ARTICLE

Discoveries

Title:

Restriction and recruitment – gene duplication and the origin and evolution of snake venom toxins

Authors:

Adam D Hargreaves[1], Martin T Swain[2], Matthew J Hegarty[2], Darren W Logan[3] and John F Mulley[1]*

Affiliations:

1. School of Biological Sciences, Bangor University, Brambell Building, Deiniol Road, Bangor, Gwynedd, LL57 2UW, United Kingdom

2. Institute of Biological, Environmental & Rural Sciences, Aberystwyth University, Penglais, Aberystwyth, Ceredigion, SY23 3DA, United Kingdom

3. Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, United Kingdom

***Corresponding author:**

Dr John Mulley

School of Biological Sciences, Bangor University, Deiniol Road, Bangor, Gwynedd LL57 2UW, United Kingdom

Tel: +44 (0)1248 383492, Email: j.mulley@bangor.ac.uk

Abstract

The genetic and genomic mechanisms underlying evolutionary innovations are of fundamental importance to our understanding of animal evolution. Snake venom represents one such innovation and has been hypothesised to have originated and diversified via a process that involves duplication of genes encoding body proteins and subsequent recruitment of the copy to the venom gland where natural selection can act to develop or increase toxicity. However, gene duplication is known to be a rare event in vertebrate genomes and the recruitment of duplicated genes to a novel expression domain (neofunctionalisation) is an even rarer process that requires the evolution of novel combinations of transcription factor binding sites in upstream regulatory regions. This hypothesis concerning the evolution of snake venom is therefore very unlikely. Nonetheless, it is often assumed to be established fact and this has hampered research into the true origins of snake venom toxins. We have generated transcriptomic data for a diversity of body tissues and salivary and venom glands from venomous and non-venomous reptiles, which has allowed us to critically evaluate this hypothesis. Our comparative transcriptomic analysis of venom and salivary glands and body tissues in five species of reptile reveals that snake venom does not evolve via the hypothesised process of duplication and recruitment of body proteins. Indeed, our results show that many proposed venom toxins are in fact expressed in a wide variety of body tissues, including the salivary gland of non-venomous reptiles and have therefore been restricted to the venom gland following duplication, not recruited. Thus snake venom evolves via the duplication and subfunctionalisation of genes encoding existing salivary proteins. These results highlight the danger of the "just-so story" in evolutionary biology, where an elegant and intuitive idea is repeated so often that it assumes the mantle of established fact, to the detriment of the field as a whole.

Introduction

Gene duplication is a rare event in eukaryotic genomes and has been suggested to be the major source of novel genetic material (Ohno 1970). Estimates of the rate of gene duplication in vertebrates vary from 1 gene per 100 to 1 gene per 1000 per million years (Lynch and Conery 2000; Lynch and Conery 2003; Cotton and Page 2005), and the most common fate for a duplicate gene is the loss of its function (nonfunctionalisation, pseudogenisation (Mighell et al. 2000; Presgraves 2005)). However, in some cases a duplicate gene is retained in the population and undergoes either subfunctionalisation (where the two duplicates divide the sum of the ancestral role(s) between them) or neofunctionalisation (where one of the duplicates assumes a new role, independent of the ancestral function (Force et al. 1999)). This latter process of evolving an entirely new function is known to be incredibly rare and there are few conclusive examples of it in the literature (Escriva et al. 2006; Van Damme et al. 2007; Deng et al. 2010).

The venom of advanced snakes has been hypothesised to have originated and diversified via gene duplication (Wong and Belov 2012). In particular, it has been suggested that both the origin of venom and the later evolution of novelty in venom has occurred as a result of the duplication of a gene encoding a non-venom physiological or “body” protein that is subsequently recruited, via gene regulatory changes, into the venom gland, where natural selection can act on randomly occurring mutations to develop and/or increase toxicity (Lynch 2007; Fry et al. 2009b; Kwong et al. 2009; Casewell et al. 2012; Fry et al. 2012a; Casewell et al. 2013; Margres et al. 2013; Vonk et al. 2013). In short, it has been proposed that snake venom diversifies via repeated gene duplication and neofunctionalisation, a somewhat surprising finding given the apparent rarity of both of these events (here we refer to neofunctionalisation with respect to the acquisition of novel sites of expression at the level of individual tissues, not the acquisition of novel functions at a molecular level, which is

separate from the claims of the duplication/recruitment hypothesis and has been shown to have occurred for only a small number of venom toxins (Kini 2002; Kini 2003; Lynch 2007; Kini and Doley 2010), whilst the majority of duplicated toxins retain ancestral bioactivity (Fry 2005; Warrell 2010)). However, there are currently several gaps in our knowledge of how this remarkable process might take place, including the mechanisms underlying repeated gene duplications and, more importantly, the gene regulatory changes that occur to facilitate "recruitment" into the venom gland. Given that whole genome duplication is a rare event in vertebrates in general and reptiles in particular (Otto and Whitton 2000; Mable 2004), it seems likely that the majority of snake venom toxin genes are duplicated via segmental duplication (Hurles 2004), where the highly repetitive nature of reptile genomes (Shedlock et al. 2007; Di-Poi et al. 2009) provides regions of pseudo-homology that facilitate unequal crossing-over during homologous recombination, producing tandemly-arranged duplicates. This process requires neither germ-line expression nor the evolution of *de novo cis*-regulatory sequences as does retrotransposition (Zhang 2003) and, if repeated so that the resulting pairs or larger clusters of genes were subsequently duplicated in the same manner, a relatively small number of duplication events could give rise to a large number of duplicate genes. Evidence for clusters of multiple SVMP, CRISP and lectin genes in the king cobra genome (Vonk et al. 2013) and for PLA_2 genes in the Okinawan habu (*Protobothrops* (now *Trimeresurus*) *flavoviridis*) (Ikeda et al. 2010) would seem to support this hypothesis, although more complete data from these and other snake whole genome sequencing projects is needed.

Whilst the above scenario explains the apparent ease with which *existing* venom toxin genes might be repeatedly duplicated along with their associated *cis*-regulatory architecture, it does nothing to explain how a non-venom gene might be "recruited" into the venom gland. The paralogous genes produced as a result of gene duplication are 100% identical and, if the

entirety of their associated *cis*-regulatory architecture has also been duplicated along with them, they will have identical temporal and spatial expression patterns (i.e. they are functionally redundant (Force et al. 1999; Lynch and Force 2000)). Therefore in order to develop a novel site of expression such as in the venom gland, a novel combination of transcriptional regulatory sequences must arise.

Eukaryotic transcription factor binding sites are the result of a trade-off between the specificity offered by longer stretches of DNA and the robustness to mutation offered by shorter sequences and vary in length between 5 and >30nt, with an average length of 10nt (Stewart et al. 2012). It has been estimated that eukaryotic promoters may contain 10-50 binding sites for 5-15 different transcription factors (Wray et al. 2003). The rarity of gene duplication, coupled with the low likelihood of evolving new combinations of transcription factor binding sites before the duplicated gene is nonfunctionalised by random mutations in coding sequences should therefore make the process of duplication and recruitment of genes encoding physiological or body proteins into the venom gland exceedingly rare. How then do we reconcile this with the apparent widespread occurrence of just this process in the origin and evolution of snake venom? One possible alternative hypothesis is that many of the genes expressed in snake venom are in fact the result of the duplication of genes that were ancestrally expressed in multiple tissues, including the venom gland. Following duplication these genes therefore evolved via subfunctionalisation, with one copy's expression being restricted to the venom gland and the other maintaining the original, multi-tissue expression pattern (possibly with subsequent loss of expression of this paralog in the venom gland). This scenario of duplication and restriction, rather than duplication and recruitment (Figure 1) is more parsimonious as it requires only the loss of transcription factor binding sites, which may occur by random mutation of single base pairs or larger insertions or deletions (indels) that may delete or disrupt the existing transcriptional regulatory sequences. In order to

differentiate between the two hypotheses gene expression data from non-venom gland tissues in venomous and non-venomous species are needed, something which has until now been missing. Here we review the existing evidence for the duplication and recruitment of genes into the venom gland and carry out a comparative transcriptomic survey of gene expression in the venom glands and body tissues of a number of reptile species, including the painted saw-scaled viper (*Echis coloratus*), a venomous snake; the corn snake (*Pantherophis guttatus*) and rough green snake (*Opheodrys aestivus*), both non-venomous colubrids which use constriction to kill prey (Kardong 2002); the royal python (*Python regius*) a non-venomous boid and the leopard gecko (*Eublepharis macularius*), a member of one of the most basal lineages of squamate reptiles. The phylogenetic position of this latter species is particularly important, as it lies outside of the proposed clade of ancestrally venomous reptiles (the Toxicofera (Fry et al. 2006; Fry et al. 2009a; Fry et al. 2012b; Fry et al. 2013)) and therefore genes found in the salivary gland of this species can be taken to represent the ancestral squamate expression pattern. We also take advantage of available transcriptomic resources for body tissues in a number of other reptile species, including king cobra (*Ophiophagus hannah*) venom gland, accessory gland and pooled tissues (heart, lung, spleen, brain, testes, gall bladder, pancreas, small intestine, kidney, liver, eye, tongue and stomach) (Vonk et al. 2013), garter snake (*Thamnophis elegans*) liver (Schwartz and Bronikowski 2013) and pooled tissue (brain, gonads, heart, kidney, liver, spleen and blood of males and females) (Schwartz et al. 2010), Burmese python (*Python molurus bivittatus*) pooled heart and liver (Castoe et al. 2011) and corn snake brain (Tzika et al. 2011).

Results and Discussion

We find the hypothesis that snake venom evolves via the duplication of physiological or body genes and subsequent recruitment into the venom gland to be unsupported by the available

data – in short, snake venom has not evolved via the recruitment of "body" genes. Indeed for a large number of the gene families claimed to have undergone recruitment we find evidence of a diverse tissue expression pattern, including the salivary gland of non-venomous reptiles (Figure 2), demonstrating that, if they do encode toxic venom components (Hargreaves *et al.* in prep), they have not been recruited into the venom gland, but restricted to it. The recently published king cobra genome paper (Vonk et al. 2013) also provides evidence for salivary (rictal) gland expression of several venom toxins in the Burmese python, *Python molurus bivittatus*, including 3ftx, cystatin, hyaluronidase and SVMP (Supplementary Table S2 in (Vonk et al. 2013)).

Therefore whilst some venom toxin genes have in the past been suggested to represent ancestral salivary proteins (notably cysteine-rich secretory proteins (CRISPs)) and Kallikrein-like serine proteases (Fry 2005; Sunagar et al. 2012), our analysis in fact shows that the majority of snake venom toxins are likely derived from pre-existing salivary proteins. Far from being an incredibly complex cocktail of proteins (Kini 2002; Wagstaff et al. 2006; Fox and Serrano 2008; Casewell et al. 2013) recruited from multiple body tissues (Fry 2005; Fry et al. 2009a; Warrell 2010; Casewell et al. 2013), snake venom should instead be considered to be simply a modified form of saliva, where a relatively small number of gene families (typically 6-14) have expanded via gene duplication, often in a lineage-specific manner (Kulkeaw et al. 2007; Wagstaff et al. 2009; Fahmi et al. 2012; Vonk et al. 2013).

The study cited most frequently in support of the duplication and recruitment hypothesis is that of Fry (Fry 2005) (see for example (Warrell 2010; Jiang et al. 2011; Casewell et al. 2012; Casewell et al. 2013)) and we therefore refer to this hypothesis as the 'genome to venome hypothesis'. In his study, Fry concluded that the evolution of snake venom was characterised by at least 24 recruitment events (Fry 2005). However, this analysis was based on assumptions that snake venom toxin sequences derived primarily from EST-based studies of

only the venom gland could be considered to be venom gland-specific and that if they were related to a gene known to be expressed in the pancreas (or another tissue) of human or other species they must therefore represent a recruitment event. It is obviously possible that the same gene may be expressed in the pancreas (or other tissue) of the snake as well and that the lack of data for these non-venom gland tissues is obscuring the true extent of their expression. It must be considered therefore that for the majority of genes Fry does not actually demonstrate any evidence for gene duplication and subsequent recruitment.

Only four examples in Fry's study include both "body" and venom gland sequences from venomous snakes and therefore only these four possibly show any evidence in support of gene duplication and recruitment into the venom gland: crotamine; complement C3; natriuretic peptide and Group IB phospholipase A_2 (Fry 2005). Of these, the South American rattlesnake (*Crotalus durissus terrificus*) *crotamine*-like sequence labelled as 'Pancreas' (accession number Q6HAA2) was in fact originally described to be highly expressed in pancreas, heart, liver, brain and kidneys (i.e. all tissues examined) with "scarce" but detectable expression in the venom gland (Rádis-Baptista et al. 2004). Our transcriptomic data shows that the toxic form of *crotamine* is derived from the duplication of a non-toxic *β-defensin*-like gene with a wider expression pattern that included the salivary/venom gland (Figure 2) and that the toxic duplicate has been restricted, not recruited, to the venom gland.

For *complement C3*, Fry's analysis utilised Indian cobra (*Naja naja*) sequences from liver (accession number Q01833) (Fritzinger et al. 1992) and venom gland (accession number Q91132) (Fritzinger et al. 1994). However, both sequences were in fact isolated from what the authors refer to as "*Naja naja kaouthia*", a synonym for the monocled cobra, *Naja kaouthia*. This inaccuracy notwithstanding, Fry's analysis does suggest that there has been a duplication of a *complement C3* gene to give rise to a new copy (often referred to as "*cobra venom factor*", more rightly called *complement C3b*) although the lack of data for other body

tissues should have precluded claims of recruitment. Analysis of our transcriptome data in fact reveals that *complement C3* is expressed in a diverse array of body tissues in multiple species, including the salivary gland of non-venomous reptiles (Figures 2 and 3) and that a paralogous copy of this gene has therefore been restricted to the venom gland following duplication. Whilst *Bothrops jararaca* does appear to possess at least two distinct forms of natriuretic peptide (Hayashi et al. 2003; Hayashi and Camargo 2005), the situation may also be more complex than that originally presented, as the sequence labelled as 'Brain' by Fry (accession Q9PW56, identical to AAD51326) in fact shows a wider expression pattern that includes brain, spleen, venom gland and, possibly, pancreas (Murayama et al. 1997; Hayashi et al. 2003; Hayashi and Camargo 2005). We find few natriuretic peptides in our dataset (Figure 2), and the low number of these sequences previously characterised would suggest that they play little role in the venom of non-*Bothrops* snakes, where they appear to have undergone duplication and subfunctionalisation. Finally, Fry used *Group IB phospholipase* A_2 (*PLA_2 IB*) sequences from the pancreas of the banded sea krait (*Laticauda semifasciata*, accession Q8JFG2) and the venom gland of the Australian coastal taipan (*Oxyuranus scutellatus*, accession P00615) to support recruitment. We find *PLA_2 IB* genes to be expressed in several body tissues, including the leopard gecko salivary gland (Figure 2 and Supplementary figure 1), suggesting a wider ancestral expression pattern than previously claimed.

It has recently been suggested that there has been a duplication of *nerve growth factor* (*ngf*) genes in some species of snake (Sunagar et al. 2013), although the presence of additional copies of *ngf* in certain species of cobra has been known for some time (Lipps 2000; Koh et al. 2004). Our data show that the non-toxic form of *ngf* (which we call *ngfa*) is expressed in a diversity of tissues, including the salivary glands of non-venomous reptiles (Figure 2 and

Supplementary figure 2). The putatively toxic version (*ngfb*) has therefore also been restricted to the venom gland following duplication.

Both coagulation *factor V* and *factor X* have been suggested to have undergone gene duplication in Australian elapids such as *Tropidechis carinatus* and *Pseudonaja textilis* with subsequent recruitment of a gene normally expressed in the liver into the venom gland (Le et al. 2005; Reza et al. 2007; Kwong et al. 2009; Kwong and Kini 2011). However, these studies do not appear to have investigated body tissues other than liver and venom gland (Le et al. 2005) and so cannot be relied upon to demonstrate the full extent of ancestral gene expression. Our analysis in fact shows *factor V* to be expressed in multiple tissues, including rough green snake scent gland, King cobra accessory gland, *Echis coloratus* scent gland, kidney, brain, ovary and skin and the scent gland, skin and salivary gland of the leopard gecko (Figure 2 and Supplementary figure 3). *Factor X* is also expressed in multiple tissues (Figure 2 and Supplementary figure 4), including the salivary or venom glands of leopard gecko, royal python, rough green snake, corn snake and *Echis coloratus*. In both cases therefore a gene with a wide expression pattern that included the salivary or venom gland has undergone duplication and restriction. The known increased expression of a *factor X* paralog following an insertion in the promoter region (Reza et al. 2007; Kwong and Kini; Kwong et al. 2009; Han et al. 2013) and the increased expression of *crotamine* in the venom gland following duplication (Rádis-Baptista et al. 2003; Rádis-Baptista et al. 2004) suggest that a possible route for pre-existing salivary proteins to become venom toxins may simply be an elevated expression level, where initial toxicity is dosage-dependent.

Interestingly, some of the key papers cited in support of the genome to venome hypothesis in fact discuss the recruitment of genes into the venom proteome, *not* the venom gland itself (Fry and Wuster 2004; Fry 2005) with such claims only becoming more common in the literature some time later (see for example (Fry et al. 2008; Durban et al. 2011; Casewell et

al. 2013)). Added to the fact that these papers show no evidence for duplication and recruitment of "body" genes it must be concluded that not only is this hypothesis not supported by our newly available data, but that it was never supported. It appears therefore that a misunderstanding of the scope of the claims of these earlier studies, together with the known role for gene duplication in the *diversification* of snake venom (Kordiš and Gubenšek 2000) is responsible for the development and propagation of the attractive, but ultimately unsupported, duplication and venom gland recruitment hypothesis. In order to fully understand the evolution of snake venom, more transcriptomic data is needed from a much greater variety of species for a much greater number of body tissues, ideally at a wider diversity of stages of venom synthesis and with consideration of sex, ontogeny, shedding and reproductive cycles and the large-scale effects on metabolism of intermittent feeding on large prey (Wall et al. 2011; Castoe et al. 2013). Even so, it will be difficult to fully account for all possible spatial and temporal influences on gene expression, and the default assumption for the fate of duplicate genes should perhaps therefore be subfunctionalisation, not neofunctionalisation.

Finally, our findings highlight the problem of 'just-so stories' (Kipling 1902) in evolutionary biology, especially when they reach the point of being considered established fact. The genome to venome hypothesis has been widely and unquestioningly cited and treated neither as a hypothesis to be tested and refuted (Popper 1959), nor as a scientific research programme to provide predictions to be investigated (Lakatos 1980). Whilst the role of gene duplication should rightly be considered as part of the core of the snake venom evolution research programme, we propose that many associated hypotheses are in need of a greater degree of scrutiny than they have hitherto received. Only after such scrutiny will we truly understand "How The Snake Got His Venom".

Materials and Methods

Total RNA was extracted from the salivary glands, scent glands and skin of two adult corn snakes (*Pantherophis guttatus*), rough green snakes (*Opheodrys aestivus*), royal pythons (*Python regius*) and leopard geckos (*Eublepharis macularius*). Only a single corn snake skin sample provided RNA of high enough quality for sequencing. RNA samples for painted saw-scaled vipers (*Echis coloratus*) were extracted from the skin, scent glands, kidney and brain of two adult specimens, and liver and ovary samples were extracted from one adult individual. Venom glands from four adult individuals were taken at different time points following venom extraction (16, 24 and 48 hours post-milking) in order to capture the full diversity of venom genes. All RNA extractions were carried out using the RNeasy mini kit (Qiagen) with on-column DNase digestion. mRNA was prepared for sequencing using the TruSeq RNA sample preparation kit (Illumina) with a selected fragment size of 200-500bp and sequenced using 100bp paired-end reads on the Illumina HiSeq2000 or HiSeq2500 platform. The quality of all raw sequence data was assessed using FastQC (Andrews 2010) and reads for each tissue pooled and assembled using Trinity (Grabherr et al. 2011) (sequence and assembly metrics are provided in Supplementary tables S1 and S2). Venom genes were identified by BLAST (Camacho et al. 2009) and maximum-likelihood-based phylogenetic analysis and tissue distribution identified by BLAST-based searches of assembled transcriptomes.

Transcriptome reads were deposited in the European Nucleotide Archive (ENA) database under accession #ERP001222 and the Sequence Read Archive (SRA) under the study accession #SRP042007.

References

Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. BMC Bioinformatics. 10:421-2105-10-421.

Casewell NR, Huttley GA, Wüster W. 2012. Dynamic evolution of venom proteins in squamate reptiles. Nature Commun. 3:1066.

Casewell NR, Wüster W, Vonk FJ, Harrison RA, Fry BG. 2013. Complex cocktails: The evolutionary novelty of venoms. Trends Ecol Evol. 28:219-229.

Castoe TA, Fox SE, de Koning APJ, Poole AW, Daza JM, Smith EN, Mockler TC, Secor SM, Pollock DD. 2011. A multi-organ transcriptome resource for the Burmese python (*Python molurus bivittatus*). BMC Res Notes. 4:310.

Castoe TA, de Koning APJ, Hall KT, Card DC, Schield DR, Fujita MK, Ruggiero RP, Degner JF, Daza JM, Gu W et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proc Natl Acad Sci U S A 110:20645-20650.

Cotton JA, Page RD. 2005. Rates and patterns of gene duplication and loss in the human genome. Proc Biol Sci. 272:277-283.

Deng C, Cheng CH, Ye H, He X, Chen L. 2010. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proc Natl Acad Sci U S A. 107:21593-21598.

Di-Poi N, Montoya-Burgos JI, Duboule D. 2009. Atypical relaxation of structural constraints in Hox gene clusters of the green anole lizard. Genome Res. 19:602-610.

Durban J, Juárez P, Angulo Y, Lomonte B, Flores-Diaz M, Alape-Girón A, Sasa M, Sanz L, Gutiérrez JM, Dopazo J. 2011. Profiling the venom gland transcriptomes of Costa Rican snakes by 454 pyrosequencing. BMC Genomics. 12:259.

Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, Cartry J, Duffraisse M, Holland L, Gronemeyer H, Laudet V. 2006. Neofunctionalization in vertebrates: The example of retinoic acid receptors. PLOS Genetics. 2:e102.

Fahmi L, Makran B, Pla D, Sanz L, Oukkache N, Lkhider M, Harrison RA, Ghalim N, Calvete JJ. 2012. Venomics and antivenomics profiles of North African *Cerastes cerastes* and *C. vipera* populations reveals a potentially important therapeutic weakness. J Proteomics. 75:2442-2453.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151:1531-1545.

Fox JW, Serrano SM. 2008. Exploring snake venom proteomes: Multifaceted analyses for complex toxin mixtures. Proteomics. 8:909-920.

Fritzinger DC, Petrella EC, Connelly MB, Bredehorst R, Vogel CW. 1992. Primary structure of cobra complement component C3. J Immunol. 149:3554-3562.

Fritzinger DC, Bredehorst R, Vogel CW. 1994. Molecular cloning and derived primary structure of cobra venom factor. Proc Natl Acad Sci U S A. 91:12775-12779.

Fry BG, Wüster W. 2004. Assembling an arsenal: Origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences. Mol Biol Evol. 21:870-883.

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.

Fry BG, Vidal N, Norman JA, Vonk FJ, Scheib H, Ramjan SR, Kuruppu S, Fung K, Hedges SB, Richardson MK. 2006. Early evolution of the venom system in lizards and snakes. Nature. 439:584-588.

Fry BG, Scheib H, van der Weerd L, Young B, McNaughtan J, Ramjan SF, Vidal N, Poelmann RE, Norman JA. 2008. Evolution of an arsenal: Structural and functional diversification of the venom system in the advanced snakes (Caenophidia). Mol Cell Proteomics. 7:215-246.

Fry BG, Vidal N, Van der Weerd L, Kochva E, Renjifo C. 2009a. Evolution and diversification of the Toxicofera reptile venom system. J Proteomics. 72:127-136.

Fry BG, Roelants K, Champagne DE, Scheib H, Tyndall JD, King GF, Nevalainen TJ, Norman JA, Lewis RJ, Norton RS. 2009b. The toxicogenomic multiverse: Convergent recruitment of proteins into animal venoms. Annu Rev Genomics Hum Genet. 10:483-511.

Fry BG, Scheib H, Junqueira de Azevedo ILM, Silva DA, Casewell NR. 2012a. Novel transcripts in the maxillary venom glands of advanced snakes. Toxicon. 59:696-708.

Fry BG, Casewell NR, Wüster W, Vidal N, Young B, Jackson TN. 2012b. The structural and functional diversification of the Toxicofera reptile venom system. Toxicon. 60:434-448.

Fry BG, Undheim EA, Ali SA, Debono J, Scheib H, Ruder T, Jackson TN, Morgenstern D, Cadwallader L, Whitehead D. 2013. Squeezers and leaf-cutters: Differential diversification and degeneration of the venom system in Toxicoferan reptiles. Mol Cell Proteomics. 12: 1881-1899

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 29:644-652.

Han X, Kwong S, Ge R, Kolatkar P, Kini RM. 2013. Transcriptional regulation of Trocarin D, a prothrombin activator from *Tropidechis carinatus*. FASEB Journal. 27:550.6.

Hayashi MA, Murbach AF, Ianzer D, Portaro FC, Prezoto BC, Fernandes BL, Silveira PF, Silva CA, Pires RS, Britto LR. 2003. The C-type natriuretic peptide precursor of snake brain contains highly specific inhibitors of the angiotensin-converting enzyme. J Neurochem. 85:969-977.

Hayashi MA, Camargo A. 2005. The bradykinin-potentiating peptides from venom gland and brain of *Bothrops jararaca* contain highly site specific inhibitors of the somatic angiotensin-converting enzyme. Toxicon. 45:1163-1170.

Hurles M. 2004. Gene duplication: The genomic trade in spare parts. PLOS Biology. 2:e206.

Ikeda N, Chijiwa T, Matsubara K, Oda-Ueda N, Hattori S, Matsuda Y, Ohno M. 2010. Unique structural characteristics and evolution of a cluster of venom phospholipase A_2 isozyme genes of *Protobothrops flavoviridis* snake. Gene. 461:15-25.

Jiang Y, Li Y, Lee W, Xu X, Zhang Y, Zhao R, Zhang Y, Wang W. 2011. Venom gland transcriptomes of two elapid snakes (*Bungarus multicinctus* and *Naja atra*) and evolution of toxin genes. BMC Genomics. 12:1.

Kardong KV. 2002. Colubrid snakes and duvernoy's “venom” glands. Toxin Rev. 21:1-19.

Kini RM. 2002. Molecular moulds with multiple missions: Functional sites in three- finger toxins. Clin Exp Pharmacol Physiol. 29:815-822.

Kini RM. 2003. Excitement ahead: Structure, function and mechanism of snake venom phospholipase A_2 enzymes. Toxicon. 42:827-840.

Kini RM, Doley R. 2010. Structure, function and evolution of three-finger toxins: Mini proteins with multiple targets. Toxicon. 56:855-867.

Koh D, Armugam A, Jeyaseelan K. 2004. Sputa nerve growth factor forms a preferable substitute to mouse 7S-beta nerve growth factor. Biochem J. 383:149-158.

Kordiš D, Gubenšek F. 2000. Adaptive evolution of animal toxin multigene families. Gene. 261:43-52.

Kulkeaw K, Chaicumpa W, Sakolvaree Y, Tongtawe P, Tapchaisri P. 2007. Proteome and immunome of the venom of the Thai cobra, *Naja kaouthia*. Toxicon. 49:1026-1041.

Kwong S, Woods AE, Mirtschin PJ, Ge R, Kini RM. 2009. The Recruitment of Blood Coagulation Factor X into Snake Venom Gland as a Toxin: The Role of Promoter *Cis*-Elements in its Expression. Thromb Haemost. 102/3:469-478

Kwong S, Kini RM. 2011. Duplication of coagulation factor genes and evolution of snake venom prothrombin activators. In: Friedberg F, editor. Gene Duplication. InTech. ISBN:978-953-307-387-3.

Lakatos I. 1980. The methodology of scientific research programmes: Volume 1: Philosophical papers. Cambridge: Cambridge University press.

Le TNM, Reza A, Swarup S, Kini RM. 2005. Gene duplication of coagulation factor V and origin of venom prothrombin activator in *Pseudonaja textilis* snake. Thromb Haemost. 93:420.

Lipps BV. 2000. Isolation of nerve growth factor (NGF) from human body fluids; saliva, serum and urine: Comparison between cobra venom and cobra serum NGF. J Nat Toxins. 9:349-356.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151-1155.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics. 154:459-473.

Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. J Struct Funct Genomics. 3:35-44.

Lynch VJ. 2007. Inventing an arsenal: Adaptive evolution and neofunctionalization of snake venom phospholipase A_2 genes. BMC Evol Biol. 7:2.

Mable B. 2004. 'Why polyploidy is rarer in animals than in plants': Myths and mechanisms. Biol J Linn Soc. 82:453-466.

Margres MJ, Aronow K, Loyacano J, Rokyta DR. 2013. The venom-gland transcriptome of the eastern coral snake (*Micrurus fulvius*) reveals high venom complexity in the intragenomic evolution of venoms. BMC Genomics. 14:1-18.

Mighell A, Smith N, Robinson P, Markham A. 2000. Vertebrate pseudogenes. FEBS Lett. 468:109-114.

Murayama N, Hayashi MA, Ohi H, Ferreira LA, Hermann VV, Saito H, Fujita Y, Higuchi S, Fernandes BL, Yamane T et al. 1997. Cloning and sequence analysis of a *Bothrops jararaca* cDNA encoding a precursor of seven bradykinin-potentiating peptides and a C-type natriuretic peptide. Proc Natl Acad Sci U S A. 94:1189-1193.

Ohno S. 1970. Evolution by gene duplication. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. Annu Rev Genet. 34:401-437.

Popper KR. 1959. The logic of scientific discovery. New York: Routledge.

Presgraves DC. 2005. Evolutionary genomics: New genes for new jobs. Curr Biol. 15:R52-R53.

Rádis-Baptista G, Kubo T, Oguiura N, Svartman M, Almeida T, Batistic RF, Oliveira EB, Vianna-Morgante ÂM, Yamane T. 2003. Structure and chromosomal localization of the gene for crotamine, a toxin from the South American rattlesnake, *Crotalus durissus terrificus*. Toxicon. 42:747-752.

Rádis-Baptista G, Kubo T, Oguiura N, Prieto da Silva A, Hayashi M, Oliveira E, Yamane T. 2004. Identification of crotasin, a crotamine-related gene of *Crotalus durissus terrificus*. Toxicon. 43:751-759.

Reza M, Swarup S, Kini R. 2007. Structure of two genes encoding parallel prothrombin activators in *Tropidechis carinatus* snake: Gene duplication and recruitment of factor X gene to the venom gland. J Thromb Haemost. 5:117-126.

Kipling R. 1902. Just so stories. London: Macmillan and Co. Limited.

Schwartz TS, Tae H, Yang Y, Mockaitis K, Van Hemert JL, Proulx SR, Choi J, Bronikowski AM. 2010. A garter snake transcriptome: Pyrosequencing, de novo assembly, and sex-specific differences. BMC Genomics. 11:694.

Schwartz TS, Bronikowski AM. 2013. Dissecting molecular stress networks: Identifying nodes of divergence between life-history phenotypes. Mol Ecol. 22:739-756.

Shedlock AM, Botka CW, Zhao S, Shetty J, Zhang T, Liu JS, Deschavanne PJ, Edwards SV. 2007. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. Proc Natl Acad Sci U S A. 104:2767-2772.

Stewart AJ, Hannenhalli S, Plotkin JB. 2012. Why transcription factor binding sites are ten nucleotides long. Genetics. 192:973-985.

Sunagar K, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A. 2012. Evolution of CRISPs associated with Toxicoferan-reptilian venom and mammalian reproduction. Mol Biol Evol. 29:1807-1822.

Sunagar K, Fry BG, Jackson TN, Casewell NR, Undheim EA, Vidal N, Ali SA, King GF, Vasudevan K, Vasconcelos V. 2013. Molecular evolution of vertebrate neurotrophins: Co-option of the highly conserved nerve growth factor gene into the advanced snake venom arsenal. PLOS One. 8:e81827.

Tzika AC, Helaers R, Schramm G, Milinkovitch MC. 2011. Reptilian-transcriptome v1.0, a glimpse in the brain transcriptome of five divergent Sauropsida lineages and the phylogenetic position of turtles. EvoDevo. 2:1-18.

Van Damme EJ, Culerrier R, Barre A, Alvarez R, Rouge P, Peumans WJ. 2007. A novel family of lectins evolutionarily related to class V chitinases: An example of neofunctionalization in legumes. Plant Physiol. 144:662-672.

Vonk FJ, Casewell NR, Henkel CV, Heimberg AM, Jansen HJ, McCleary RJ, Kerkkamp HM, Vos RA, Guerreiro I, Calvete JJ et al. 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. Proc Natl Acad Sci U S A. 110:20651-20656.

Wagstaff SC, Laing GD, Theakston RDG, Papaspyridis C, Harrison RA. 2006. Bioinformatics and multiepitope DNA immunization to design rational snake antivenom. PLOS Medicine. 3:e184.

Wagstaff SC, Sanz L, Juárez P, Harrison RA, Calvete JJ. 2009. Combined snake venomics and venom gland transcriptomic analysis of the ocellated carpet viper, *Echis ocellatus*. J Proteomics. 71:609-623.

Wall CE, Cozza S, Riquelme CA, McCombie WR, Heimiller JK, Marr TG, Leinwand LA. 2011. Whole transcriptome analysis of the fasting and fed Burmese python heart: Insights into extreme physiological cardiac adaptation. Physiol Genomics. 43:69-76.

Warrell DA. 2010. Snake bite. Lancet. 375:77-88.

Wong ES, Belov K. 2012. Venom evolution through gene duplications. Gene. 496:1-7.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 20:1377-1419.

Zhang J. 2003. Evolution by gene duplication: An update. Trends Ecol Evol. 18:292-298.

Acknowledgements

The authors wish to thank R. Morgan, A. Barlow and C. Wüster for technical assistance and S. Webster and W. Wüster for comments on the manuscript. We would also like to acknowledge the always enthusiastic help and support of the late Ashley Tweedale. We are very grateful to the staff of High Performance Computing (HPC) Wales for enabling and supporting our access to their systems. This work was partially supported by a Royal Society Research Grant awarded to JFM (grant number RG100514) and Wellcome Trust funding to DWL (grant number 098051). JFM, MJH and MTS are supported by the Biosciences, Environment and Agriculture Alliance (BEAA) between Bangor University and Aberystwyth University and ADH is funded by a Bangor University 125th Anniversary Studentship.

Figure Legends

Figure 1. Restriction and recruitment. Duplicated genes may be either restricted or recruited to the venom gland, with the latter process dependent on the evolution of new combinations of transcription factor binding sites in upstream regulatory regions. Mutation/loss of regulatory regions is indicated with an X.

Figure 2. Tissue distribution of putative toxin gene families. Many proposed toxin gene families are expressed in a wide range of tissues, including the salivary or venom gland and have therefore been restricted to the venom gland following duplication, not recruited. Tissue abbreviations: Sal, salivary gland; VG, venom gland; Bra, brain; Liv, liver; K, kidney; O, ovary; P, pooled tissue (see text for details). Species abbreviations: Ema, leopard gecko (*Eublepharis macularius*); Pre, royal python (*Python regius*); Oae, rough green snake (*Opheodrys aestivus*); Pgu, corn snake (*Pantherophis guttatus*); Eco, painted saw-scaled viper (*Echis coloratus*); Oha, king cobra (*Ophiophagus hannah*); Tel, garter snake (*Thamnophis elegans*).

Figure 3. Maximum likelihood tree of *complement C3* genes. *complement C3* genes are expressed in a diversity of tissues, including venom and salivary glands. Following a gene duplication event (marked with *, shaded dark grey) one paralog has been restricted to the

venom gland in the king cobra (*Ophiophagus hannah*) and the monocled cobra (*Naja kaouthia*). The two distinct king cobra sequences most likely represent geographic variation between Indonesian and Chinese populations. An additional gene duplication event appears to have occurred in the *Austrelaps superbus* lineage (marked with +, shaded light grey). Lineages for which body (non-venom gland) sequences are available are coloured blue and bootstrap values for 500 replicates are shown above branches.