGLYCEE  HYKKEGN*--  448
```

SOI Figure 9

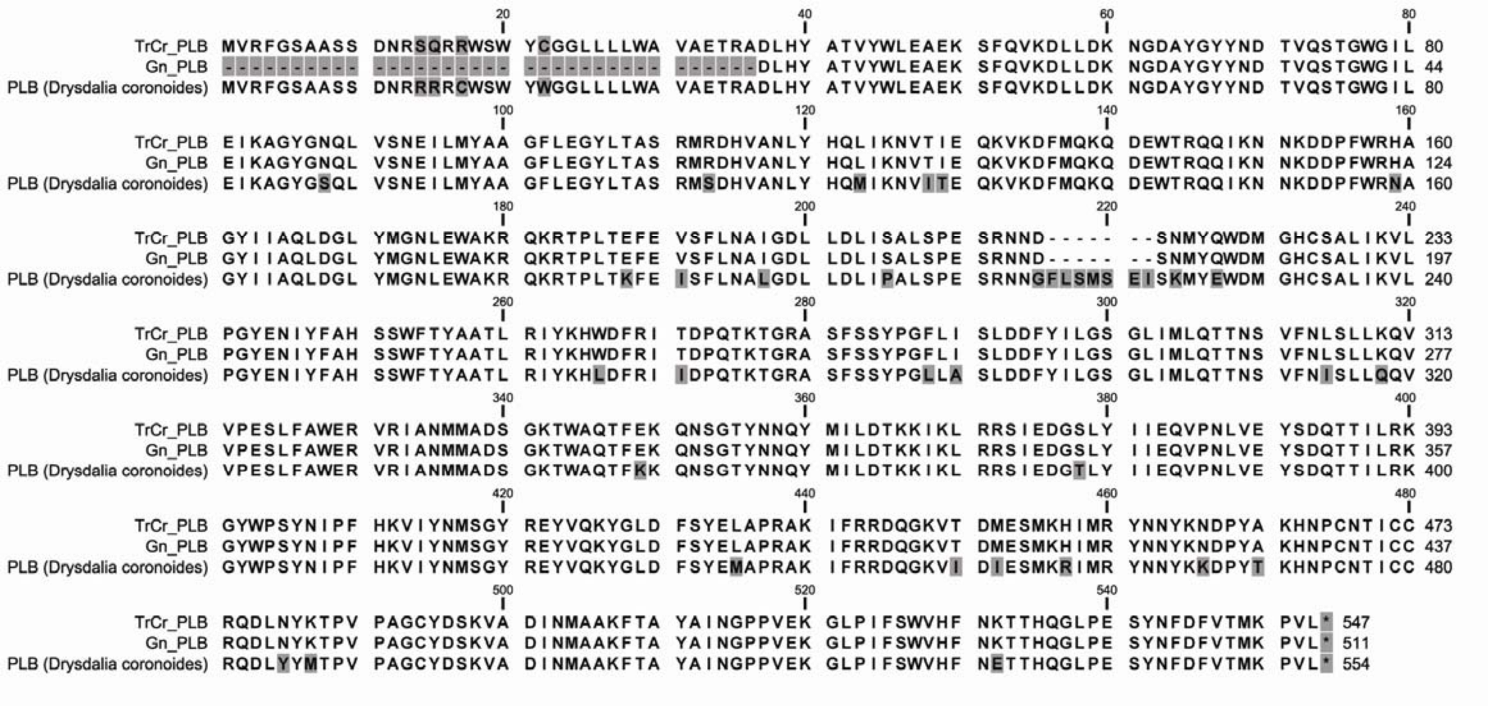
a) scheme of the genomic synteny of the PL-B genes in the *Anolis*, *Gallus* and king cobra. *Anolis*, *Gallus* genomic sequences from www.ensembl.org; **b**) Mapping of the transcriptome reads onto one scaffolds containing the PL-B gene shows that this gene is expressed in the venom gland; **c**) alignment of the PL-B gene with reference sequences showing that our identified genes belong to the PL-B gene family.



b



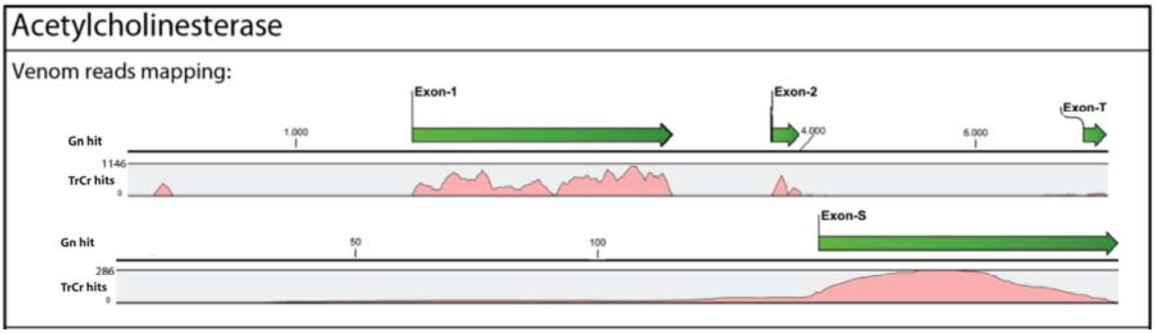
c



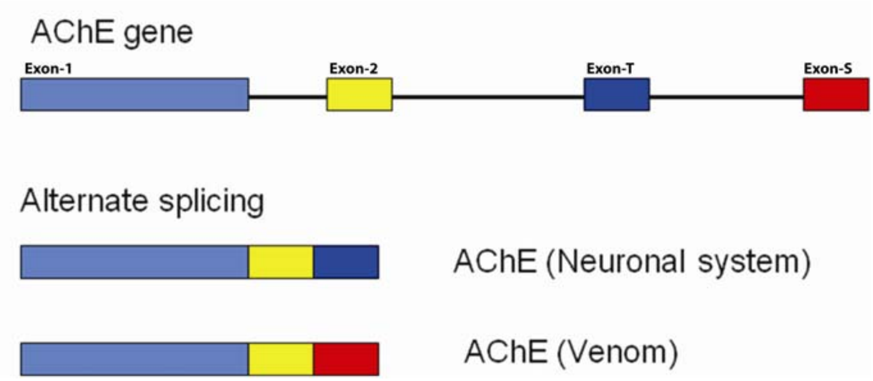
```
                                          20                    40                    60                    80
                 TrCr_PLB MVRFGSAASS DNRSQRRWSW YCGGLLLLWA VAETRADLHY ATVYWLEAEK SFQVKDLLDK NGDAYGYYND TVQSTGWGIL 80
                   Gn_PLB ---------- ---------- ---------- ------DLHY ATVYWLEAEK SFQVKDLLDK NGDAYGYYND TVQSTGWGIL 44
PLB (Drysdalia coronoides) MVRFGSAASS DNRRRRCWSW YWGGLLLLWA VAETRADLHY ATVYWLEAEK SFQVKDLLDK NGDAYGYYND TVQSTGWGIL 80
                                         100                   120                   140                   160
                 TrCr_PLB EIKAGYGNQL VSNEILMYAA GFLEGYLTAS RMRDHVANLY HQLIKNVTIE QKVKDFMQKQ DEWTRQQIKN NKDDPFWRHA 160
                   Gn_PLB EIKAGYGNQL VSNEILMYAA GFLEGYLTAS RMRDHVANLY HQLIKNVTIE QKVKDFMQKQ DEWTRQQIKN NKDDPFWRHA 124
PLB (Drysdalia coronoides) EIKAGYGSQL VSNEILMYAA GFLEGYLTAS RMSDHVANLY HQMIKNVITE QKVKDFMQKQ DEWTRQQIKN NKDDPFWRNA 160
                                         180                   200                   220                   240
                 TrCr_PLB GYIIAQLDGL YMGNLEWAKR QKRTPLTEFE VSFLNAIGDL LDLISALSPE SRNND----- --SNMYQWDM GHCSALIKVL 233
                   Gn_PLB GYIIAQLDGL YMGNLEWAKR QKRTPLTEFE VSFLNAIGDL LDLISALSPE SRNND----- --SNMYQWDM GHCSALIKVL 197
PLB (Drysdalia coronoides) GYIIAQLDGL YMGNLEWAKR QKRTPLTKFE ISFLNALGDL LDLIPALSPE SRNNGFLSMS EISKMYEWDM GHCSALIKVL 240
                                         260                   280                   300                   320
                 TrCr_PLB PGYENIYFAH SSWFTYAATL RIYKHWDFRI TDPQTKTGRA SFSSYPGFLI SLDDFYILGS GLIMLQTTNS VFNLSLLKQV 313
                   Gn_PLB PGYENIYFAH SSWFTYAATL RIYKHWDFRI TDPQTKTGRA SFSSYPGFLI SLDDFYILGS GLIMLQTTNS VFNLSLLKQV 277
PLB (Drysdalia coronoides) PGYENIYFAH SSWFTYAATL RIYKHLDFRI IDPQTKTGRA SFSSYPGLLA SLDDFYILGS GLIMLQTTNS VFNISLLQQV 320
                                         340                   360                   380                   400
                 TrCr_PLB VPESLFAWER VRIANMMADS GKTWAQTFEK QNSGTYNNQY MILDTKKIKL RRSIEDGSLY IIEQVPNLVE YSDQTTILRK 393
                   Gn_PLB VPESLFAWER VRIANMMADS GKTWAQTFEK QNSGTYNNQY MILDTKKIKL RRSIEDGSLY IIEQVPNLVE YSDQTTILRK 357
PLB (Drysdalia coronoides) VPESLFAWER VRIANMMADS GKTWAQTFKK QNSGTYNNQY MILDTKKIKL RRSIEDGTLY IIEQVPNLVE YSDQTTILRK 400
                                         420                   440                   460                   480
                 TrCr_PLB GYWPSYNIPF HKVIYNMSGY REYVQKYGLD FSYELAPRAK IFRRDQGKVT DMESMKHIMR YNNYKNDPYA KHNPCNTICC 473
                   Gn_PLB GYWPSYNIPF HKVIYNMSGY REYVQKYGLD FSYELAPRAK IFRRDQGKVT DMESMKHIMR YNNYKNDPYA KHNPCNTICC 437
PLB (Drysdalia coronoides) GYWPSYNIPF HKVIYNMSGY REYVQKYGLD FSYEMAPRAK IFRRDQGKVI DIESMKRIMR YNNYKKDPYT KHNPCNTICC 480
                                         500                   520                   540
                 TrCr_PLB RQDLNYKTPV PAGCYDSKVA DINMAAKFTA YAINGPPVEK GLPIFSWVHF NKTTHQGLPE SYNFDFVTMK PVL* 547
                   Gn_PLB RQDLNYKTPV PAGCYDSKVA DINMAAKFTA YAINGPPVEK GLPIFSWVHF NKTTHQGLPE SYNFDFVTMK PVL* 511
PLB (Drysdalia coronoides) RQDLYYMTPV PAGCYDSKVA DINMAAKFTA YAINGPPVEK GLPIFSWVHF NETTHQGLPE SYNFDFVTMK PVL* 554
```

SOI Figure 10

The scaffolds containing the acetylcholinesterase gene. The gene consists of exon-1, exon-2 and two exons that are alternatively spliced for both the neuronal AChE (exon-T) and the venom AChE (exon-s). **a**) the mapping of the venom transcriptome reads onto the two scaffolds shows that exon-s is expressed in the venom gland, but exon-t is not; **b**) a diagram showing how AChE is alternatively spliced in the neuronal form and the venom form (from).



b



METHODS

King cobra tissue acquisition and processing

All animal procedures were approved by the local ethics committee. 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.

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 SOI Table 1). 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 SOI Table 1). After the first gel purification the fragment length was analyzed by 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.

mRNA-Seq 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. The RNA used for the mixed tissue mRNA-Seq library was obtained by mixing of 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.

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 libraries were single-read sequenced with a read length of 51 nucleotides. Image analysis and base calling were done by the Illumina pipeline.

Genome assembly strategy

In assembling the King cobra genome, we largely followed the strategy pioneered by Li *et al.* for the assembly of the giant panda genome(*10*). 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. 7% 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, 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. Briefly, SSPACE aligns paired reads to the contigs (using Bowtie), 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 infividual 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 225511 bp.

Genome annotation strategy and mRNA-Seq analysis

To predict genes on the scaffolds we used AUGUSTUS (version 2.4). To make prediction more accurate hint files were constructed from the available transcriptome data using BLAT and the scripts provided with AUGUSTUS. The output of AUGUSTUS was used to annotate the scaffolds. For subsequent manual annotation of selected genes, transcriptome reads were aligned and quantified using the CLC bio Genomics Workbench (version 4).

Mitochondrial phylogeny

To reconstruct the phylogenetic relationship of the family Elapidae to which the king cobra belongs, seven elapid mitochondrial genomes available from Genbank were gathered, as well as the mtDNA sequence of *Agkistrodon piscivorus*, a member of the related family Viperidae, as an outgroup (summarized in SOI table 2).

The mitochondrial genome of the King cobra under study was identified in the final scaffolds by BLAST search. Most snake mtDNA genomes contain a duplication of the control region, hence this scaffold (16215 bp) does not directly correspond to the complete mtDNA genome: the control region is essentially a ~1 Kb repeat that cannot be resolved using our general assembly and scaffolding strategy. Therefore, the Velvet *de novo* assembler was used to reassemble all reads aligning (using Bowtie) to either this scaffold or to a published elapid snake mtDNA genome. Based on this assembly, a 17263 bp circular genome was reconstructed, which was annotated using results from a tRNAscan-SE server and based on homology with the genomes listed in SOI table 2.

All mitochondrial genomes under consideration contain 13 protein coding genes, which were aligned at the amino acid level using the CLC bio Genomics Workbench. The alignment was manually checked and ambiguous regions were removed.; based on this amino acid alignment an alignment at the nucleic acid level was produced. RAxML was used to construct a maximum likelihood (ML) phylogenetic tree based on 11268 sites using a GTR + Γ model, with all parameters estimated independently for all genes and codon positions by the algorithm. Statistical support of branches was evaluated by 1000 ML bootstrap replicates. Monophyly of each genus was supported by 98–100% bootstrap probability, whereas the intergenic relationships in the family Elapidae were not fully resolved.

SOI TABLES

SOI Table 1. Sequencing libraries.

Library name	Library type	Insert size (bp)1	Read length	Raw sequence	Clean sequence2	Scaffolding links2
PE200	Paired-ends	60–157	2×76 nt	21.9 Gbp	16.8 Gbp	n.a.
PE500	Paired-ends	122–478	2×50 nt	8.5 Gbp	7.9 Gbp	4.3 M
			2×151 nt	10.8 Gbp	9.8 Gbp	
MP2K	Mate pair	1600–2400	2×36 nt	5.4 Gbp	n.a.	3.4 M
MP7K	Mate pair	2500–6000	2×51 nt	2.3 Gbp	n.a.	181 K
MP10K	Mate pair	6500–10000	2×51 nt	5.3 Gbp	n.a.	1.4 M
MP15K	Mate pair	9000–13000	2×51 nt	3.8 Gbp	n.a.	1.2 M
Venom	mRNA-Seq	n.a.	51 nt	0.83 Gbp	n.a.	n.a.
Pooled organs	mRNA-Seq	n.a.	51 nt	0.91 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 distribution are based on inspection of a histogram.

2. Clean sequence is filtered for adapter sequences and low quality nucleotides, and preassembled (see text). Scaffolding links are pairs of which both reads align to different initial contigs at unique positions. n.a., not applicable.

SOI Table 2. Mitochondrial genomes used in phylogeny reconstruction

Species	**Family**	**Common name**	**Accession**	**Length**	**Reference**
Agkistrodon piscivorus	Viperidae	Cottonmouth	NC_009768	17213 bp	(*26*)
Bungarus fasciatus	Elapidae	Banded krait	NC_011393	17234 bp	(*27*)
Bungarus multicinctus	Elapidae	Taiwanese banded krait	NC_011392	17144 bp	(*27*)
Micrurus fulvius	Elapidae	Eastern coral snake	NC_013481	17506 bp	(*28*)
Naja naja	Elapidae	Indian cobra	NC_010225	17213 bp	(*29*)
Naja atra (1)	Elapidae	Chinese cobra	NC_011389	17216 bp	(*27*)
Naja atra (2)	Elapidae	Chinese cobra	EU921898	17214 bp	(*27*)
Ophiophagus hannah	Elapidae	King cobra (China)	NC_011394	17267 bp	(*27*)
Ophiophagus hannah	Elapidae	King cobra (Indonesia)	-	17263 bp	This study