DOCTOR OF PHILOSOPHY

Simplifying complexity: Investigating snake venom evolution using comparative transcriptomics and genomics

Hargreaves, Adam

Award date:
2014

Awarding institution:
Bangor University

SIMPLIFYING COMPLEXITY:

INVESTIGATING SNAKE VENOM EVOLUTION USING COMPARATIVE TRANSCRIPTOMICS AND GENOMICS

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

By Adam David Hargreaves

Bangor University

September 2014

"Simplicity is a great virtue but it requires hard work to achieve it and education to appreciate it. And to make matters worse: complexity sells better."

-Edsger Wybe Dijkstra

ABSTRACT

Snakes represent a diverse reptile lineage, with the evolutionary innovation of venom allowing them to exploit and thrive in different ecological niches. Snake venom has frequently been proposed to be highly complex, having evolved a single time at the base of squamate reptiles with venom genes being "recruited" from multiple body tissues. However, the genetic mechanisms behind these processes and those underpinning the regulation of venom gene expression are poorly understood. Therefore, 2nd generation sequencing was utilised in order to investigate and evaluate these processes. Several methods were first used in order to assess the optimal methodology for genome and transcriptome assembly. Comparative transcriptomic analyses revealed that venom is likely to be a simple mixture containing products from a few gene families. The Toxicofera hypothesis, proposing a single early origin of venom in reptiles, was found to be unsupported in a number of regards. Evaluation of venom gene recruitment revealed that this hypothesis was never supported originally, and newly generated data suggests that venom genes are ancestrally expressed in multiple tissues, including the salivary gland of non-venomous reptiles. Venom genes likely arise through restriction of their expression to the venom gland following gene duplication, and venom can be considered to simply be a modified version of saliva. The transcription factors (TFs) and signalling pathways in the venom and salivary glands are highly similar, with only one transcription factor found in the venom gland but not the salivary gland. Utilising a comparative genomics approach it was found that transcription factor binding sites are conserved between members of the same venom gene families, but not between families, suggesting multiple gene regulatory networks are behind snake venom production. Temporal variation in venom gene expression also appears to support this hypothesis. In short, snake venom has evolved via much simpler processes than previously thought.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor Dr. John Mulley for providing me with the opportunity to complete my PhD at Bangor University. It's been a hugely rewarding experience, never boring, always fun, frequently challenging and *mostly* unstressful! I am extremely grateful for his guidance, support, openness to ideas, generosity with time, and seemingly bottomless knowledge of biology which, taken together, make him a great mentor.

I am grateful to several collaborators who have been extremely generous with their time and expertise. Dr. Justin Pachebat and Dr. Matthew Hegarty for letting me spend time in their labs, showing me the ropes with library preparation, and letting me near some extremely expensive machines without any hesitation. I am indebted to Dr. Martin Swain for his time, patience and bioinformatics wisdom when turning the computer off and on again just didn't solve the problem. Thanks also to Dr. Darren Logan for undertaking a large amount of reptile RNA sequencing, especially as snakes eat the species he normally works with…

Thank you to the members of my PhD thesis committee, Professor Simon Webster and Dr. Michael Knapp, for constructive comments and useful suggestions to keep this work on-course.

Huge thanks to High Performance Computing Wales, especially Laura Redfern for dealing with the endless paperwork and Ade Fewings for being a wizard when it comes to computing. I'm also really sorry that I crashed the system. Twice.

Thank you to Rhys Morgan, Wolfgang and Cathy Wüster and Axel Barlow for taking excellent care of the animals, and for technical help with things such as milking. Thank you to my MRes supervisor Dr. Anita Malhotra for telling me about this PhD studentship!

I would also like to acknowledge and express my sincere gratitude and respect to the late Ashley Tweedale for his enthusiasm, dedication and strength.

Quite a few folks have made living in Bangor and working in the looney-bin of a lab that is D2 so much better: Rhys and Amy for the hilarity, Jess for the amazing baking and foot-fives, Emma for keeping the office innuendos alive, Rosie for throwing a tonne of demonstrating work and cake my way, Karim, Rick, Dylan, Becky, Widad, Marie, Isabelle, Liz, Gaz, Rolf, Louise, Katie, Rich, and Dan and Gem (and the pack!) for being party animals.

Thank you of course to my friends and family back home for their constant support and encouragement. In particular, thank you to my parents for being therapists, extremely forgetful loan sharks, and bartenders all rolled into one.

AUTHOR DECLARATION

The work presented in this thesis was performed entirely by myself and any work done in collaboration is specifically indicated in the text. Several perl and python scripts were used during this study, and the websites which they were obtained from are cited in the text. However, they are also provided on an additional material CD in case the websites are discontinued.

PUBLICATIONS AND PRESENTATIONS ARISING FROM THIS WORK

Chapters 4, 5 and 6 represent publications either in press or in review, and as such are presented as modified versions of publication manuscripts. The published versions of several journal articles can be seen in the Appendices section.

1. **Hargreaves, A.D.**, Swain, M.T., Hegarty, M.J., Logan, D.W., Mulley, J.F. Restriction and recruitment - gene duplication and the origin and evolution of snake venom toxins. *Genome Biology and Evolution*. **6** (8): 2088-2095.

2. **Hargreaves, A. D.** and Mulley, J. F. A plea for standardized nomenclature of snake venom toxins. *Toxicon* **90**: 351-353.

3. **Hargreaves, A.D.**, Swain, M.T., Logan, D.W., Mulley, J.F. Testing the Toxicofera: comparative reptile transcriptomics casts doubt on the single, early evolution of the reptile venom system. *Toxicon* **92**: 140-156.

4. **Hargreaves, A.D.**, Swain, M.T., Hegarty, M.J., Logan, D.W., Mulley, J.F. Genomic and transcriptomic insights into the regulation of snake venom production. (In review).

Various aspects of this work have been presented as oral and poster presentations at the first snake genomics and integrative biology meeting (Colorado, USA) in 2011 and the UK Next-Generation Sequencing meeting (Nottingham, UK) in 2012 and the Bangor University Herpetological Society Venom Day in 2013 and 2014.

LIST OF FIGURES

LIST OF TABLES

Chapter 1
General introduction

This PhD project set out to achieve two broad goals. Firstly, to utilise 2nd generation sequencing technology to generate genomic and transcriptomic resources for an assortment of reptile species, snakes in particular. In so doing, a range of bioinformatics tools were employed in an attempt to determine the optimal method for assembling and analysing these data, as no gold standard currently exists. Secondly, the evolution of snake venom is proposed to have evolved once early in the evolutionary history of squamate reptiles, the so-called "Toxicofera hypothesis". Additionally, venom genes have been proposed to arise through the duplication and "recruitment" of genes to the venom gland where natural selection can then act to develop or increase toxicity. However, both of these hypotheses are based on the analysis of mainly venom gland-derived data, suggesting that their conclusions are perhaps premature and not supported by sufficient evidence. In light of this, these hypotheses were re-evaluated using the newly generated genomic and transcriptomic data.

The purpose of this introduction is to outline the phylogeny and diversity of venomous snakes, the varying definitions of what constitutes a venom, the global medical burden posed by snake envenoming, the inter- and intra-specific variation in snake venom composition and its proposed causes, the potential pharmaceutical uses of snake venom, the proposed origins of venom in snakes, and the Toxicofera hypothesis. As the overarching theme of this thesis includes the extensive use of DNA and RNA sequencing, I will also outline the rationale and methodology behind some of these techniques.

1.1 Snakes

The majority of extant reptiles belong to the order Squamata (Figure 1.1), a name referencing the scaled skin of these animals, which is made up of all lizards and snakes. Snakes are limbless ectothermic animals of the suborder Serpentes, comprising of approximately 3,000 species. Snakes are proposed to have evolved from terrestrial lizards, which consequently lost their limbs as an adaptation to a fossorial lifestyle (Vidal and Hedges 2004). The majority of all snakes (~2,500 species) make up the Caenophidia or "advanced snakes", a sub-order containing 4 major lineages; the Atractaspidinae (such as stiletto snakes), the Viperidae (vipers and pit vipers), the Elapidae (such as cobras and mambas) and the Colubridae (a polyphyletic group which is constantly undergoing taxonomic revision, but does contain venomous species such as the boomslang, *Dispholidus typus*) (Vonk et al. 2011). Approximately 600 species are traditionally considered to be "venomous" in that they possess venom glands surrounded by compressor muscles, tubular fangs at the front of the mouth and are of medical significance to humans, and these species belong to the former three lineages mentioned. Members of the Colubridae are opisthoglyphous (rear fanged) and do not generally pose a threat to humans (although fatalities have been attributed to the boomslang and twig snakes, *Thelotornis spp.*) and so historically have not been considered to be venomous in the traditional sense.

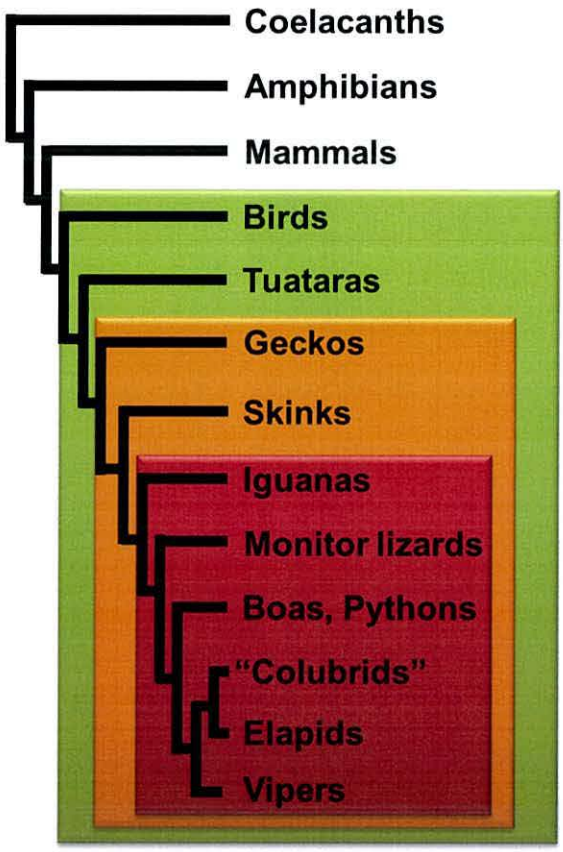


Figure 1.1. Simplified phylogenetic positions of the major groups of the Reptilia (which includes birds) shaded in green, the order squamata shaded orange, and the Toxicofera clade (Vidal and Hedges 2005) (see later section) shaded in red. Due to persistent taxonomic uncertainty within the Colubridae, the term colubrids is placed within inverted commas.

1.2 Venom

The use of venom is ubiquitous throughout the animal kingdom, with venomous animals being found in the Arthropoda (Remipede crustaceans, spiders, scorpions, centipedes, bees, wasps), Cephalopda (octopi), Cnidaria (jellyfish and sea anemones), Gastropoda (cone snails), Actinopterygii (stone fish and lion fish), Reptilia (lizards and snakes) and Mammalia (platypus, shrews and lorises) (Vetter and Visscher 1998; Fry et al. 2006; Fry et al. 2009b; Undheim and King 2011; Ligabue-Braun et al. 2012; Ruder et al. 2013; Nekaris et al. 2013; von Reumont et al. 2014). Venom is predominantly used as a prey capture method but it can also be used defensively (for example in spitting cobras (Westhoff et al. 2005)) or for combat during breeding seasons (such as the platypus, *Ornithorhynchus anatinus* (Whittington and Belov 2007)). Other functions such as acting as an antimicrobial agent (Nair et al. 2007; Ciscotto et al. 2009) or to aid the digestion of large prey items (McCue 2005) have also been suggested. As many venoms in snakes are known to have prey-specific effects (see section 1.4.1), it is likely that venom in snakes has evolved to aid in the incapacitation and killing of prey.

Venom contains an array of active toxic components (and indeed, non-toxic components also), the majority of which are enzymatic proteins. The main characterised toxins found in snake venoms include: snake venom metalloproteinases, serine proteases, phospholipase A_2s, 3 finger toxins and C-type lectins (Ren et al. 1999; Kini and Chan 1999; Lu et al. 2005; Kini and Doley 2010; Markland Jr and Swenson 2013). In general the venom of snakes can be described as being either haemotoxic or neurotoxic (these definitions are by no means mutually exclusive), with viper envenomation largely causing haemotoxic effects and the elapids causing neurotoxic effects. Venom is defined in the Oxford English dictionary as:

"*A poisonous substance secreted by animals such as snakes, spiders, and scorpions and typically injected into prey or aggressors by biting or stinging*"

Therefore it is distinct from a poison which requires ingestion in order to be toxic. Over the years there has been some debate over the definition of what constitutes a venom, especially

within the snake venom research community. Some have specified that a venom should cause "rapid prey death" (Kardong 1980), a definition focusing on the biological role of a venom which can be quantified and easily compared to other uses of oral secretions in snake species classically described as being non-venomous (Kardong 2012). Others have proposed definitions less focused on biological function and more on shared homology across taxa. For example, the most recent proposed definition:

"*...a secretion, produced in a specialised tissue (generally encapsulated in a gland) in one animal and delivered to a target animal through the infliction of a wound (regardless of how tiny it is). A venom must further contain molecules that disrupt normal physiological or biochemical processes so as to facilitate feeding or defence by/of the producing animal.*" (Fry et al. 2012b)

Alongside this definition, and the fact that many of the toxic proteins found in snake venom have evolved convergently in other taxa, the authors also suggest that haematophagus (feeding on blood) animals are encapsulated by this definition (Fry et al. 2012b), and propose that vampire bats (Low et al. 2013), mosquitoes, ticks and even lampreys (Fry et al. 2009b) should be classed as being "venomous".

1.3 The global burden of snakebite and treatment

It has been estimated that globally there are at least 421,000 envenomings by snakes each year resulting in 20,000 deaths (Kasturiratne et al. 2008) (Figure 1.2). However, these figures are based on a conservative estimate, and the annual global mortality due to snakebite could in fact be as high as 94,000 deaths (Kasturiratne et al. 2008). Even so, snakebite has only recently be classified as a neglected tropical disease (NTD) (Warrell 2010) which constitutes a disease affecting predominantly inhabitants of the rural tropics, and the development of a treatment has been largely ignored by funding bodies and research efforts (Feasey et al. 2010). The majority of deaths caused by snakebite occur in Sub-Saharan Africa, Asia and the Americas, and poverty has been shown to be directly associated with an increased mortality rate (Harrison et al. 2009).

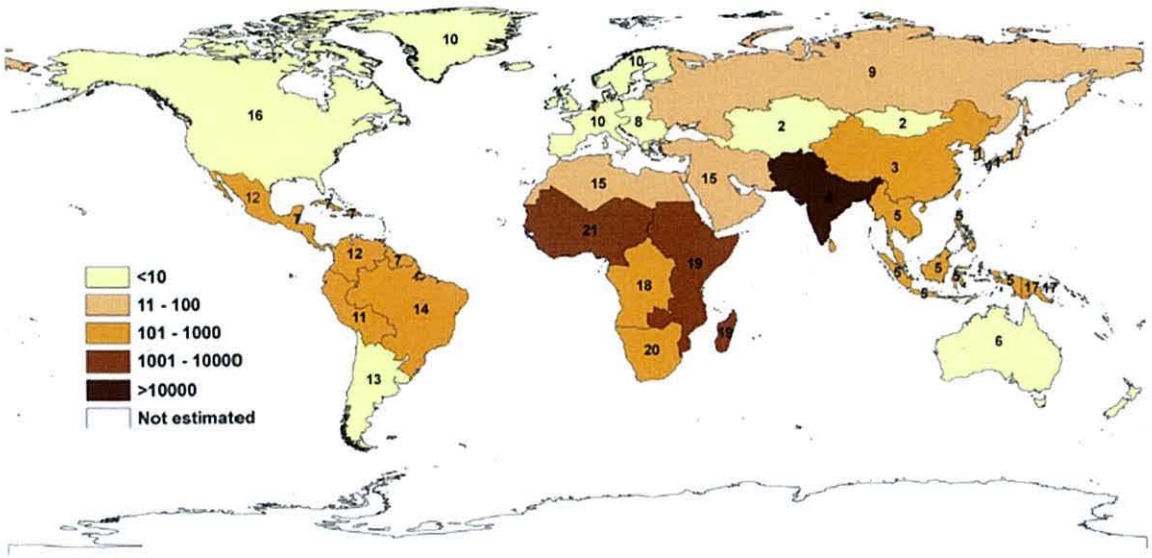


Figure 1.2. Global regional estimates of annual mortality caused by envenomation by snakes. Figure is an adapted version of Figure 7 from (Kasturiratne et al. 2008).

The first treatment against snake envenomation was developed in 1895 by Albert Calmette, an immunologist working under Luis Pasteur (Hawgood 1999), which at the time was called Calmette's serum. His work was later expanded upon by Vital Brazil, who later went on to become the director of the Instituto Butantan in Brazil (Hawgood 1992). The currently used method of producing antivenom involves the hyperimmunisation of a large animal (usually sheep or horses), with the raised IgG antibodies to the venom proteins being extracted and purified to give antivenom (Warrell 2010). Unfortunately there are several downsides to this treatment: the produced antivenom must be kept in cold-chain storage making the transport of antivenoms to rural communities with poor transport infrastructure and, more obviously, a lack of electricity an impossibility (unless freeze-dried, which significantly increases cost). Whilst the IgG molecules present in antivenom are capable of binding toxin molecules and abating the systemic effects of envenomation, they are too large to diffuse into tissues and prevent the local tissue damage caused by some venom toxins (Cook et al. 2010). Ultimately this means that although death may have been avoided there still may be tissue necrosis, which could lead to the development of secondary infections, gangrene, and the necessity of amputation. The use of animal antibodies as a therapeutic treatment also poses the risk of serum sickness, the development of hypersensitivity and in worst cases, anaphylaxis (Warrell 2010). However, it is perhaps the venoms themselves which cause the most confounding problems. The efficacy of an antivenom is completely dependent on the venom(s) used for immunisation which, coupled with the fact that not all proteins found in venom are toxins, can be significantly affected by the choice of animals used to obtain samples. As venoms are known to be highly variable, even

within the same species across different geographical regions, this can have significant implications for antivenom manufacture (Fry et al. 2001; Fry et al. 2003b; Gutiérrez et al. 2009; Casewell et al. 2014; Sunagar et al. 2014).

1.4 Venom variation

Venom composition can display both inter- and intra-specific variation, and the exact causes of this are either not known or poorly understood. Examples include ontogenetic variation between juveniles and adults, sexual differences between males and females and the presence or absence of particular toxins caused by the geographical location of different populations (Chippaux et al. 1991). Venom composition has also been proposed to be adapted to the diet of the snake. These factors are discussed below.

1.4.1 Adaptation to diet

As snake venom is primarily used to aid in prey capture it has been suggested that the venom of a particular species will be tailored to its choice of prey (Daltry et al. 1996) and several studies have shown evidence for prey-specific toxic effects (Gibbs and Rossiter 2008; Gibbs and Mackessy 2009; Barlow et al. 2009; Gibbs et al. 2011; Richards et al. 2012). Instances of resistance to venom toxins in several mammal species have also been well documented (Kilmon Sr 1976; Perez et al. 1978; Menchaca and Perez 1981; Perales et al. 1986; Biardi et al. 2000; Neves-Ferreira et al. 2000; Jurgilas et al. 2003; Ho 2005; Biardi et al. 2006; Biardi and Coss 2011), leading to the formation of the idea that venom evolution constitutes a co-evolutionary "arms race" between predator and prey (Lynch 2007). Indeed venom is postulated to have allowed the evolutionary diversification of the advanced snakes during the Cenozoic era (Vidal and Hedges 2002; Fry et al. 2012b). It can be envisaged that the switch from capturing prey via a mechanical method (i.e. constriction) to a biochemical one (i.e. venom) means a reduction in selective pressure for a large, muscular body. However, the risk of injury from prey items (especially those with teeth and claws) means that there is still a need to quickly incapacitate prey, which will consequently also increase foraging efficiency (the prey does not stray far following envenomation). Furthermore, the evolution of venom toxins which target different receptors or physiological systems may open up possibilities to feed on new prey items, allowing snakes to exploit a new ecological niche. Further evidence for dietary adaptation is the observation that the snake venom system degenerates following a switch in diet to non-threatening prey such as eggs (colloquially known as "use it or lose it") (Fry et al. 2008; Fry et al. 2012b). A prime example of this is the marbled sea snake, *Aipysurus eydouxii*. Investigation

found that after a dietary switch from fish to feeding exclusively on fish eggs the fangs were lost in this species and the venom gland became atrophied. Molecular analysis of venom gland expression found that a neurotoxic three finger toxin expressed by this species had undergone a dinucleotide deletion, resulting in a complete loss of neurotoxic function (Li et al. 2005a). The PLA_2 genes of this species had also accumulated deleterious mutations in their protein coding regions rendering them non-functional (Li et al. 2005b). The question of whether venom aids in the digestion of prey has been subject to some debate, with studies finding that it does (Thomas and Pough 1979) and some that it does not (McCue 2007; Chu et al. 2009). As venom toxins are primarily enzymes, many of which are proteolytic or lipolytic, it is possible that this digestion occurs as a side-effect of envenomation, allowing the ingestion of large prey items.

1.4.2 Ontogenetic, sexual and geographic variation of venom composition

Venom composition also shows ontogenetic variation, in particular a switch in venom composition appears to commonly occur between juvenile and adult (Mackessy 1988; Furtado et al. 1991; López-Lozano et al. 2002; Guércio et al. 2006; Zelanis et al. 2008; Alape-Girón et al. 2008) and is thought to again be due to a shift in diet from small prey items (such as insects and amphibians) to larger ones (such as mammals) (Andrade and Abe 1999; Mackessy et al. 2003; Gibbs et al. 2011).

Venom composition has also been shown to display sexual variation, again further confounding the manufacture of effective antivenoms via current methods as absolute specificity is required for maximum efficacy (Warrell 2010). Proteomic analysis has previously found variation in the presence or absence of proteins between male and female snakes, which results in their respective venoms having different activities (Menezes et al. 2006; Pimenta et al. 2007), and it is possible that this is an adaptation to observed sexual variation in prey choice (Daltry et al. 1998).

Finally, geographic variation in venom composition has also been shown, which can have profound effects for antivenom treatment. For example, Mojave rattlesnake (*Crotalus scutulatus scutulatus*) populations exhibit one of three venom types: A (neurotoxic), B (haemorrhagic) or A+B (a combination of both effects) (Wilkinson et al. 1991). This variation is believed to be due to the absence of the acidic subunit of Mojave toxin (Mtx) (Wooldridge et al. 2001), a neurotoxic phospholipase A_2 heterodimer, in certain populations.

1.5 Pharmaceutical uses of venom

Due to the significant medical burden caused by snake envenoming, treatments to counteract its effects are usually the focus of biological research. However, as snake venom contains a plethora of proteins and peptides specifically targeted to different receptors, it also represents a potential goldmine of compounds with utility in the development of novel pharmaceutical treatments (Vonk et al. 2011). Several components of snake venoms have already been developed either into drugs or medical diagnostic tests. The first Angiotensin-converting enzyme inhibitors (ACE inhibitors) were developed from a venom component (a bradykinin-potentiating factor (Ferreira 1965)) found in *Bothrops jararaca* venom, and subsequently used to treat high blood pressure. Components from venoms are also used as a diagnostic test for blood clotting efficiency, for example Ecarin, a venom disintegrin derived from the venom of the saw-scaled viper *Echis carinatus*, is used to assess blood clotting following treatment with anticoagulants in an assay known as "Ecarin clotting time" or ECT (Fabrizio 2001). Some venom components also have potential uses in cancer therapy, either as a cell marker or as a delivery device for anti-cancer drugs. Crotamine, a venom component found in the venoms of *Crotalus* rattlesnakes, is a modified version of an ancestral β defensin protein. It has been shown to have cell-penetrating properties, including into cells which are actively proliferating, and nucleolar targeting peptides have been derived from the crotamine molecule and are being investigated for use in the translocation of therapeutic treatments directly into cancer cells (Radis-Baptista and Kerkis 2011; Hayashi et al. 2012). Whilst the toxins in snake venoms can cause catastrophic health effects for humans, it is clear that they may also possess new ways to improve healthcare and medical treatment.

1.6 Proposed origins of snake venom

Historically, venom has been suggested to have evolved in a number of ways and originated from a number of sources. Kochva et al. (1983) suggested that as some venom components shared homology with enzymes secreted by the pancreas (namely Phospholipase A_2), and enzymes found in mammalian saliva are also found in the pancreas (α-amylase), venom could have originated from digestive enzymes in snakes. The example of mammalian α-amylase is of significance as the salivary version of this gene is the result of an insertion into the regulatory region of one copy (there are 5 copies arranged in tandem) which originated from the duplication of a pancreatic amylase gene, with a salivary-gland specific expression pattern being caused by the insertion (Meisler and Ting 1993). This process of a duplicate gene evolving a novel expression pattern is known as neofunctionalisation (Force et al. 1999) and

has been hypothesised to be the way in which snake venom genes arise. Previous phylogenetic analysis of snake toxin gene families (Fry and Wuster 2004) led to the hypothesis that genes encoding venom toxins in snakes have evolved via the recruitment of non-toxic physiological genes from a wide range of body tissues into the venom gland following gene duplication where natural selection can then act to develop or increase toxicity (Fry 2005) (Chapter 5). As such, gene duplication is considered to be a key process in the evolution and diversification of reptile venoms (Wong and Belov 2012). Of course not all proteins can become venom toxins, and so not all genes will be recruited. Firstly the gene must be short enough to allow its duplication. Secondly the protein it encodes must be secreted, and so must also possess a signal peptide (Fry et al. 2009b) and the protein must possess a stable molecular scaffold, resulting in a bias towards the selection of cysteine-rich proteins as toxins (Fry et al. 2012a). The strict maintenance of the molecular scaffold of the protein is coupled with the modification of surface residues, sometimes leading to new binding specificities which in turn may lead to novel activities (Fry et al. 2008).

Venom genes are frequently described as evolving via the birth-and-death model of evolution (Nei et al. 1997; Nei and Rooney 2005) whereby a multigene family evolves via gene duplication followed by some copies being maintained in the genome, some being rendered non-functional by pseudogenisation, and others being deleted. It is interesting that so far, the incidence of venom pseudogenes appears to be relatively rare (John et al. 1996; Li et al. 2005a; Li et al. 2005b; Ikeda et al. 2010).

1.7 Theories of reptile venom evolution and the Toxicofera hypothesis

1.7.1 Toxicofera hypothesis foundations

Proteomic analysis of the saliva of the radiated rat snake (*Coelognathus radiatus*), a non-venomous snake reliant on constriction for prey capture, led to the discovery of a post-synaptic neurotoxin belonging to the three finger toxin (3Ftx) family (Fry et al. 2003c). This protein was found to possess the typical ten conserved cysteine residues of elapid 3Ftxs and when functionally tested led to antagonism of nicotinic acetylcholine receptors. As such this protein was considered to be structurally and functionally homologous to the elapid three finger toxins (Fry et al. 2003c). Phylogenetic analysis showed strong support for the nesting of the rat snake 3Ftx within a clade of previously categorised 3Ftxs (Fry et al. 2003a) which was interpreted to mean that the three finger toxins were recruited into the venom repertoire of the advanced snakes early in their evolutionary history prior to the divergence of the Elapidae and Colubridae (Fry et al. 2003c). Indeed the analysis of other colubrid "venoms" (Mackessy 2002; Fry et al.

2003d; Lumsden et al. 2005) added further support that the use of venom in the advanced snakes pre-dated their radiation.

1.7.2 The Toxicofera hypothesis and its propagation and expansion

The Toxicofera is a hypothetical clade of squamate reptiles consisting of the Iguania, Anguimorpha and Serpentes (Vidal and Hedges 2005), and its name refers to the presence of venom within these groups. Phylogenetic analysis utilising 9 nuclear genes found this clade to be strongly supported (Vidal and Hedges 2005) , however phylogenetic relationships within the Toxicofera are unresolved based on nuclear data, although the use of Sauria SINEs (short interspersed nuclear elements) suggests a clustering of snakes with anguimorph lizards (Piskurek et al. 2006). It was previously believed that venom evolved twice independently in squamate reptiles, once in Serpentes (snakes) and once in the Helodermatid lizards (Gila monsters and beaded lizards) (Pough et al. 2004). This belief was mainly due to the distant phylogenetic relatedness of these animals and the clear differences in the morphology of their respective venom delivery systems (Fry et al. 2010b). The presence of putative toxin proteins in the saliva of species usually regarded as non-venomous, and the expression of venom gene homologs in their salivary glands, led to the hypothesis that venom evolved a single time in squamate reptiles, and not twice independently as had been previously believed (Fry et al. 2006). The "single, early origin" hypothesis is synonymous with the Toxicofera clade, and therefore will subsequently be referred to as "the Toxicofera hypothesis". Based on fossil records the authors deduced that the emergence of venom in squamate evolution dates back to 200 Myr ago, and not 100 Myr ago as had previously been postulated (Fry et al. 2006). The main findings of the original Toxicofera study (Fry et al. 2006) was the detection of putative venom toxins expressed in the salivary glands of non-Helodermatid lizards, namely a Monitor lizard (*Varanus varius*) and an Iguanian (*Pogona barbata*). Further phylogenetic analysis demonstrated that nine toxin families were shared between what are frequently regarded as non-venomous lizards and advanced snakes, which are AVIT peptide, B natriuretic peptide, cysteine-rich secretory protein (CRISP), cobra venom factor (which is in fact complement component c3 (Alper and Balavitch 1976)), crotamine, cystatin, kallikrein, nerve growth factor and vespryn. Based on the presence or absence of these genes, timings of venom gene recruitment events were estimated, which can be seen in Figure 1.3.

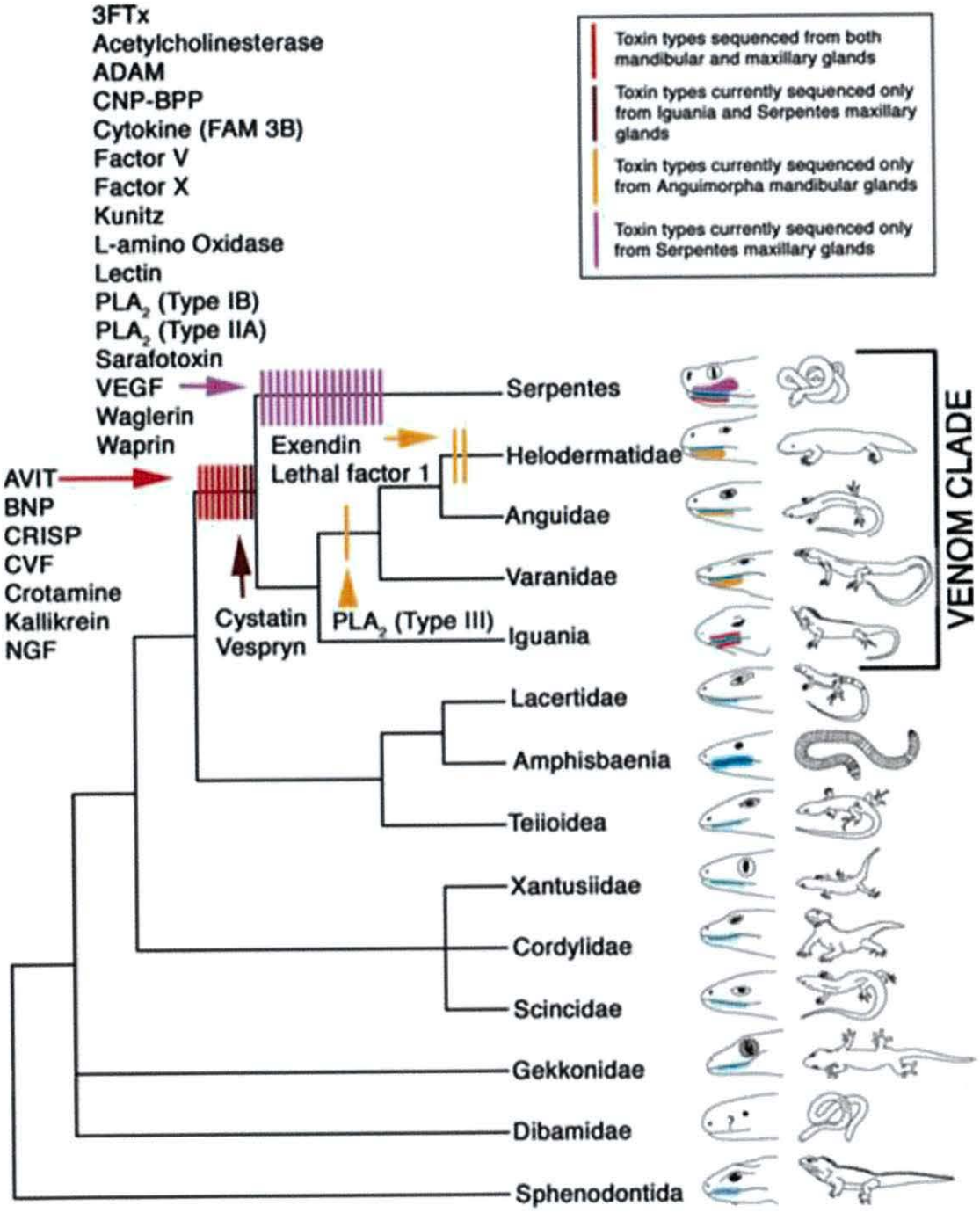


Figure 1.3. Cladogram of the Reptiles with the timings of venom gene recruitment events indicated for the Toxicofera clade. Figure is taken from (Fry et al. 2006).

The expression of these putative toxins in the salivary glands of non-Helodermatid lizards ultimately led to the proposal that venom evolved once in reptiles, and that extant members of the Toxicofera clade share a venomous ancestor. This early venomous squamate is proposed to have possessed primitive toxin-secreting glands located on both the upper and lower jaw (Fry et al. 2006). The venom delivery systems in advanced snakes and lizards are therefore homologous but morphologically distinct derivatives of this primitive system, with snakes retaining the maxillary venom glands and venomous lizards maintaining the mandibular glands (Fry et al. 2006; Fry et al. 2010a), with the opposing glands being secondarily lost by each lineage. The authors propose that members of the Iguania (such as the green anole lizard, *Anolis carolinensis*) diverged whilst this venom system was in an incipient stage (Fry et al. 2012b), and so lack any form of specialised toxin secreting glands. Furthermore, snakes which now use alternative prey capture methods such as constriction are proposed to have secondarily lost venomous function (Fry et al. 2012b).

The timing of proposed venom gene recruitment events has undergone significant modification over the course of subsequent Toxicofera-related studies (Fry et al. 2009a; Fry et al. 2010b; Koludarov et al. 2012; Fry et al. 2012a; Fry et al. 2012b; Fry et al. 2013). Further sampling has led to the detection of an increased number of proposed basal Toxicoferan genes, leading to a much more complicated view of venom gene recruitment throughout the evolution of the Toxicofera clade (Figure 1.4). These studies have also led to the inclusion of charismatic megafauna such as the Komodo dragon, *Varanus komodoensis* which is now considered to be venomous (Fry et al. 2009c).

Whilst the Toxicofera hypothesis and its claims have been met with some resistance (Kardong 2012; Weinstein et al. 2012), especially as many of these putative toxins have not been functionally characterised, the Toxicofera hypothesis has become widely accepted amongst the toxinological community.

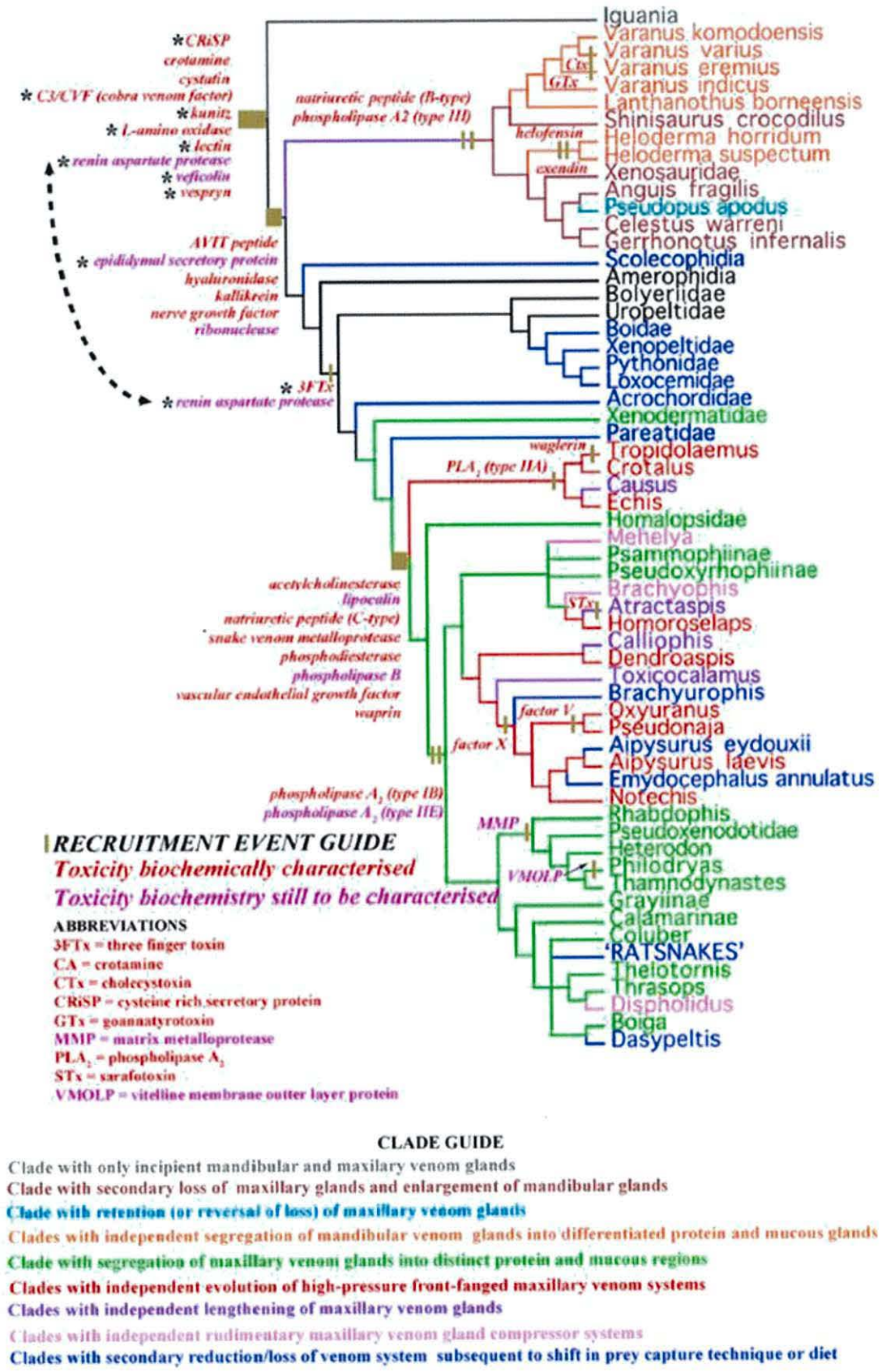


Figure 1.4. The most recent cladogram of the Toxicofera clade with timings of venom gene recruitment events indicated. Figure taken from (Fry et al. 2013).

Several other details are pertinent to the Toxicofera hypothesis and the overarching theme of this thesis. Firstly, all sampling carried out in support of the Toxicofera hypothesis was from either venom or salivary glands, and no other body tissues were sequenced. Additionally, only 384 ESTs (expressed sequence tags, see later section) were sequenced per sample (Fry et al. 2006; Fry et al. 2010b; Fry et al. 2012a) (later studies switched to using 454 pyrosequencing), a minimal amount of sequencing considering the frequently cited complexity of reptile venoms (Li et al. 2005b; Wagstaff et al. 2006; Kini and Doley 2010). None of the detected homologous toxins were purified and functionally characterised. Indeed, Renin aspartate protease, which was initially detected expressed in the venom gland of *Echis ocellatus* (Wagstaff and Harrison 2006), was further suggested to be a Toxicoferan toxin (Fry et al. 2008) despite never being shown to actually be toxic other than possessing protease activity. Furthermore the authors state that "A number of frameworks expressed in the venom glands are known only from the mRNA transcripts or corresponding bioactivities remain to be elucidated", implying that some of the toxins they use to support the Toxicofera have actually never been shown to be toxic, in venomous species or otherwise.

The detection of nerve growth factor (*ngf*) in the mandibular salivary gland transcriptome of *Abronia graminea* led to the conclusion that because it (and in fact other venom gland *ngf* transcript sequences from snakes) was highly homologous to nuclear gene sequences of *ngf* sequenced from genomic DNA used previously for phylogenetic analysis (Wiens et al. 2010), it must represent a venom toxin without the requirement of gene duplication (i.e. an incidence of pleiotropy, where a single gene fulfils multiple roles) (Koludarov et al. 2012). Here it seems that the presence of a sequence has been concluded to imply homology and a toxic function. It is possible that a gene may be used pleiotropically as a toxin, but unless its expression is elevated in the salivary gland, there would be little evidence to suggest that it was anything more than a housekeeping or maintenance gene, expressed at similar levels in multiple tissues. It is mentioned in a later study that a limitation of the methodology used is potential incomplete sampling due to the exclusion of non-venom gland data, which could give rise to incorrect monophyletic "venom clades" during phylogenetic analysis (Koludarov et al. 2012). The authors then state that Casewell et al. (2012) abated this in their study by including non-venom gland tissue data in their phylogenetic analyses, which provided further support for the single early origin of venom hypothesis. On examination of this paper, which did indeed conduct phylogenetic analyses of venom toxin sequences alongside non-venom gland sequences derived from Burmese python (*Python molurus bivittatus*) pooled tissues (heart and liver) (Castoe et al. 2011) and garter snake (*Thamnophis elegans*) pooled tissue (brain, gonads, heart, kidney, liver,

spleen and blood of males and females) (Schwartz et al. 2010), the findings are intriguing. Casewell et al. (2012) found that non-toxin sequences nested within clades of toxin gene families "...providing strong evidence for the non-monophyly of Toxicoferan toxins" and that "...the results of our phylogenetic analyses would strongly refute the key prediction of the 'SEO' (single early origin) hypothesis...". They then go on to propose that the origin of venom toxins via recruitment is not a one-way process, suggesting that a venom toxin may again duplicate and be recruited back into the body to fulfil a physiological role, so-called "reverse recruitment" (Casewell et al. 2012). However, the more parsimonious hypothesis that these sequences actually represent *reptile* body sequences (which have never been toxins) forming *reptile* clades rather than body sequences nesting within *venom* clades is not considered.

In the most recent study (Fry et al. 2013) the most highly expressed transcripts within the Iguanian species sampled were found to be crotamine/β defensin and cystatin, and the authors acknowledge that these peptides are known to possess antimicrobial function. The fact that these two types of protein are commonly found to be expressed in saliva and the salivary glands (Mathews et al. 1999; Dickinson 2002; Abiko et al. 2003) should perhaps make their detection in the salivary glands of Iguanian lizards (and others) unsurprising.

Therefore, on further scrutiny of the studies proposing support for the Toxicofera hypothesis, there appears to be a number of discrepancies warranting further investigation, particularly in cases where presence and homology is concluded to represent ancestral toxicity.

1.8 Frederick Sanger and the start of DNA sequencing

During the 1970's, Frederick Sanger and colleagues developed a technique for sequencing DNA, which was, and still is, a widely used DNA sequencing method. Dideoxy chain termination sequencing (more commonly referred to as "Sanger sequencing") (Figure 1.5) involves the extension of a single-stranded DNA molecule using a sequencing primer, DNA polymerase, normal deoxynucleotidetriphosphates (dNTPs) and modified di-deoxynucleotidetriphosphates (ddNTPs) which are fluorescently labelled and lack a 3'-OH group causing the cessation of DNA polymerisation following their incorporation (i.e. chain termination) (Sanger et al. 1977). As each ddNTP is labelled with its own fluorescent signal, every fragment arising from the sequencing reaction can be "read" in order to determine the progressing DNA sequence, for example by a capillary sequencing machine. This technology opened up genetics and molecular biology as a whole, perhaps most significantly being essential for the sequencing of the human genome (Venter et al. 2001).

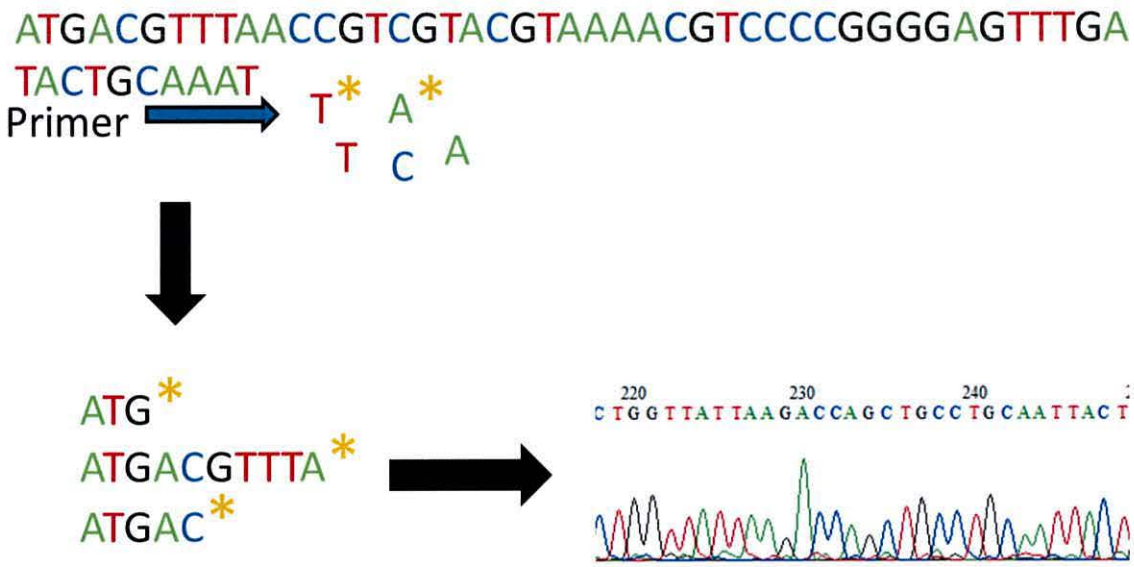


Figure 1.5. Di-deoxy chain termination sequencing.

1.9 Traditional methodologies in venom research

Much of the previous research efforts into snake venom have employed either proteomics (so-called "venomics" (Calvete et al. 2007a)) or the cloning and sequencing of a relatively small number (≤1000) of expressed sequence tags (ESTs) by sanger sequencing, for example (Wagstaff and Harrison 2006; Casewell et al. 2009; Siang et al. 2010). Even recently ESTs have been used for venom research (Casewell et al. 2014). Expressed sequence tags are short sequences of clones derived from a cDNA library prepared from a sample of RNA, which are then sequenced using Sanger sequencing. Whilst ESTs have proved useful in the past (for example in the discovery of new genes for the human genome project (Adams et al. 1991) they represent a limited approach in a number of ways. Firstly, the cDNAs cloned (and therefore the ESTs sequenced) are completely dependent on what genes are being actively transcribed in the tissue/cell of interest at the time. Thus the absence of a gene from a sample does not necessarily mean it is not expressed in that tissue. Secondly, the ESTs sequenced are dependent on the cDNA library. As the expression of genes is a dynamic process, so too is the presence and levels of mRNA molecules transcribed by a cell. As such the amount of mRNA transcripts per expressed gene will not be equal, leading to bias within the cDNA library (over- and under-represented transcripts) (Nagaraj et al. 2007). This gives some information in terms of what genes are highly expressed, but lowly expressed genes may not be captured unless sufficient numbers of ESTs are sequenced. Secondly, the reverse transcriptase used to synthesise the cDNA and the sequencing reaction can both incorporate error into the generated sequence data.

Coupled with the fact that ESTs are generally partial, low quality fragments of a cDNA molecule, a high amount of sequencing is needed to correctly infer a nucleotide sequence with minimal chance of error. Finally, cDNA libraries will contain many redundant sequences (i.e. identical fragments sequenced multiple times) making the assembly of ESTs computationally problematic (Parkinson and Blaxter 2004). More recently, the use of new sequencing methods, for a long time dubbed "next-generation sequencing" but now referred to as 2nd generation sequencing, has begun to increase in venom research. Beginning with the sequencing of snake "body tissue" transcriptomes (Schwartz et al. 2010; Castoe et al. 2011), venom gland trancriptomes have become more and more commonplace (Durban et al. 2011; Rokyta et al. 2012; Margres et al. 2013; Aird et al. 2013) and even whole genome sequences for venomous snakes have begun to emerge (Vonk et al. 2013).

1.10 The genomics era

Following the sequencing and completion of several genome projects (for example *Caenorhabditis elegans* (Sequencing Consortium 1998), *Drosophila melanogaster* (Adams et al. 2000) and *Takifugu rubripes* (Aparicio et al. 2002)) and of course the human genome project (Venter et al. 2001; McPherson et al. 2001; Lander et al. 2001; Ross et al. 2005; Gregory et al. 2006), the number of genomes sequenced has increased exponentially. The number of planned genome projects for the future has also drastically increased (Haussler et al. 2009) aided by the rapid improvement in sequencing technology and its reducing cost (Figure 1.6).

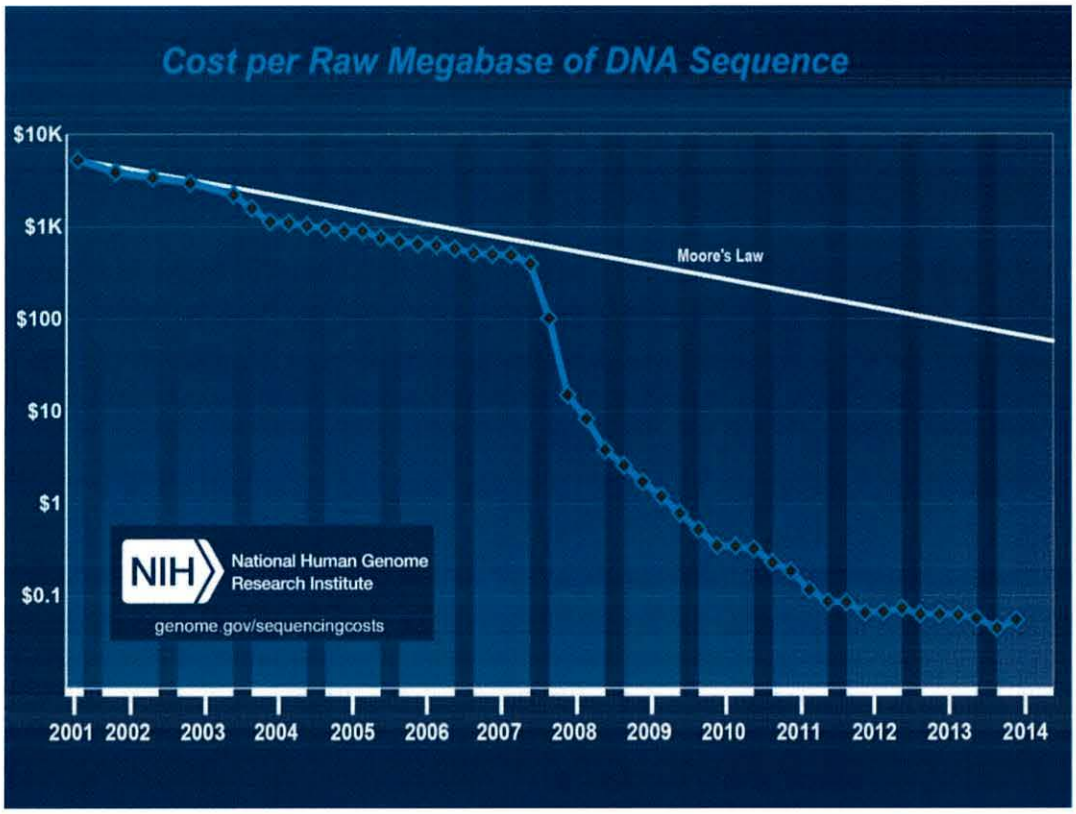


Figure 1.6. Decreasing costs of DNA sequencing since the start of the human genome project in 2001 (www.genome.gov/sequencingcosts).

The entire human genome project had a total cost in the region of 2.7 billion US dollars (http://www.genome.gov/11006943), a figure in stark contrast to the recently publicised "$1,000 genome" following the release of the Illumina HiSeq X Ten system (a package where ten sequencers must be purchased together, costing in the region of ten million dollars, with the proviso that only human samples must be sequenced on the machines) (http://systems.illumina.com/systems/hiseq-x-sequencing-system.ilmn).

1.11 The dawn of shotgun sequencing

Whilst the utility and accuracy of traditional Sanger sequencing should not be understated, it is ultimately limited by the length of sequence it produces and throughput, making the sequencing of large genomic regions or whole genomes very inefficient and time consuming. In shotgun sequencing, DNA is fragmented randomly into many small pieces which are then sequenced. The sequenced fragments of shotgun sequencing are referred to as "reads" (Figure 1.7) and can either be single-end (sequenced from one end of the fragment) or paired-end (sequenced from both ends of the fragment). In this format, reads provide little utility. However, if enough reads are sequenced there will be the same region of DNA present in the sample of reads multiple

times, and reads are likely to overlap at their ends. As a result, reads can be overlapped to assemble longer lengths of contiguous sequence, which are known as "contigs" (Figure 1.7).

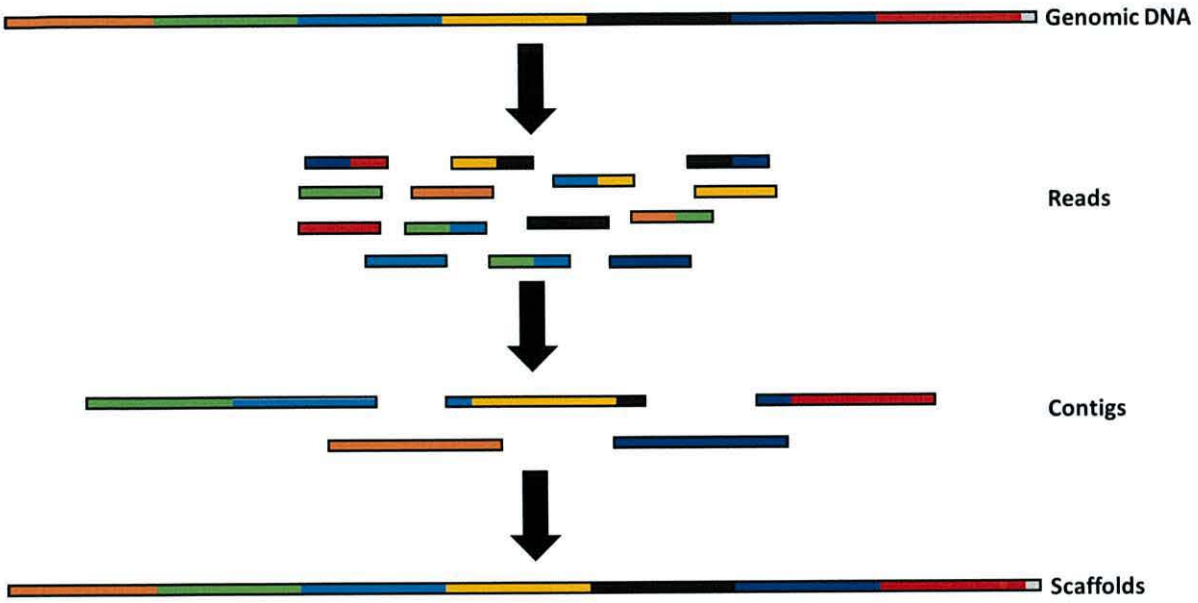


Figure 1.7. Shotgun sequencing to give reads which are assembled into contigs, which can further be assembled into scaffolds.

If reads are generated which confer positional information (such as from the sequencing of mate-pair libraries, see next section), contigs can be orientated and overlapped together to form longer lengths of sequence known as "scaffolds". If gaps between contigs cannot be filled with known sequence, but the distance between them is known, these gaps are usually filled with N's until such a time as the missing sequence can be elucidated.

When preparing genetic material (either DNA or RNA) for sequencing, it is usually fragmented to fragments of a desired size and subsequently modified to be compatible with the sequencing technology to be used (see next section). Once this is complete the resulting prepared fragments are said to be a "library", in the sense that it contains fragments of all the genetic material in the original sample. Library preparation can be carried out using small sized fragments or much longer fragments, the length of which is known as the "insert size".

Mate-pair libraries (Figure 1.8) are constructed using much longer fragments of DNA. Essentially through the biotinylation and circularisation steps, fragments are produced whose ends are a known distance apart. As such the sequencing of fragments in a mate-pair library (using paired-end sequencing, to gain the DNA sequence from regions which are distantly

located from each other) provides long-distance positional information which can be used to inform the position and orientation of contigs to construct scaffolds.

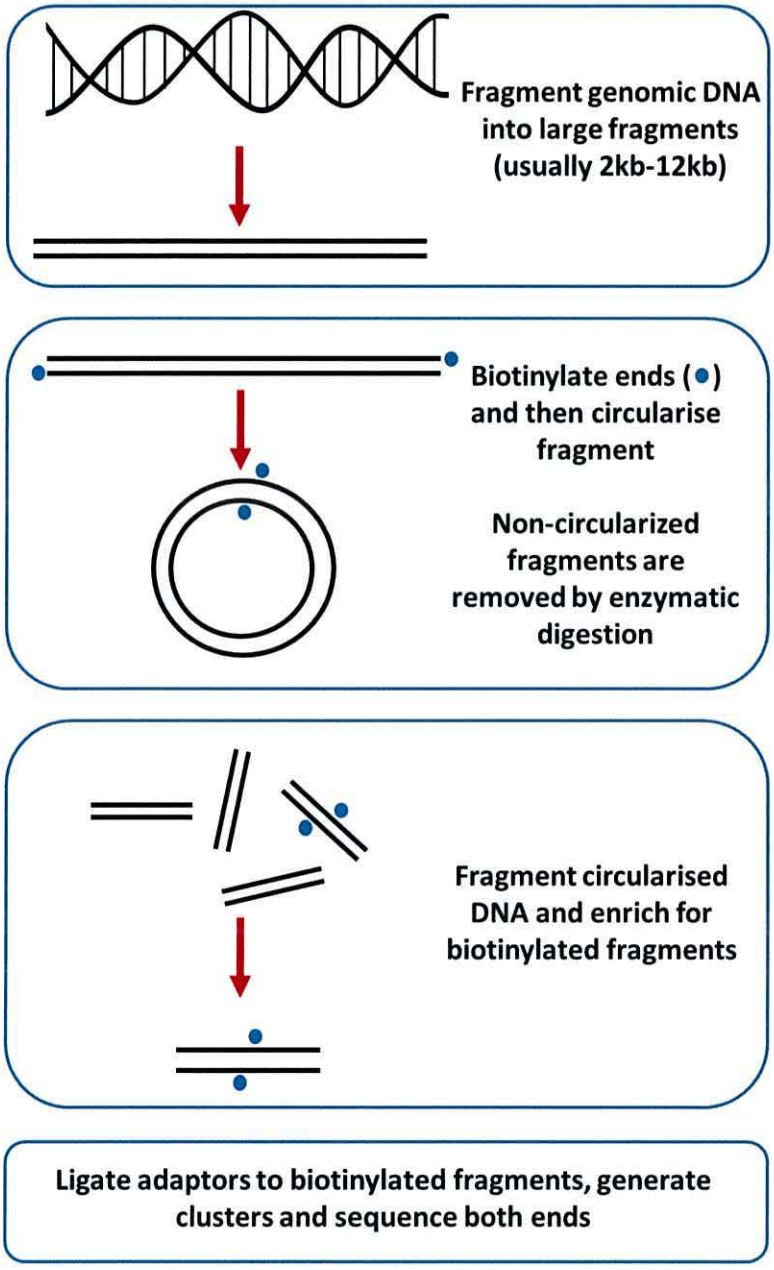


Figure 1.8. Workflow of a typical mate-pair library preparation.

Nowadays, the number of reads generated can be in the billions, and so must be assembled computationally from files output by a sequencing machine. The most commonly used file format (although certainly not the only one) is FASTQ format (herein denoted as .fastq which is the suffix for files of this type).

In fastq format, the information for each sequenced read is placed on four consecutive lines. The first line begins with the character "@" and then is proceeded by a sequence identifier. The identifier usually contains information such as the name of the sequencing instrument, the lane number of the flow cell where the sequence came from, and an indication of whether the read is the "forward" or "reverse" strand of a DNA fragment (for paired-end sequencing) usually indicated with the number 1 or 2. Line number 2 contains the raw nucleotide sequence of the read. The third line begins with the "+" symbol and may or may not contain a repeated sequence identifier. Finally, line 4 contains the quality values for each base present in line number 2 (Cock et al. 2010).

Following the assembly of reads into contigs or scaffolds, sequences are usually output in another format, known as FASTA (herein denoted as .fasta which is the suffix for files of this type). .fasta format files usually contain sequence (DNA, RNA or protein) data, with each line beginning with the symbol ">", followed by an identifier, and then the sequence.

1.12 2nd generation sequencing technologies

Several 2nd generation (formerly referred to as "next-generation") sequencing technologies are described below. As there are a multitude of different technologies not all have been included (such as sequencing by oligonucleotide ligation and detection, or SOLiD sequencing) as they are not directly relevant to the subject of this thesis. Instead, those that have been used previously in snake venom research have been described, namely Roche/454 pyrosequencing and Illumina sequencing.

1.12.1 Pyrosequencing

The Roche/454 FLX sequencing platform was the first commercially available sequencing machine, utilising a method known as pyrosequencing. At one time this system offered the longest sequencing reads available (~400bp) but has since been eclipsed by other competitors. Due to these long read lengths this technology was preferentially used during the advent of 2nd generation sequencing use in reptile research, with the first snake "next-gen" transcriptomes (Schwartz et al. 2010; Castoe et al. 2011) sequenced using 454.

DNA library preparation for 454 is similar to most other sequencing technologies in that the DNA is first fragmented and has adapters ligated to both ends of the fragments. Fragments are then mixed with agarose beads which have oligonucleotides complementary to the library adapter sequences on their surface, resulting in the binding of a single DNA fragment to each

bead. The bead/fragment complexes are then amplified using emulsion PCR, generating roughly one million copies of the DNA fragment on the surface of each bead. Each bead is subsequently placed in a well of a picotiter plate (PTP) (one bead per well) and reagents are added prior to centrifugation which catalyse the subsequent pyrosequencing reaction.

Pyrosequencing relies on the incorporation of one of four dNTPs (A, T, G, C) to the fragment molecule by DNA polymerase. After each addition a pyrophosphate (PPi) molecule is released. In the presence of ATP sulfurylase and adenosine 5' phosphosulfate, PPi is converted into ATP (which is a quantitative reaction, for example in the case of the dNTP molecules binding to a stretch of consecutive nucleotides along the fragment) which is further used in the reaction converting luciferin to oxiluciferin by the enzyme luciferase. This final reaction produces light which is detected by the sequencer and analysed. Any unincorporated nucleotides are degraded by the addition of the enzyme apyrase and the reaction is repeated using a different dNTP (Ronaghi et al. 1996; Margulies et al. 2005; Mardis 2008).

1.12.2 Sequencing by synthesis (Illumina)

The sequencing by synthesis (SBS) offered by Illumina (formerly Solexa) is currently the most popular and widely used sequencing technology, with 90% of all published next-generation sequencing studies using this technology. It is of particular relevance to this thesis as Illumina sequencing has been used in the sequencing of all published snake whole genome sequences (Bradnam et al. 2013; Vonk et al. 2013; Castoe et al. 2013) and several snake venom gland transcriptomic studies (Rokyta et al. 2012; Margres et al. 2013; Aird et al. 2013).

First, sample DNA is fragmented and adapters are ligated to each end of the fragment, as with pyrosequencing (section 1.12.1). The sequencing libraries are then loaded onto a flow cell, which has oligonucleotides complementary to the library adapter sequences bound to its inside surface. In this way single-stranded DNA molecules are hybridised to the surface of the flow cell, allowing the access of enzymes to the molecule whilst also allowing chemical reagents to be passed through the flow cell. The bound strands then undergo bridge amplification using unlabelled nucleotides, creating up to a million local copies of the DNA strand, generating "clusters" of the same molecule on the flow cell surface (Figure 1.9).

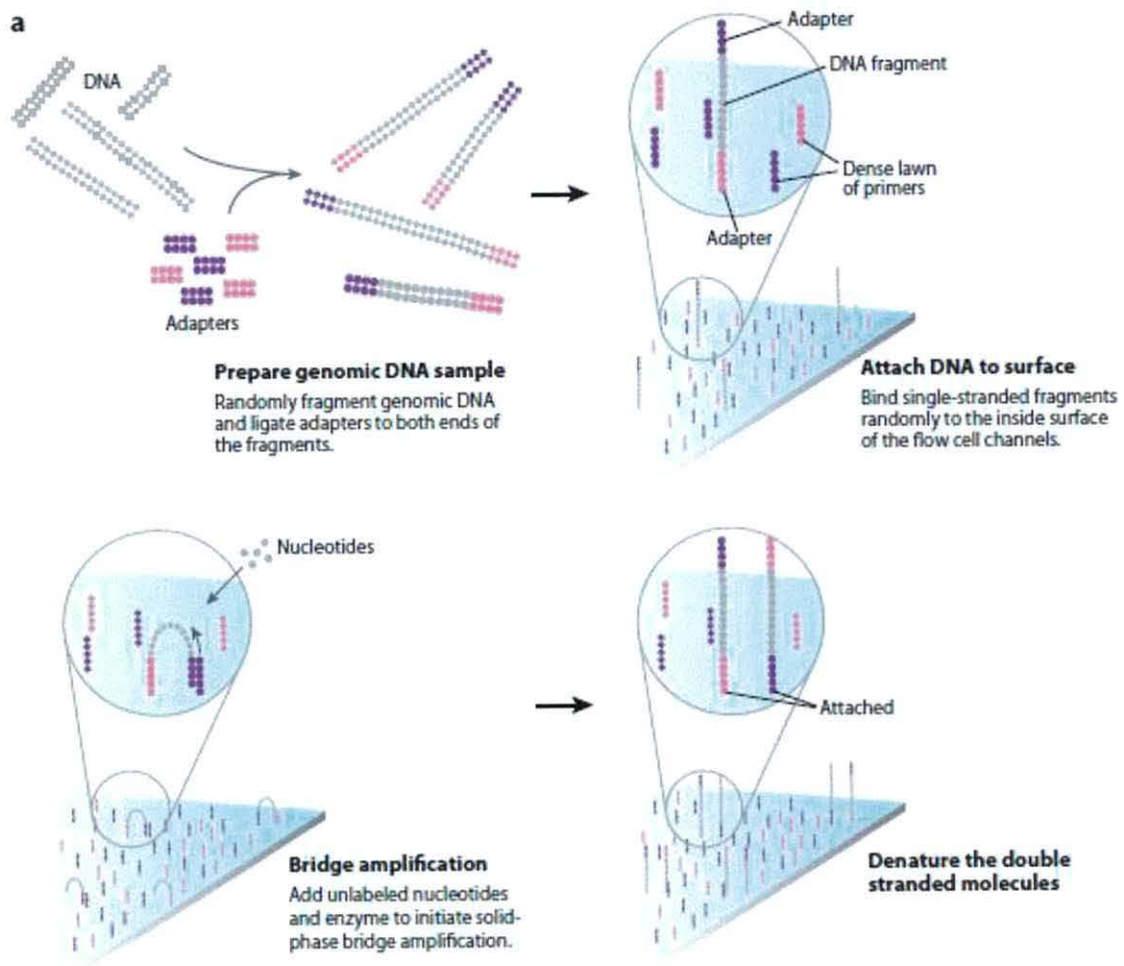


Figure 1.9. Adhesion of Illumina prepared libraries to the surface of the flow cell which then undergo bridge amplification to generate clusters. Figure taken from (Mardis 2008).

The sequencing-by-synthesis can then take place which is similar in concept to Sanger sequencing (section 1.8), as a sequencing primer is added to one end of each molecule along with the addition of fluorescently labelled nucleotides (which are base-specific). It is different to Sanger sequencing in the sense that each of these nucleotides contains a 3'-OH group terminator which halts any further polymerisation following incorporation, thus the reaction proceeds base-by-base. For the sequencing reaction, all four labelled nucleotides, sequencing primers and DNA polymerase are added to the flow cell and a laser is shone onto it. The light emitted indicates which base was incorporated onto each cluster, and recorded by a sensor. The blocking terminator and fluorescent tag are then removed from each incorporated base on the DNA strand and the reaction is then repeated (Figure 1.10) (Mardis 2008). Although the sensor used to record the fluorescence emission is not sensitive enough to detect the fluorescent signal

from a single molecule, the generation of clusters means that the signal for each molecule is greatly increased, thus overcoming this problem.

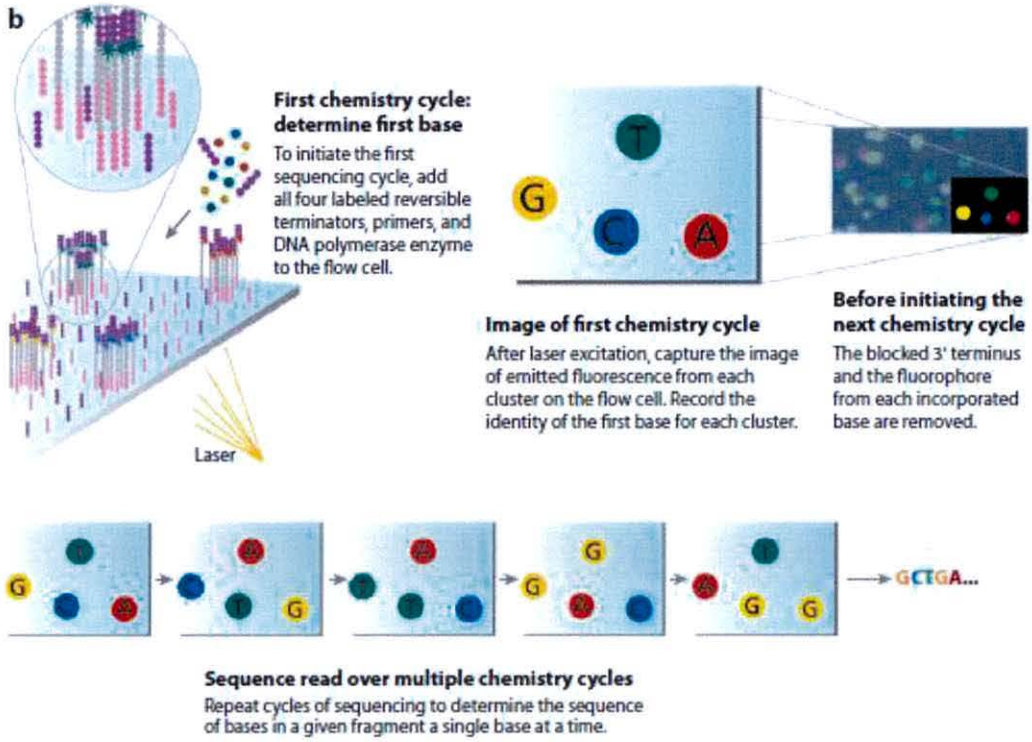


Figure 1.10. Illumina sequencing-by-synthesis method. Figure taken from (Mardis 2008).

Illumina offers a number of sequencing machines, each one designed to fulfil a specific application. For example the MiSeq is a low-coverage desktop sequencer designed for sequencing small genomes. Its current iteration with the latest Illumina reagents is capable of generating 25 million paired end reads with 2x300bp read length. The HiSeq (the latest version is the HiSeq2500) (Figure 1.11) is Illumina's ultra-high-throughput system designed for sequencing genomes and transcriptomes and is capable of running two flow cells with an output of up to 8 billion paired-end reads with 2x125bp read length. The most recent release, the NextSeq 500 can output up to 800 million paired-end reads with 2x150bp read lengths and is designed for medium-throughput studies such as whole genome and exome sequencing (all specifications listed were taken from the Illumina website (http://www.illumina.com/) which was last accessed on the 27/9/14).

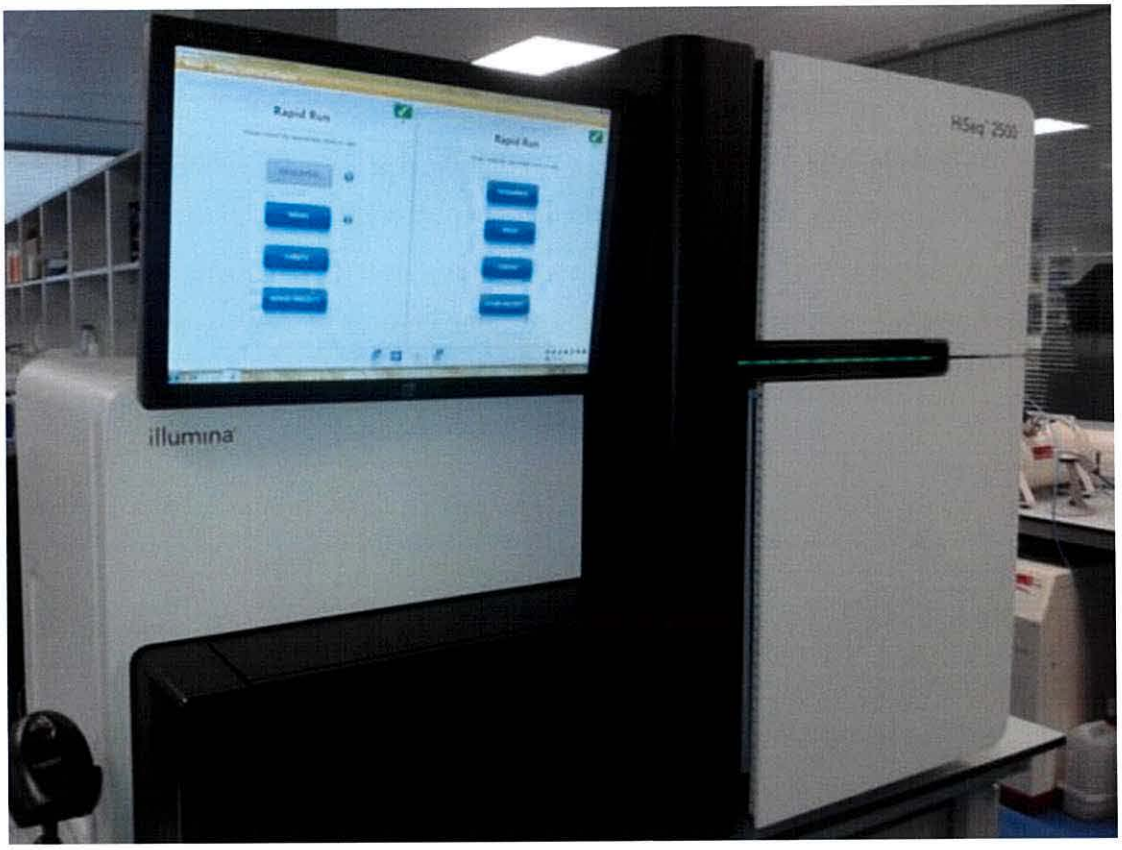


Figure 1.11. Photograph of the Illumina HiSeq2500 at the Institute of Biological, Environmental and Rural Sciences (IBERS) phenomics centre at Aberystwyth University.

1.13 Overview of this thesis

The use of venom as a means of prey capture represents a key evolutionary innovation in snakes, allowing the safe debilitation of prey items with minimal energy expenditure and without the need for large, muscular bodies. In terms of human impact, venomous snakes represent both killer and cure (with significantly more sway towards the former), constituting a huge medical burden globally but especially in the developing world. Without the development of new, more refined antivenoms, death and injury due to envenomation is likely to remain unchanged, with snakebite remaining a neglected tropical disease. Conversely, venoms also represent a potential goldmine of therapeutic compounds, and could aid the development of novel treatments for a range of diseases and medical conditions.

The evolution of venom in reptiles appears (based on the literature) to be extremely complex, comprising hundreds of proteins and peptides, originating from multiple recruitments of genes encoding non-toxic physiological genes from different body tissues into the venom gland. Venom is proposed to have evolved once at the base of squamate reptiles, with multiple

secondary losses occurring in a range of reptile taxa in favour of other prey capture methods such as constriction.

These ideas have been widely accepted for almost a decade, despite being based on low-coverage sequencing of only venom gland samples coupled with a lack of genomic resources for reptiles. With the advent of next-generation sequencing technologies, and the huge amount of data they can produce, it is now possible to re-examine these hypotheses of venom evolution in reptiles.

The first step in this process is to sequence and assemble both genomic (Chapter 2) and transcriptomic (Chapter 3) data. As there is a multitude of computing programs available to carry out these tasks, and no widely accepted "gold standard" methodology to do so, several approaches will be used and evaluated. The resulting data can then be used to assess current theories of venom evolution in reptiles, namely the Toxicofera hypothesis (Chapter 4) and the theory of venom gene recruitment (Chapter 5). The combination of genomic and transcriptomic data can also be used to investigate the transcriptional regulation of venom genes (Chapter 6), an area which has largely been neglected in previous studies. Finally, the implications of these re-evaluations are discussed (Chapter 7).

Chapter 2

De novo reptile genome assembly and optimisation

The potential applications of 2nd generation sequencing technology are numerous, with perhaps the most widely used being the sequencing of whole genomes. Until recently there has been a significant dearth of available genome sequences for reptile species, hindering studies into the evolution of this diverse lineage and amniote genome evolution as a whole. Confounding this problem is the fact that once genomic sequencing data has been generated, it must then be assembled, which poses many computational challenges and there is currently no "gold standard" in methodology. In an attempt to create a reference for future reptile (particularly snake) genomic studies, low coverage draft genomes of three species of snake were sequenced for the painted saw-scaled viper, *Echis coloratus*; the Egyptian saw-scaled viper, *Echis pyramidum* and the corn snake, *Pantherophis guttatus* using the Illumina sequencing platform. Multiple methods of *de novo* genome assembly were carried out and evaluated in order to determine the optimal approach to this bioinformatics endeavour. Several factors were obvious in their positive effect on genome assembly. Others however appeared to be specific to each data set, and therefore each genome assembly should be considered as unique and optimised independently. Evaluation of generated assemblies revealed that basic assembly metrics, such as the commonly used N50 value, are not sufficient to gain a thorough picture of assembly quality and that multiple methods are required for assessment. Analysis of the newly generated genome assemblies and three published snake genomes revealed that the addition of sequencing reads which confer long-range positional information to the assembly program, such as mate-pair data, can greatly improve overall assembly quality.

2.1 The genome

The term "genome" was first used by Hans Winkler in 1920 (Winkler 1920), and is believed to be a portmanteau of the words "gene" and "chromosome". However, it has been suggest that the suffix "-ome", referring to the totality of units contained within its prefix (in the case of genome, the total collection of genes), was Winkler's intended rationale behind the word (Lederberg and Mccray 2001; Gregory 2005). The genome can be defined as the entirety of an organisms DNA, including all of its genes and all non-coding regions. Historically, the size of an organism's genome was thought to be reflective of its development, with more "advanced" organisms presumed to have larger genomes. However, it was later found that some simple organisms had disproportionately large genomes, similar groups of organism showed diversity in genome size, and genome size tended to be higher than the predicted number of genes contained within it, leading to the "C-value paradox" (Thomas Jr 1971). Here C-value is a reference to the haploid size of an organism's genome (Swift 1950).

Traditionally genome sequencing projects have focussed on generating physical or genetic linkage maps in order to understand where DNA sequences are located across the whole genome (for example (Donis-Keller et al. 1987; Weissenbach et al. 1992)), the gaps between these regions can then be filled via sequencing. The dawn of shotgun sequencing (Chapter 1) has drastically changed this approach, potentially allowing whole genomes to be sequenced in a single sequencing experiment.

The choice of species to be sequenced requires careful consideration. Model organisms such as the mouse *Mus musculus* (Chinwalla et al. 2002), and pathogenic organisms such as *Mycobacterium tuberculosis* (Cole et al. 1998) have largely received preference in being sequenced due to their utility in medical research. The constant evolution of DNA sequencing technology means that the process of genome sequencing is rapidly becoming quicker, more accurate, and more cost effective, ultimately giving researchers free-reign to sequence their genome of choice. Since the sequencing of human genome (Venter et al. 2001) and other earlier genome projects such as the fruit fly *Drosophila melanogaster* (Adams et al. 2000), there has been an enormous increase in the amount of whole genomes sequenced from a wide diversity of organisms, including the mosquito *Anopheles gambiae* (Holt et al. 2002) and the puffer fish *Takifugu rubripes* (Aparicio et al. 2002), with many others in varying stages of completion.

In more recent years, plans for genome sequencing projects have become more and more ambitious, with the sequencing of ten thousand vertebrate genomes being proposed by the Genome 10K Community Of Scientists (G10KCOS) (Haussler et al. 2009) and a collaboration

between Genomics England, the English national health service and the Wellcome Trust planning and committing to the sequencing of 100,000 human genomes using the newly released Illumina HiSeq X 10 sequencing system.

2.1.1 Reptile genomes

At the time of commencement of this PhD project, the only whole genome sequence available for a non-avian reptile was that of the green anole lizard, *Anolis carolinensis* (Alföldi et al. 2011). However, in recent years more genome sequences have become available for a range of reptile species including the painted turtle (*Chrysemys picta*) (Shaffer et al. 2013), soft-shell turtle (*Pelodiscus sinensis*) (Wang et al. 2013), green sea turtle (*Chelonia mydas*) (Wang et al. 2013) and two snake species, the Burmese python (*Python molurus bivittatus*) (Castoe et al. 2013) and the king cobra (*Ophiophagus hannah*) (Vonk et al. 2013). It is likely that the ever-improving technology associated with DNA sequencing, and also the reduced financial cost of this, will result in many more reptile whole genome sequences being produced. Indeed, several more have already been proposed (Castoe et al. 2011a; St John et al. 2012) or completed (Card et al. 2014). An increase in genomic resources for the Reptilia will bridge the current gap between a dearth of reptile genome sequences in comparison to the abundance of Avian (for example chicken, *Gallus gallus* (Hillier et al. 2004), zebra finch, *Taeniopygia guttata* (Warren et al. 2010) and collared flycatcher, *Ficedula albicollis* (Ellegren et al. 2012)) and Mammalian (for example human, *Homo sapiens* (Venter et al. 2001), mouse, *Mus musculus* (Chinwalla et al. 2002), giant panda, *Ailuropoda melanoleuca* (Li et al. 2009)) genome sequences, thus allowing a full spectrum look at amniote genome evolution and the transition from living in water to land (Alföldi et al. 2011).

Reptile genomes show moderate variation in size, with the genome size of crocodiles and turtles being slightly larger than squamate genomes (Figure 2.1), and an overall average haploid genome size of 2.3Gb. Reptile genomes have been shown to possess several unique features which differentiate them from bird or mammal genomes. The whole genome sequence of the green anole lizard, *Anolis carolinensis* (Alföldi et al. 2011), was the first step in enabling a comparison between members of all three amniote lineages.

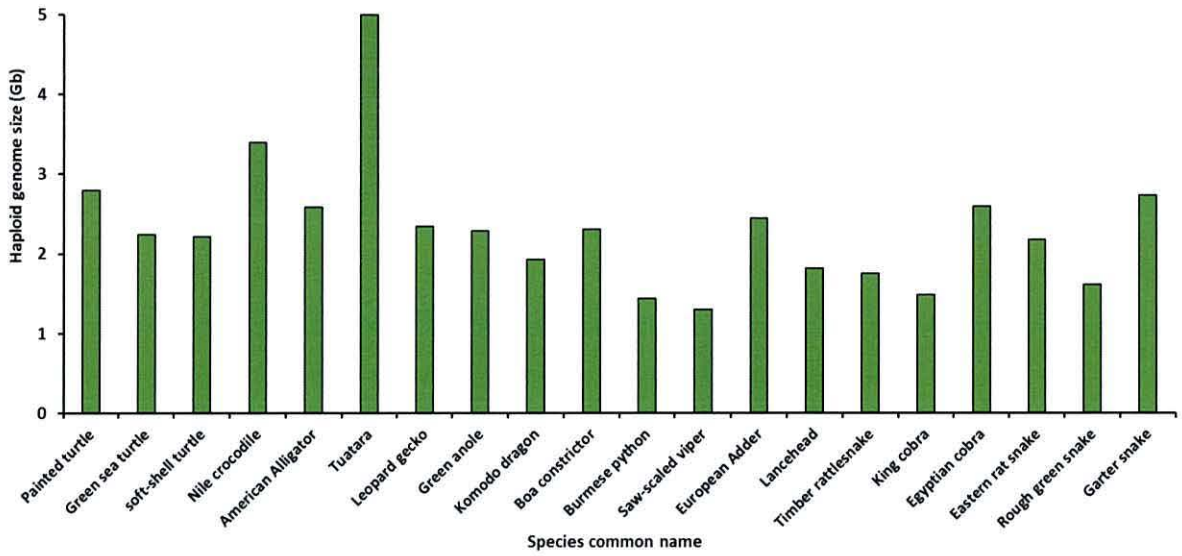


Figure 2.1. Estimated haploid genome sizes (or C-values) of key members of the Reptilia. Included species include members of the Testudines (painted turtle, *Chrysemys picta*; green sea turtle, *Chelonia mydas*; soft-shell turtle, *Pelodiscus sinensis*), Crocodylia (Nile crocodile, *Crocodylus niloticus*; American alligator, *Alligator mississippiensis*), Rhynchocephalia (Tuatara, *Sphenodon punctatus*), and members of the order squamata including representatives from the Gekkonidae (leopard gecko, *Eublepharis macularius*), Iguania (green anole lizard, *Anolis carolinensis*), Varanidae (Komodo dragon, *Varanus komodoensis*), and Serpentes (boa constrictor, *Boa constrictor constrictor*; Burmese python, *Python molurus bivittatus*; saw-scaled viper, *Echis carinatus*; European adder, *Vipera berus*; lancehead pit viper, *Bothrops atrox*; timber rattlesnake, *Crotalus horridus*; king cobra, *Ophiophagus hannah*; Egyptian cobra, *Naja haje*; Eastern rat snake, *Elaphe obsoleta*; rough green snake, *Opheodrys aestivus* and the garter snake, *Thamnophis elegans*). All C-values were obtained from the animal genome size database (www.genomesize.com).

Most notably the *Anolis* genome was found to be lacking isochores (large regions of the genome rich in Guanine and Cytosine (GC) content) (Fujita et al. 2011), with a homogeneous GC content throughout the genome unlike the genomes of birds and mammals (Alföldi et al. 2011). Additionally it was found to contain far more tandem repeats than the genomes of turtles and Archosaurs (Shedlock et al. 2007).

Reptile genomes are known to be highly repetitive, containing a high degree of transposable elements (TEs) and simple sequence repeats (SSRs) (Shedlock et al. 2007). For example, the genomes of squamates and the closely related Tuataras (*Sphenodon spp.*) contain a novel family of Short Interspersed elements (SINEs) known as Sauria SINEs (Piskurek et al. 2006). Analyses

have shown that the degree of genomic repeat content can vary between snake lineages (Castoe et al. 2011b), or that certain TEs are specific to certain lineages of snake (Kordiš and Gubenšek 1997). Most notably it was found that the more primitive Burmese python (*Python molurus bivittatus*) genome contains a low amount of genomic repeats similar to the genomes of birds, whereas the copperhead (*Agkistrodon contortix*), a venomous pit viper, contains a much higher amount similar to mammalian genomes (Castoe et al. 2011b). As the presence and expansion of TEs throughout the genome can affect gene regulation and duplication (Levinson and Gutman 1987; Hurles 2004), it is possible that the increase of them in the genomes of venomous snakes means that their genomes are "primed" to gene duplication which is thought to be a fundamental mechanism in the origin and expansion of venom genes (Chapter 5).

Previous cytogenetic studies and a more recent genomic study (Vicoso et al. 2013) have found that sex chromosome heteromorphism (the degree to which sex chromosomes differ to each other) is variable between groups of snakes, with members of the Boidae (such as boas and pythons) having homomorphic (morphologically identical) sex chromosomes, but more advanced snakes such as colubrids and vipers having completely heteromorphic (morphologically distinct) sex chromosomes.

The sequencing of the Burmese python (*Python molurus bivittatus*) genome (Castoe et al. 2013) was carried out in order to investigate the extreme phenotypic and physiological adaptations of snakes. Indeed, this study had several interesting findings. It was found that genes involved in metabolism have undergone positive selection in snakes which, coupled with previous findings that snake mitochondrial genomes have been significantly re-structured (Castoe et al. 2008), suggests extensive modification of the metabolic pathways in snakes allowing them to ingest very large prey items intermittently (Secor and Diamond 1998; Secor 2008). Variation within multigene families was also discovered, with an expansion of olfactory receptor, vomeronasal receptor and ephrin-like genes indicating a genetic basis behind the enhanced chemoreception displayed by snakes. Concurrently, the loss of opsin genes in the Burmese python and king cobra genomes is supportive of the hypothesis that snakes were once fossorial (Vidal and Hedges 2004) and thus selection for light perception was not maintained (Castoe et al. 2013). Analysis of the repeats/TEs found in the genome of this and ten other species found that, whilst the amount of repeat content within genomes varied between species, the types of repeats were relatively constant except for a family of CR1 LINEs (Long Interspersed Nuclear Elements) which were only detected in the genomes of advanced snakes. Finally, unlike the *Anolis* genome, snakes do possess GC isochores but less so than Archosaurs and mammals (Castoe et al. 2013).

The genome sequence(s) assembled for the boa constrictor (*Boa constrictor constrictor*) were assembled as part of the Assemblathon 2 competition (Bradnam et al. 2013) and have not (yet) been utilised to investigate characteristics of the genome of this species.

The king cobra (*Ophiophagus hannah*) genome paper (Vonk et al. 2013) appears to be more focused on the associated transcriptomic (both mRNA and miRNA) and proteomic analyses carried out, rather than the genome sequence itself. This study does seem to suggest however that genes from several venom toxin gene families (namely snake venom metalloproteinases, cysteine-rich secretory proteins and lectins) are clustered on genomic scaffolds, hinting that gene duplication through unequal crossing-over during homologous recombination may be involved in the expansion of toxin gene families (Chapter 5).

It is apparent that reptiles represent a unique group of animals, both in terms of their sometimes extreme phenotypic manifestations and the structure and content of their genomes. Many new discoveries have been made by sequencing a small number of reptile genome sequences, suggesting that the inclusion of more species will only add to the intrigue towards, and the understanding of, these weird and wonderful animals.

2.2 Genome study species

Any genome sequencing project requires a logical rationale, both in terms of the species chosen and the sequencing strategy. In this way, the utility of the resulting genome sequence is maximised, making it a worthy endeavour for both the required research effort and the cost involved. Ideally the species should be chosen with a biological question (or questions) in mind, be relatively easy to obtain for sampling, and the individual sequenced should be representative of the entire genome (for example have all sex chromosomes).

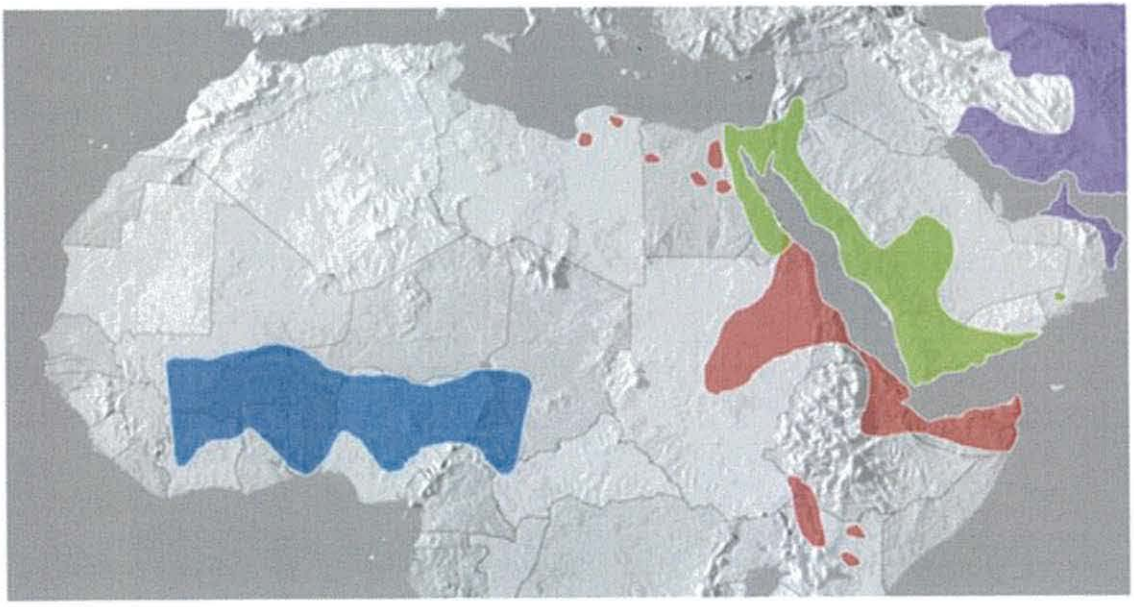
The recently sequenced genome of the king cobra (*Ophiophagus hannah*) (Vonk et al. 2013) represents a poor example of a well-planned sequencing strategy, for several reasons. Firstly, the king cobra is the world's largest venomous snake (Vonk et al. 2013) , reaching in excess of 4-5 metres in length. As such it is difficult to maintain and handle safely in captivity, especially in large numbers. The king cobra is the sole member of the genus *Ophiophagus*, and its inclusion in a clade with other cobra genera (including *Naja*, *Aspidelaps* and *Hemachatus*) is not supported by phylogenetic analysis (Wüster et al. 2007), therefore there is little utility of this genome sequence in comparative studies between cobra species. The sequenced genome was generated from a male specimen which is homogametic (ZZ), and so an entire W chromosome has not been sequenced. Finally, the transcriptomic data generated for this species

was derived from a different specimen to the genome animal (making genome annotation with this data more difficult) and was sequenced from pooled tissue samples, making the allocation of transcripts to a particular tissue impossible.

With this in mind, several factors were considered when choosing study species for this work, namely that species chosen should be easy to maintain in large numbers and easily obtainable, all specimens for genome sequencing should be female, tissues for transcriptome sequencing should be preferentially obtained from the genome animal (Chapter 3), and the species should be relevant to a biological question. The details for the choice of study species is outlined below.

2.2.1 Saw-scaled vipers of the genus Echis

The saw-scaled vipers of the genus *Echis* are small, venomous vipers found distributed across North Africa, the Middle East, Sri Lanka, Pakistan and India (Pook et al. 2009) (Figure 2.2).



E. ocellatus – blue
E. pyramidum – red
E. carinatus – purple
E. coloratus – green

Figure 2.2. Geographic distribution of the four main clades of the *Echis* species complex. This figure is based on figure found in (Pook et al. 2009).

The name "saw-scaled" is a reference to the serrated keeled scales possessed by these species, which produce a hissing sound when the snake stridulates (rubs its body against itself to produce

sound) as a threat display (Escoriza et al. 2009). The taxonomy of this genus has received significant revision over the years, with the most recent phylogenetic analyses (Barlow et al. 2009; Pook et al. 2009) finding four supported clades within this species complex (Figure 2.3).

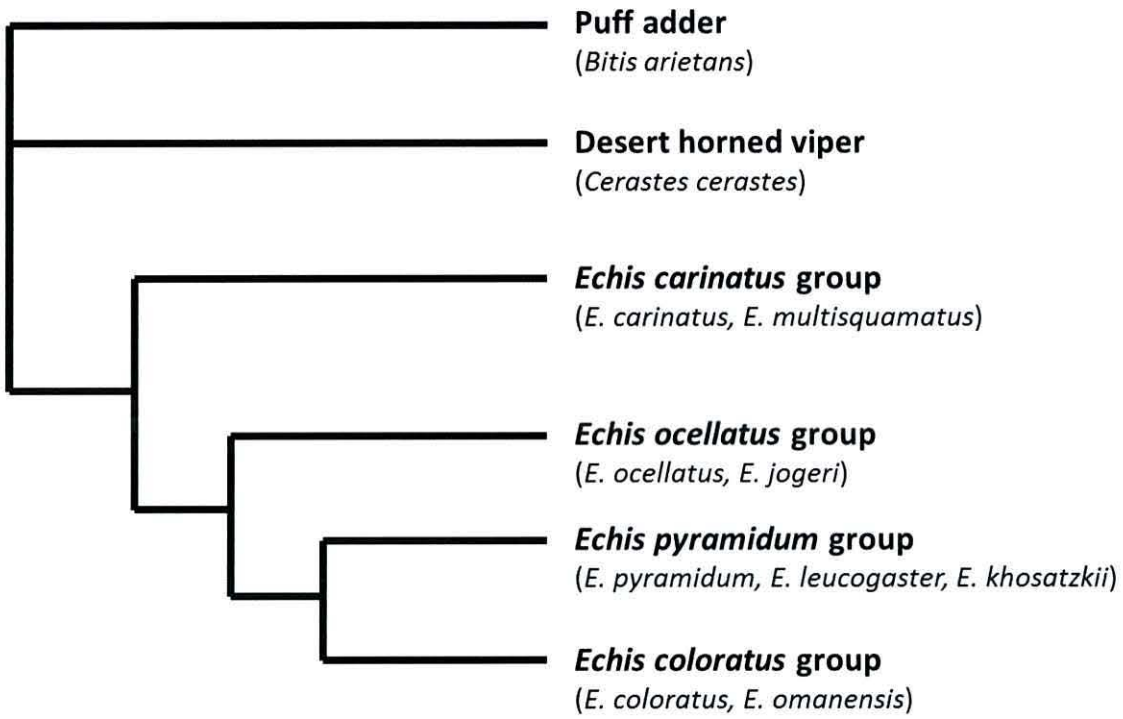


Figure 2.3. Cladogram of the four main monophyletic *Echis* species groups based on a phylogenetic analysis of four mitochondrial (cytochrome b, NADH dehydrogenase subunit 4, 12s rRNA and 16s rRNA) and one nuclear (recombination activating gene 1) genes described in (Barlow et al. 2009).

The genus *Echis* presents itself as an ideal laboratory model for several reasons. Firstly, the small size of these species make them easy to keep in captivity in large numbers, and are relatively safe to handle and easy to obtain. Venom components from *E. carinatus* are already used for pharmaceutical applications (see Chapter 1), supporting further investigation into the venom of these species for medical uses. Most notably, the venom composition of *Echis* species has been shown to display interspecific variation (Casewell et al. 2009), which is possibly reflective of an adaptation to the diet of each species (Barlow et al. 2009; Richards et al. 2012). This variation could be caused by factors at the genome level (e.g. a differing presence or absence of venom toxin genes between species), the transcriptome level (e.g. different genes are expressed in different species or the expression level of genes show interspecific

differences) or the proteome level (e.g. post-translational modification varies between species) which in turn has implications for antivenom efficacy (Fry et al. 2003). For this study the two sister taxa *E. coloratus* (which primarily feeds on vertebrates) (Figure 2.4) and *E. pyramidum* (which feeds mostly on invertebrates) (Figure 2.5) (Barlow et al. 2009; Richards et al. 2012) were chosen as genome study species.

Figure 2.4. Photograph of a juvenile painted saw-scaled viper, *Echis coloratus*

Figure 2.5. Photograph of an adult Egyptian saw-scaled viper, *Echis pyramidum*. Photo taken by R. Morgan and used with permission.

Genome size in the Viperidae ranges from 1.3Gb to 2.71Gb, with the average genome size being 2.06Gb (based on available data) (Figure 2.6). Within the Viperinae ("true vipers") the average genome size is 2.05Gb, with that of *Echis carinatus* predicted to be 1.27Gb (Desmet 1981). The genomes of *E. coloratus* and *E. pyramidum* are therefore assumed to be 1.3Gb in size.

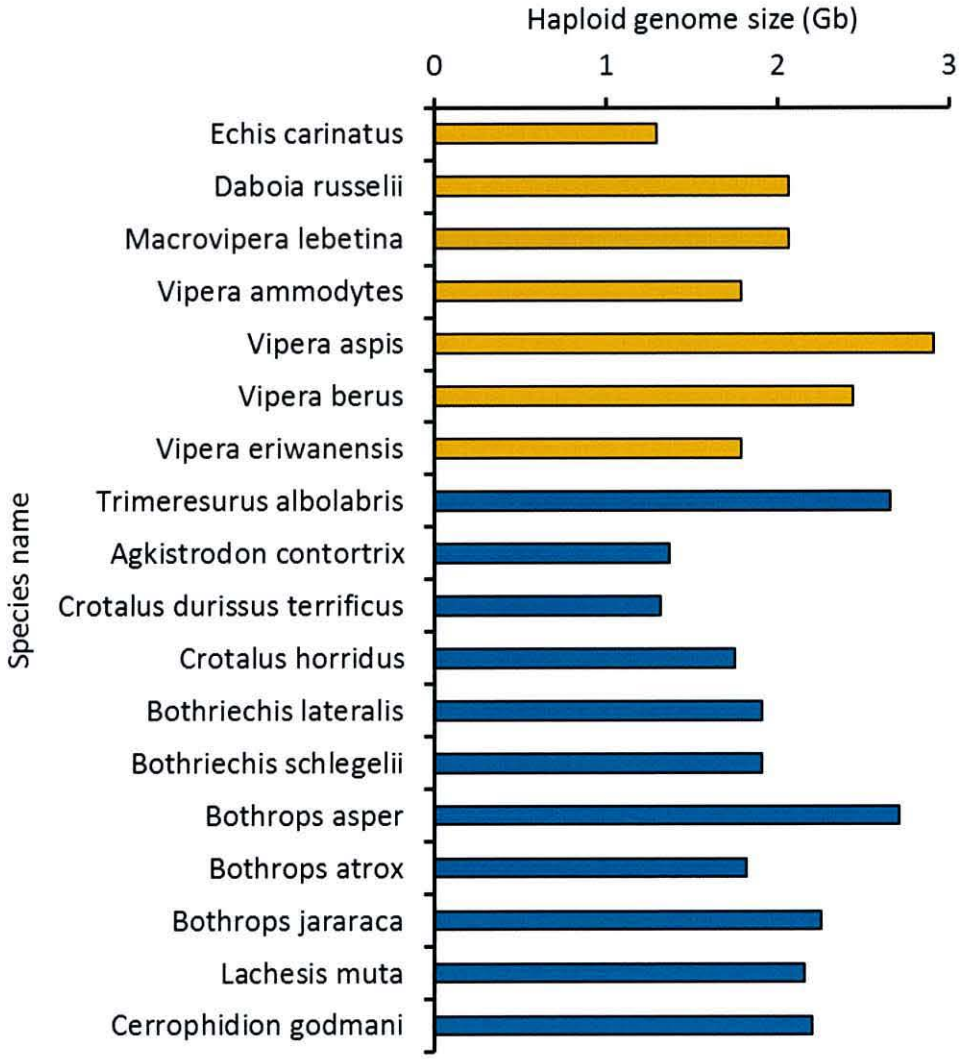


Figure 2.6. Haploid genome size of species within the Viperidae. "True vipers" (members of the Viperinae) are shaded in yellow and pit vipers (members of the Crotalinae) are shaded in blue. All genome size data was obtained from the animal genome size database (www.genomesize.com).

2.2.2 The corn snake (Pantherophis guttatus)

The corn snake, *Pantherophis* (formerly *Elaphe* (Pyron and Burbrink 2009)) *guttatus*, is a member of the Colubridae ("colubrids") and is commonly sold in the commercial pet trade. Their availability, ease of keeping in captivity and ease to breed make corn snakes a prime snake species to use as a research model animal. Previously, corn snakes have been used in developmental studies, both to investigate Hox gene expression in snakes (Woltering et al. 2009; Di-Poï et al. 2010; Liang et al. 2011; Mansfield 2013) and to investigate an increased rate of somitogenesis (Gomez et al. 2008; Vonk and Richardson 2008; Gomez and Pourquié 2009) to determine their effect on the loss of limbs and the change in the body plan of snakes (Woltering 2012). Additionally, this species has been selectively bred to create pigmentation and pattern mutants (colloquially referred to as "morphs") for many years, meaning mutant specimens to investigate pigmentation pattern development in snakes are already widely available. More specifically, these mutations are likely to either affect the migration and differentiation of neural crest cells giving rise to pattern mutants, or the biosynthetic pathway of a particular pigment such as melanin, giving rise to phenotypic traits such as amelanism (Figure 2.7).

Figure 2.7. Photograph of an adult amelanistic corn snake (*Pantherophis guttatus*).

The genome size of colubrids ranges from 1.43Gb to 2.73Gb, with an average haploid genome size of 2.11Gb (Figure 2.8), giving them a slightly larger genome size than vipers. The genome

size of the corn snake is predicted to be 1.965Gb based on the mean of the four *Elaphe* species estimated genome sizes on the Animal genome size database (www.genomesize.com).

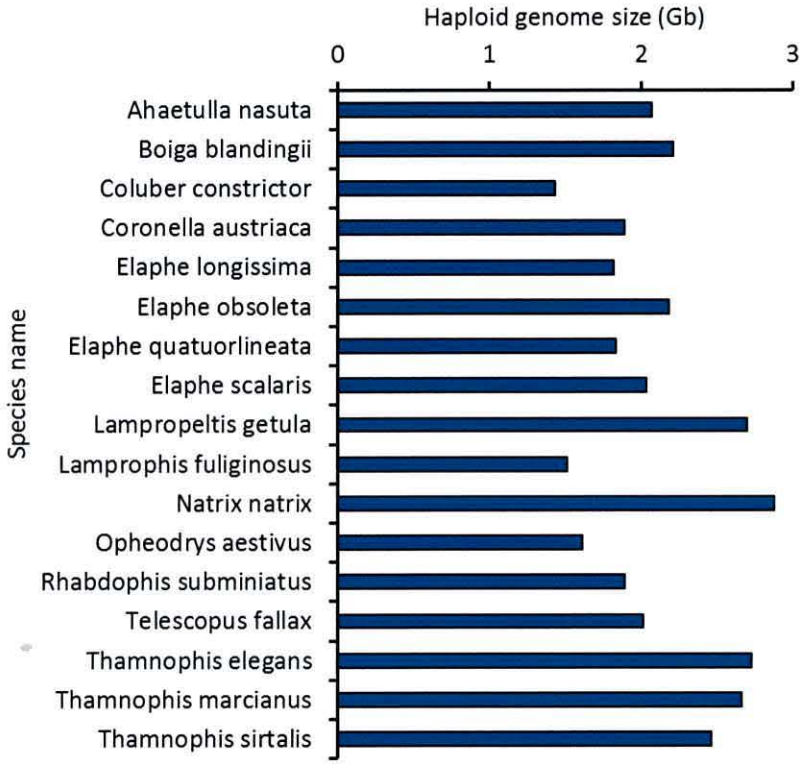


Figure 2.8. Haploid genome size of members of the Colubridae. All genome size data was obtained from the animal genome size database (www.genomesize.com).

2.3 Genome assembly

2.3.1 What is an assembly?

Once genomic DNA libraries have been sequenced to give sequencing "reads" (see Chapter 1), they must then be assembled. An assembly can be defined as:

"*...a hierarchical data structure that maps the sequence data to a putative reconstruction of the target*" (Miller et al. 2010).

In simpler terms, an assembly is a jigsaw puzzle with millions of pieces, and there is no image of what the final puzzle should look like on the box. However, many of the pieces will overlap with each other, informing where they should be placed. As discussed in Chapter 1, overlapping

sequencing reads are grouped into lengths of contiguous sequence (or "contigs"), which represent the consensus sequence of a region of DNA. Contigs can further be assembled into scaffolds, which are longer stretches of contiguous sequence whose constituent contigs are orientated and spaced apart based (usually) on sequencing reads from mate-pair libraries (Miller et al. 2010).

Due to the enormous amount of data produced from modern sequencing technologies, this process must be carried out computationally using a dedicated assembly computer program. There are a considerable number of short-read assembly programs (referred to from now on as "Assemblers") available, examples of which include Velvet (Zerbino and Birney 2008), ABySS (Simpson et al. 2009), SOAPdenovo2 (Luo et al. 2012), ALLPATHS-LG (Gnerre et al. 2011), CLC (CLC bio, http://www.clcbio.com/) and SGA (Simpson and Durbin 2012). The majority of genome assemblers are based on a De Bruijn graph-based algorithm (SGA is the exception to this in the list of assemblers mentioned previously), which are explained in the following section.

2.3.2 Assembly algorithms- De Bruijn graphs and k-mers

A De Bruijn graph is a concept used in graph theory, but has been adapted for use in genome assembly. It can be defined as a directional graph (a set of nodes connected by edges which have a direction associated with them) (Khan and Kamal 2014) representative of overlaps between sequences. A commonly used example is that of balls connected by directional arrows (Figure 2.9). Each edge represents a connection (or "path") between one node to the other, which will increase in length depending on the number of connections between subsequent nodes.

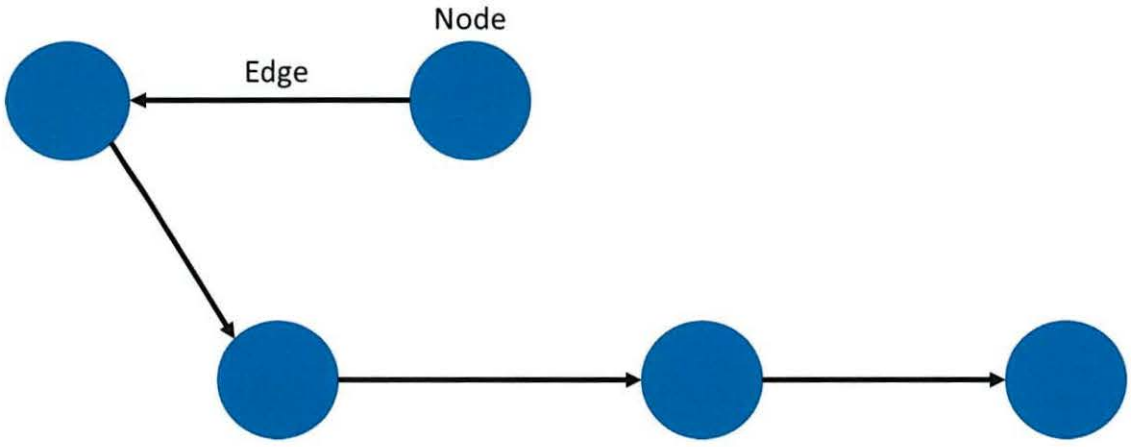


Figure 2.9. Graphical representation of the formation of a De Bruijn graph, with "nodes" being linked to each other by directional "edges".

In genome assembly, a De Bruijn graph is constructed, based on k-mers from sequencing reads, and the genome assembly is then constructed from the De Bruijn graph. k-mers must overlap by k-1 in order to form a connection (an edge) between nodes (Figure 2.10). In short, the graph is extended by k-mers which overlap and are identical apart from their first or last position (Khan and Kamal 2014).

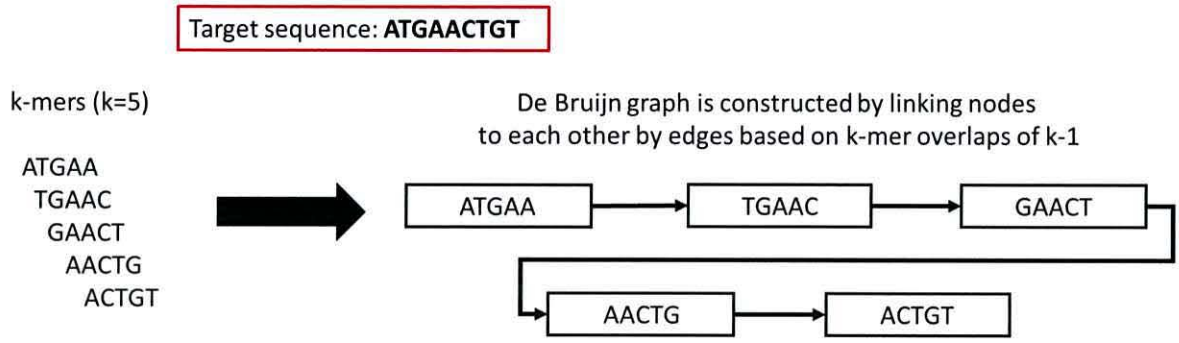


Figure 2.10. De Bruijn graph formation from k-1 overlapping k-mers to assemble a target sequence.

A k-mer can be defined as any number of sub-sequences of length *k* (which is defined by the user) contained within a DNA sequencing read (Compeau et al. 2011). For example, if a 100bp sequencing read were considered on its own, it would represent a 100-mer i.e. a k-mer where *k*=100. However, due to the asymmetrical nature of shotgun sequencing (i.e. not all regions will be sequenced due to compositional bias, GC-rich regions etc.) not all sections of the genome will be present as 100-mers. As such, the value of *k* is chosen as a smaller value, in order to increase the probability that all smaller sized k-mers will be present. Put simply, if considering two random 100bp reads, it is more likely that there will be an overlap between them of a smaller number (for example 4, based on a k-mer value of 5 (k-1 overlap)) than of a larger number (for example 99 if considering k=100). Therefore, assemblies using a smaller k-mer value will result in more connections formed between sequencing reads (and therefore more contigs), but a larger k-mer value will result in fewer but longer, more accurate contigs.

The use of De Bruijn graphs does pose some challenges, especially when assembling sequence data derived from DNA. The sheer amount of data produced by next generation sequencing means that the construction of the graph must be done efficiently, otherwise in terms of the computing memory required, it would be an impossible task. Next-generation sequencing

inherently leads to sequencing errors contained within the data (Miller et al. 2010). Sequencing errors can disrupt the De Bruijn graph process in two ways: errors within k-mers can lead to the formation of incorrect nodes in the graph, and this can result in wasted computing memory making the assembly process longer (Compeau et al. 2011). There are several ways to overcome the issues caused by sequencing errors in short read data. Firstly, reads can be pre-processed in order to correct or fully remove sequencing error, either by using a dedicated program such as Quake (Kelley et al. 2010), or by the assembler itself (for example ALLPATHS (Butler et al. 2008)). Secondly, many assemblers use the original sequencing reads as weighted support for the formation of edges between nodes during De Bruijn graph construction, with poorly supported edges being collapsed or "eroded" (Zerbino and Birney 2008). The over- or under-representation of genome regions in a set of DNA sequencing reads can lead to issues when constructing De Bruijn graphs. In particular, genomic repeats can lead to the formation of multiple poorly supported edges which are consequently eroded. As a result, repetitive regions of a genome can be difficult to sequence by short-read shotgun sequencing. Finally, assembly using this method is highly computationally intensive, meaning it must be as efficient as possible and requires special computer hardware, especially if there is a large dataset of reads.

2.3.3 Assemblers used in this chapter

In this chapter, two genome assembly programs will be used: the CLC genomics workbench (CLC bio) and ABySS (Assembly By Short Sequences) (Simpson et al. 2009).

The CLC assembler was used most notably in the assembly of the king cobra genome (Vonk et al. 2013), and so stands as a benchmark assembler for snake genomes. It utilises a De Bruijn graph approach for assembly and has several features which make it a popular choice. Firstly, the program is implemented through a GUI (Graphical User Interface) meaning that running an assembly can be achieved through a click of a button, and is much easier than running other assemblers which must be run via the Linux command line. The run time for an assembly is also exceedingly quick. An additional advantage is that CLC can assemble sequencing data of different read lengths from multiple sequencing platforms including Illumina, Sanger, Roche 454 and Life technologies SOLiD and Ion Torrent, and a hybrid assembly can be generated from all of these data types simultaneously (CLC bio, http://www.clcbio.com/products/clc-genomics-workbench/). The downside however is that this assembly program is not freely available and is subject to an annual license fee.

The assembler ABySS is a short-read *de novo* assembler also based on a De Bruijn graph algorithm. Because ABySS implements MPI (Message Passing Interface), its processes can be run in parallel on multiple CPUs (Central Processing Units), meaning it is capable of assembling mammalian-sized genomes efficiently (it was trialled on the human genome (Simpson et al. 2009)). Additionally, ABySS can perform the scaffolding of contigs without using any additional software, and so is ideal for generating *de novo* genome assemblies.

2.3.4 A good assembly- a matter of semantics?

Whilst there is currently no gold standard method in producing a good genome assembly, there is also no definitive way of evaluating how good an assembly actually is. Ideally, a perfect assembly would contain only two sequences (one for each strand of DNA), both being the exact length of the genome with no gaps, and zero sequencing errors. Based on current technology this is simply impossible, especially for large and repetitive genomes. Traditionally, the contig or scaffold N50 was considered to be the main metric on which to judge an assembly. The N50 can be defined as the contig/scaffold length at which all contigs/scaffolds of that length or longer represents at least 50% of the total length of all contigs/scaffolds (Miller et al. 2010). Therefore 50% of the entire assembly will be contained within contigs/scaffolds of this length or greater, meaning a higher N50 value is indicative that the assembly contains a higher proportion of longer contigs/scaffolds. A similar metric, the NG50 has been proposed (Earl et al. 2011), based on the fact that the N50 is representative of the size of the *assembly* and not the size of the sequenced *genome*. This means that comparisons of assemblies generated from different organisms is possible. Other commonly used metrics to evaluate a genome assembly include the number of contigs/scaffolds in the assembly (the smaller the number, the better the assembly) and the maximum contig/scaffold length.

Two events have now been held in the form of a competition in order to evaluate *de novo* assembly methods and the metrics and methods used to evaluate the generated assemblies, namely the Assemblathon 1 and 2 (Earl et al. 2011; Bradnam et al. 2013). The Assemblathon 1 had 17 teams of researchers assemble synthetic data from a simulated genome using their assembly method of choice, with assemblies subsequently being evaluated. However, whilst this produced an extensive catalogue of methods and metrics reflecting their performance with which to compare future genome assembly approaches, this assessment was ultimately limited by the use of a relatively small simulated dataset (Earl et al. 2011). In response to this the Assemblathon 2 was held, this time using paired-end and mate-pair library Illumina data

derived from a lake Malawi cichlid (*Maylandia zebra*, Fish), a budgerigar (*Melopsittacus undulatus*, Bird) and a boa constrictor (*Boa constrictor constrictor*, Snake). This study gave a great deal of insight into *de novo* genome assembly in several ways. Firstly, it highlighted the enormous abundance of *de novo* genome assemblers available (21 were trialled). Second, over 100 different metrics were used to assess genome assemblies, each one providing meaningful information about aspects of the genome assembly, but ultimately demonstrating that judging an assembly solely on its N50 value is not sufficient. And finally, based on all of these metrics, there was still no "winning" assembly program/method (except for Snake which was the assembler SGA (Simpson and Durbin 2012)) showing that each assembly is unique and each metric indicates a specific thing about it. As such, the purpose of the assembly (e.g. gene discovery, Single nucleotide polymorphism (SNP) detection, analysis of genome rearrangements) should be carefully considered and the metrics relevant to that purpose should be used to inform how to achieve the "best" assembly as the end result.

This chapter aimed to sequence low-coverage draft whole genome sequences for three species of snake: the painted and Egyptian saw-scaled vipers and the corn snake. Inspired by the Assemblathon 2 competition, a variety of assembly and evaluation methods was then used to try and determine the optimal approach to generating a *de novo* draft whole genome sequence for a snake. Whilst only the Illumina sequencing platform was used due to its cost-effectiveness, ease of sequencing library preparation, and high output; two different sequencing machines were used with different library insert sizes and different read lengths. Coupled with the ability to use sub-sets of this sequencing data, and alter parameters of the assembly programs, it was possible to gain an insight into several aspects of genome sequencing and assembly. As the CLC Genomics workbench assembler has been used previously to assemble a snake genome, it was considered here as a benchmark to beat. ABySS offers much more versatility, and so was used to test various parameters to try and determine the best approach to the *de novo* sequencing of snake genomes in the hopes of providing a reference for future snake genome sequencing projects. Read trimming, single- or paired-end sequencing, read length, library insert size and k-mer size were all assessed. Assemblies were then evaluated using basic metrics along with others suggested by the Assemblathon 2. Finally, the newly generated whole genome sequences were compared to the three currently available snake genome sequences.

2.4 Methods

2.4.1 Tissue sampling

All research involving animals was carried out in accordance with institutional and national guidelines and was approved by the Bangor University Ethical Review Committee. All study animals were sacrificed according to Schedule 1 procedures as stipulated in The Animals (scientific procedures) Act 1986. Where possible all body tissues were dissected and preserved for use in future experiments in accordance with the 3Rs (Replacement, Reduction and Refinement) which encourage and offer guidelines to minimise the use of animals in scientific research (Guhad 2005). All tissue samples were snap frozen immediately in either liquid Nitrogen or dry ice and stored at -80°C until required.

All three genome animals were adult female (heterogametic ZW) specimens to allow the sequencing of a full complement of chromosomes. The painted saw-scaled viper (*Echis coloratus*) and the Egyptian saw-scaled viper (*Echis pyramidum*) were wild caught from Israel and Egypt, respectively. The corn snake (*Pantherophis guttatus*) specimen was a captive-bred specimen obtained through the commercial pet trade in the UK.

2.4.2 Genomic DNA extraction and Quality control

Genomic DNA extractions were carried out using Phenol/Chloroform/Isoamyl alcohol (Fisher Scientific) according to the manufacturer's protocol and eluted into 30µl of 1M Tris-EDTA (TE). All DNA samples were extracted from muscle samples.

Prior to library preparation, genomic DNA samples were assessed for integrity and the absence of inhibitory compounds by agarose gel electrophoresis and performing a test restriction digest using the restriction enzyme EcoRI and NotI (Promega) according to the manufacturer's instructions (an example is shown in Figure 2.11). All genomic DNA samples were treated with 5µl Ribonuclease A (Sigma-Aldrich) and then re-precipitated in order to digest any RNA molecules present in the extractions.

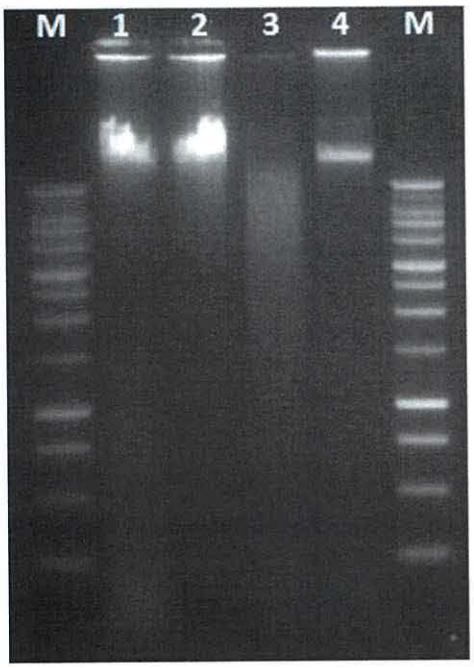


Figure 2.11. Agarose gel of genomic DNA and test restriction digests from *Echis coloratus*. M, DNA size marker. Lane 1 is untreated genomic DNA extract, Lane 2 is RNase treated genomic DNA which shows less smearing at lower molecular weights, Lane 3 is a test restriction digest using EcoRI and Lane 4 is a test restriction digest using NotI.

DNA samples were quantified using the Qubit fluorometer and the Nanodrop spectrophotometer to give DNA sample concentration and the 260/280 ratio (as a marker of purity). Samples were then diluted to contain 10µg of DNA in 50µl of TE as per specifications required by the GenePool sequencing centre at the University of Edinburgh (http://genepool.bio.ed.ac.uk/) which is now a part of Edinburgh Genomics (https://genomics.ed.ac.uk/). Samples were re-quantified to ensure the correct concentration of genomic DNA.

2.4.3 Genomic library preparation and sequencing

To assess and evaluate potential problems which could arise (such as a high GC content or high amount of repetition) during sequencing, initial libraries were prepared at the IBERS (Institute of Biological, Environmental and Rural Sciences) at Aberystwyth University under the supervision of Dr. Justin Pachebat. Libraries were prepared using the Illumina paired-end sample preparation kit, the iteration of Illumina reagents before the TruSeq sample preparation kits. Two libraries with insert sizes of 200-300bp and 500-600bp were made for the corn snake, *P. guttatus*. An overview of the Illumina library sample preparation workflow can be seen in

Figure 2.12 (the diagram is based on the TruSeq library preparation guide, but all major steps are essentially the same).

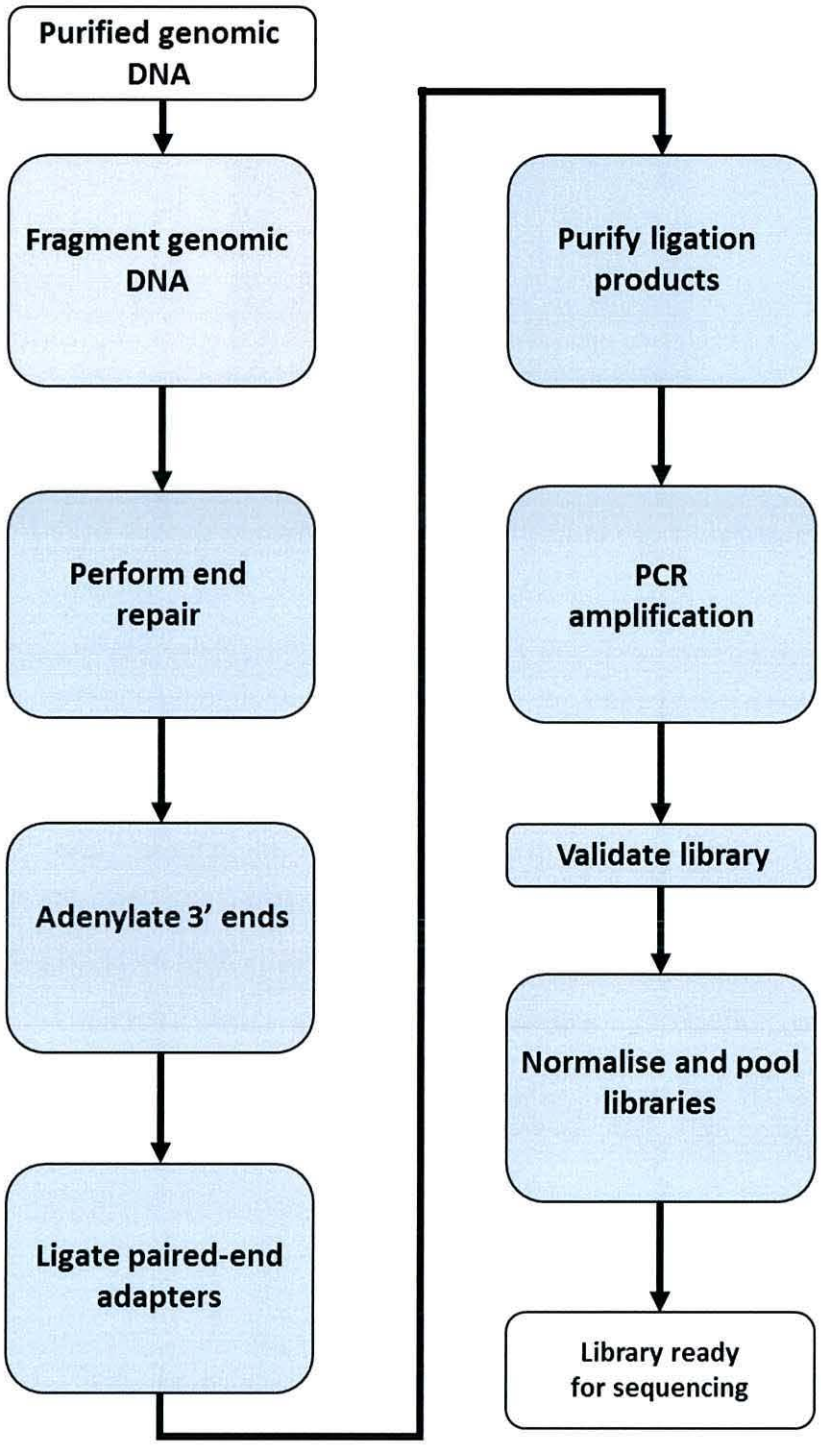


Figure 2.12. Workflow diagram of Illumina DNA sequencing library sample preparation. Figure is adapted from figure 29 of the Illumina TruSeq DNA sample preparation guide.

Firstly, genomic DNA was fragmented by sonication using a Diagenode Bioruptor standard for 15 cycles (30 seconds on, 30 seconds off), followed by purification using a PCR purification kit (Qiagen). The ends of fragments were then repaired to remove overhangs and result in blunt ends. The 3' ends of the DNA fragments were then adenylated (an adenine is added to them) in order to prevent self-annealing during subsequent steps, which could lead to the formation of chimeric molecules. Sequencing adapters were subsequently ligated to each end of the DNA fragments, and the library samples were run out on an agarose gel. Gel sizes between 200-300bp and 500-600bp were taken and then extracted using the gel extraction kit (Qiagen) and purified using the PCR purification kit (Qiagen). Genomic libraries were then enriched by 14 cycles of polymerase chain reaction (PCR) according to the Illumina protocol in order to enrich for fragments which had adapters ligated to them. After PCR purification was performed again using the Qiagen PCR purification kit, and libraries were again run out on an agarose gel, extracted and purified. The final libraries were then quantified using a Nanodrop spectrophotometer and stored at -20°C.

Blunt-end ligations were set up using pUC19 linearised with SmaI (Thermo Scientific) as a cloning vector according to the manufacturer's guidelines and incubated at 4°C overnight. Ligations were transformed by heat shock into NovaBlue Singles competent *E. coli* cells (Novagen) at a ratio of 10µl ligation to 50µl competent cells. 250µl SOC medium was added to each sample and they were then incubated at 37°C for 30 minutes. Samples were then plated onto XIA (Xgal, IPTG and Ampicillin) plates and incubated at 37°C overnight. The following day a total of 46 white colonies were picked and used to inoculate 3ml of LB broth with Ampicillin per reaction, which were incubated at 37°C with shaking overnight. Plasmid DNA was then extracted and purified using the QIAprep spin miniprep kit (Qiagen), eluting into 50µl EB (elution buffer). Samples were then sent to be sequenced on the ABI 3130 capillary sequencer at the Department of Zoology at the University of Oxford.

Analysis of the resulting sequences using the NCBI (National Centre for Biotechnology Information) VecScreen tool (http://www.ncbi.nlm.nih.gov/tools/vecscreen/) revealed that only 24 of the samples contained insert DNA. These sequences were analysed using BLAST, only 9 sequences had any significant BLAST hits, and only 5 of these matched to a reptile sequence. In particular only one 600bp fragment from the 500-600bp library was cloned and sequenced correctly. Optimisation was attempted with the addition of T4 PNK (Polynucleotide Kinase) (Promega) prior to ligation, but the number of transformants for the 500-600bp library was still minimal (1-2 white colonies). The mean GC content of the successfully sequenced libraries was found to be 41.96%.

It became apparent that this methodology would be very inefficient for genome sequencing, especially as the larger insert size library was proving problematic to clone and sequence. Therefore subsequent sequencing was carried out using the Illumina TruSeq library preparation kits and reagents. Whilst the chemistry of the reagents and the adapter sequences used are different to those used to construct the corn snake "test" libraries, the TruSeq library preparation methodology is fundamentally the same except that purification steps are carried out using Agencourt AMPure XP beads rather than Qiagen PCR purification kits.

Genomic DNA samples for *E. coloratus* and *P. guttatus* were sent for library preparation and sequencing by the GenePool at the University of Edinburgh (http://genepool.bio.ed.ac.uk/). Two libraries per species were made with selected insert sizes of 300bp and 600bp using the Illumina TruSeq DNA sample preparation kit. These were then pooled and sequenced on one lane of the Illumina HiSeq2000 sequencing platform.

The genomic library for *E. pyramidum* was prepared by me at the IBERS (Institute of Biological, Environmental and Rural Sciences) phenomics laboratory at Aberystwyth University, again using the Illumina TruSeq sample preparation kit but in this instance using an insert size of 400bp. This library was then loaded onto the Illumina MiSeq sequencing platform by me and two runs were carried out, one using 2x150bp reads and another using 2x250bp reads.

2.4.4 Read quality control

All genomic reads were assessed for quality using the software program FastQC (Andrews 2010).Reads were trimmed of any poor quality bases using the python script `fastqTrim.py` (see additional material CD) using the syntax:

```
python fastqTrim.py file1.fastq file2.fastq
Shuffled_file.fastq 15 99
```

Where `file1.fastq` and `file2.fastq` represent forward and reverse read files respectively. The last two numbers of the script indicate where the reads are to be trimmed, in the above example reads are trimmed to leave bases between positions 15 and 99.

Metrics for the number of paired-end reads per sample and the total number of bases sequenced per library were obtained using the perl script `prinseq-lite.pl` (Additional material CD, available online at http://prinseq.sourceforge.net/manual.html) using the syntax:

```
perl prinseq-lite.pl -fastq file.fastq -out_format 3 -out_good
test.fastq
```

2.4.5 CLC

The CLC genome assemblies for corn snake and *Echis coloratus* were carried out using default parameters by the Genepool at Edinburgh University. The CLC assemblies for *Echis pyramidum* were carried out by me at Aberystwyth University. All parameters were kept as default to allow comparison to the assemblies of the other two study species, except for one assembly which was carried out using a k-mer size of 31 as used in the assembly of the king cobra genome (Vonk et al. 2013).

2.4.6 ABySS

Assemblies using single-end reads only, paired-end reads from one genomic library, and paired-end reads from multiple sequencing libraries were assembled using ABySS. Examples of the commands for each are given below. In all cases the k-mer size is given as k=60 and the minimum mean k-mer coverage (c) of a unitig is given as 5.

Single-end assembly:

```
ABYSS -k60 reads.fastq -o assembly.fasta
```

Paired-end using one insert size library:

```
abyss-pe -j1 k=60 n=5 np=12 c=5 mpirun=mpirun
name=output_directory lib=Left_reads.fastq Right_reads.fastq'
```

Paired-end using multiple libraries:

```
abyss-pe -j1 k=60 n=5 np=12 c=5 mpirun=mpirun
name=output_directory lib='lib1 lib2' lib1='
Left_reads_300bp.fastq Right_reads_300bp.fastq'
lib2='Left_reads_600bp.fastq Right_reads_600bp.fastq'
```

2.4.7 Basic assembly metrics

Basic assembly metrics such as maximum contig/scaffold length and contig/scaffold N50 values were obtained using the perl script `contig-stats.pl` (Additional material CD, available online at http://milkweedgenome.org/?q=node/2) using the command:

```
perl contig-stats.pl Assembly_file.fasta
```

Contig and scaffold NG50 values were assessed using the perl script `Assemblathon-stats.pl` (Additional material CD, also available online at http://korflab.ucdavis.edu/datasets/Assemblathon/Assemblathon2/Basic_metrics/assemblathon_stats.pl) using the command:

```
perl assemblathon-stats.pl Assembly_file.fasta
```

To assess the number of gene sized scaffolds (≥25Kb), the perl script `selectSeqsAboveMinLength.pl` (Additional material CD, available online at http://nebc.nerc.ac.uk/tools/code-corner/scripts/sequence-processing#-selectseqsaboveminlength-pl) which extracts sequences of a designated size was used using the command:

```
perl selectSeqsAboveMinLength.pl Assembly_file.fasta Gene-
sized-output.fasta 25000
```

2.4.8 CEGMA

All generated genome assemblies were assessed for completeness using the Core Eukaryotic Genes Mapping Approach (CEGMA) (Parra et al. 2007) using the pipeline version v2.4.010312. Available genome sequences for the Burmese python (Castoe et al. 2013) and the king cobra (Vonk et al. 2013) were also incorporated into the CEGMA analysis to allow subsequent comparison between assemblies.

Firstly, all sequence .fasta headers required to be altered in order to be compatible for the CEGMA pipeline (solely numerical headers are not accepted). This was completed using the Linux command:

```
sed 's/^>/>SampleA/g'
Genome_assembly.fasta>CEGMA_Genome_assembly.fasta
```

This results in all fasta sequences within the file `Genome_assembly.fasta` being amended with "SampleA" at their start, and all sequences being output into the file `CEGMA_Genome_assembly.fasta`. The compatible assembly was then run through the CEGMA pipeline using the command:

```
cegma --genome CEGMA_Genome_assembly.fasta -o CEGMA_OUTPUT
```

2.5 Results

2.5.1 Raw sequencing metrics and initial assessment

Metrics for the raw sequencing data can be seen in Table 2.1. A total of ~58Gb was sequenced for the painted saw-scaled viper, ~11Gb for the Egyptian saw-scaled viper, and ~29.6Gb for the corn snake (Figure 2.13). It is immediately apparent that the Illumina MiSeq provides a much lower sequencing coverage than the Illumina HiSeq2000 platform (Figures 2.13 and 2.14).

Table 2.1. Overall genome sequencing data produced from Illumina sequencing

Species	Insert size (bp)	Read length (bp)	Total Paired-End reads	Total bases
Echis coloratus	300	2x100	225,870,544	22,812,924,944
Echis coloratus	600	2x100	353,897,282	35,389,728,200
Total			579,767,826	58,202,653,144
Echis pyramidum	400	2x150	22,252,762	3,294,553,256
Echis pyramidum	400	2x250	39,541,006	7,998,447,172
Total			61,793,768	11,293,000,428
Pantherophis guttatus	300	2x100	202,756,304	20,478,386,704
Pantherophis guttatus	600	2x100	90,602,508	9,150,853,308
Total			293,358,812	29,629,240,012

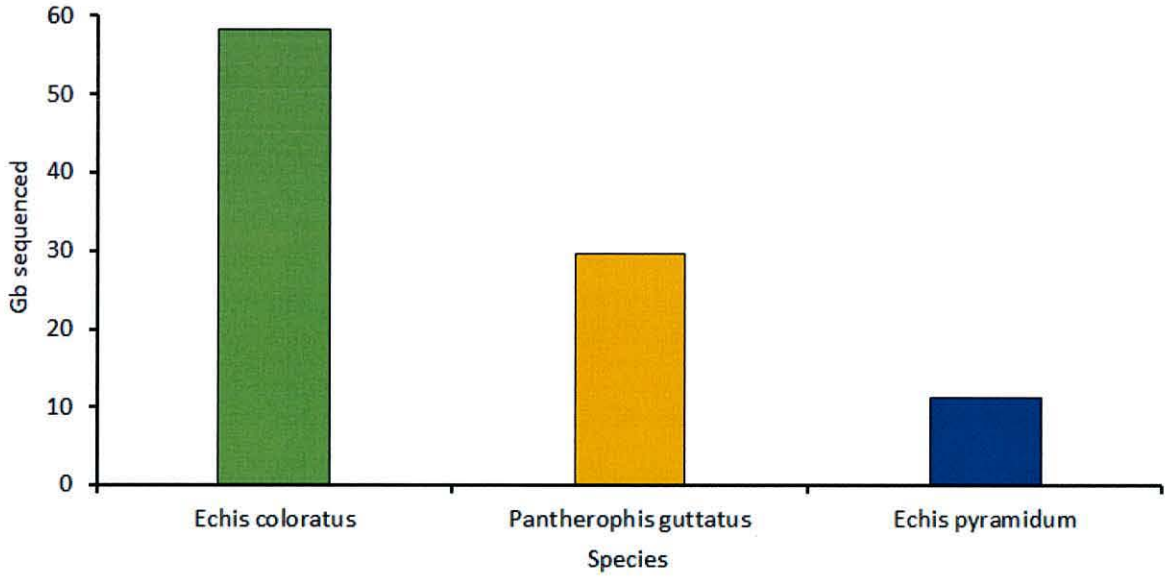


Figure 2.13. Number of bases sequenced in Gigabases (Gb) for each of the three genome species.

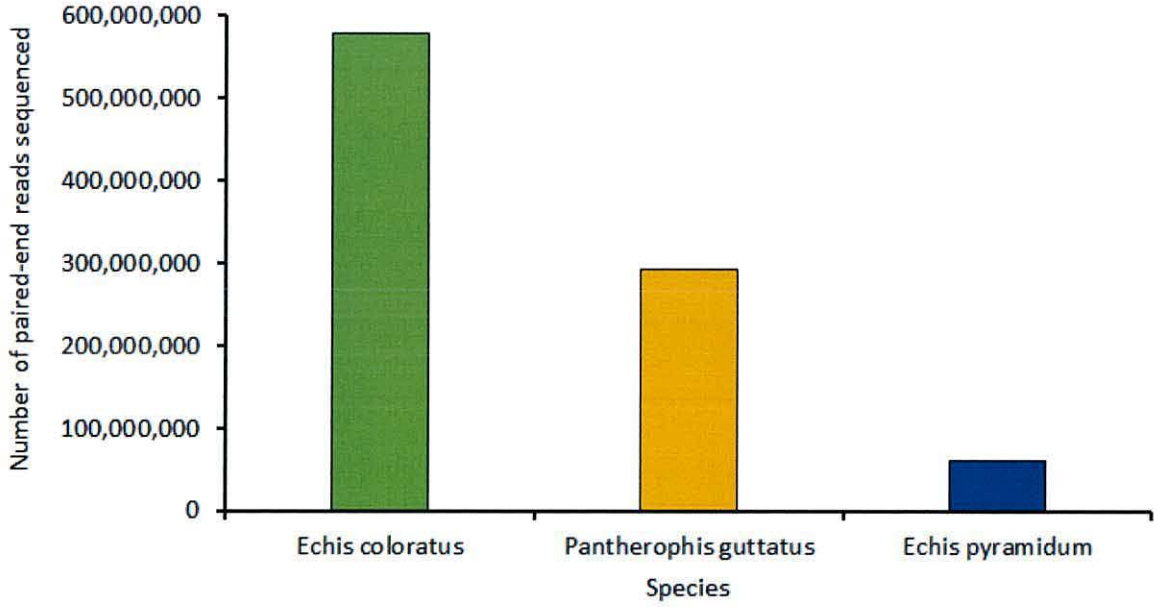


Figure 2.14. Number of paired-end reads sequenced for each of the three genome species.

It has long been known that biases in the GC content (guanine-cytosine content) of genomes can have significant effects on *de novo* genome assembly (Chen et al. 2013). The majority of assembly programs operate based on the assumption that the reads they are using to generate an assembly are distributed evenly across the target genome. However, as mentioned previously due to compositional biases within a genome there is asymmetry of sequencing coverage. Both low and high coverage k-mers will have an effect on the construction of a De Bruijn graph

during assembly: edges which have low weighted k-mer support due to low k-mer coverage may be treated as the result of sequencing error, and be collapsed. High coverage k-mers will be treated as repetitive elements, again leading to the collapse of the De Bruijn graph (Miller et al. 2010). A strong bias in GC content can result in shorter assembled transcripts, inaccuracy, or possibly a failure to assemble the genome at all (Kozarewa et al. 2009; Chen et al. 2013).

Biases in GC content can be due to several factors involved with Illumina sequencing, especially during the PCR enrichment step of library preparation and from bridge amplification during cluster generation on the flow cell (Aird et al. 2011). Chen et al. (2013) found that issues caused by GC bias could be overcome by the addition of more sequencing data as it equalised the read coverage of GC biased regions. With this in mind, the percentage GC content of sequencing reads was assessed prior to assembly. The percentage GC content for the newly generated data were found to be 40.88% for *E. pyramidum*, 42.17% for *E. coloratus*, with corn snake having the highest GC content with 44.00%. These findings are in accordance with the Burmese python genome paper (Castoe et al. 2013) which found, based on analysis of GC content at third codon positions, an increase in AT content in snakes compared to other amniote genomes.

The sequencing coverage or sequencing depth of a genome can be defined as the theoretical amount of times that a particular nucleotide in a genome is sequenced by a sequencing experiment (Sims et al. 2014). Extrapolating from this definition, the higher the sequencing coverage the increased number of times a particular nucleotide is sequenced, improving the confidence that that nucleotide is correct within an assembly and not the result of sequencing error. Based on an estimated genome size of 1.3Gb for the painted saw-scaled viper and the Egyptian saw scaled viper (derived from the haploid genome size of 1.27Gb for the related *Echis carinatus* (Desmet 1981)) these genomes have been sequenced to a depth of roughly 89x and 20x respectively according to the Lander/Waterman equation ((number of reads x read length)/genome size (Lander and Waterman 1988)). An estimated genome size of 1.965Gb was used for the corn snake, which is an average of the four *Elaphe* species available on the genome size database (the corn snake was formerly a member of the *Elaphe* genus (Pyron and Burbrink 2009)). Based upon this figure the genome of the corn snake was sequenced to a depth of roughly 30x.

It is apparent that there was a difference in sequencing coverage between libraries for each species as shown in Figure 2.15. The 600bp insert library for *E. coloratus* appears to have sequenced exceptionally well, accounting for 54.4x coverage by itself. The 600bp insert library for the corn snake however only represents a 9.2x sequencing coverage for this genome.

Therefore there appears to have been some inequality in the sequencing coverage across libraries, despite having the same insert size. An explanation for this may be that the *E. coloratus* 600bp library was sequenced on a different run to the other libraries sequenced at the GenePool, and could therefore have been sequenced to a higher coverage depending on how many other samples were sequenced on the flow cell. The two MiSeq sequencing runs also show variation in the amount sequenced, most likely due to the difference in read length between them, with the 2x150bp read length run having 5.1x coverage but the 2x250bp read run having approximately three times more coverage at 15.2x.

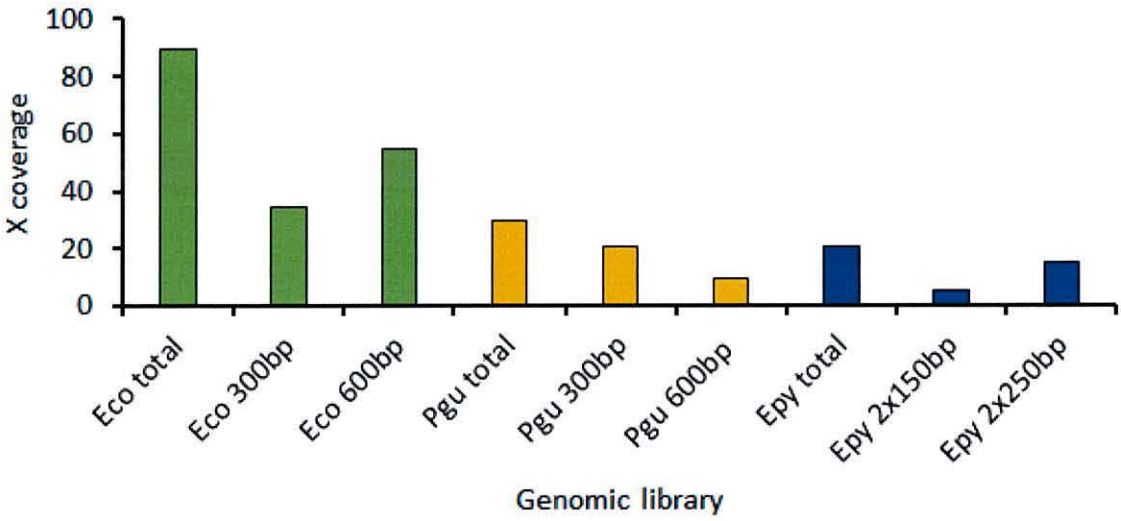


Figure 2.15 Estimated sequencing coverage for the genomes of the three study species. Epy, *Echis pyramidum*; Eco, *Echis coloratus*; Pgu, *Pantherophis guttatus*.

Raw sequencing reads were then assembled using CLC and ABySS. Full results tables of assembly metrics can be seen in Appendix 1-6.

2.5.2 CLC assemblies

CLC Genomics workbench (CLC bio) assemblies were here used as a benchmark to beat. Analysis of these assemblies shows that they are broadly similar across all species (Figure 2.16) with only a slight increase in metrics for *E. coloratus* which is likely a consequence of a higher sequencing depth for this species. Even assemblies of the *E. pyramidum* genome appear to be highly similar, despite using different sets of sequencing reads and selecting a k-mer size of 31.

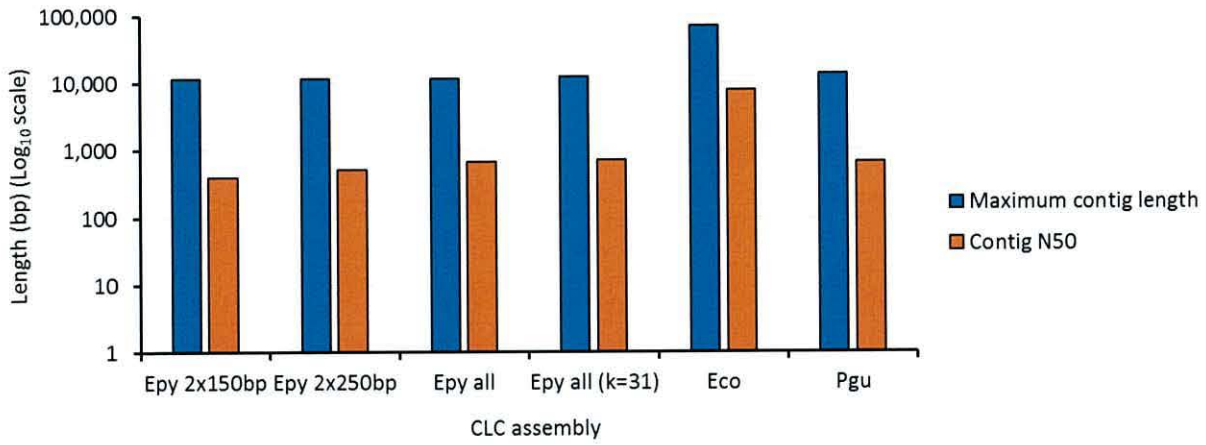


Figure 2.16. Maximum contig length and contig N50 values for genome assemblies produced by CLC. Epy, *Echis pyramidum*; Eco, *Echis coloratus*; Pgu, *Pantherophis guttatus*.

2.5.3 Trimmed vs untrimmed reads

Whilst the sheer amount of data produced by modern DNA sequencing methods is extremely beneficial, it also increases the likelihood of sequencing errors occurring within the data produced. The inclusion of errors into the assembly process may lead to the formation of incorrect k-mers, which in term will cause complications in De Bruijn graph-based assembly. The removal of low-quality bases from read datasets, whilst aiming to maintain as much of the original read data, is consequently beneficial in improving assembly (Del Fabbro et al. 2013). Whilst some assemblers have in-built facilities to remove lowly supported k-mers (such as ABySS (Simpson et al. 2009)), it is usually necessary to carry out read trimming using a different program. Del Fabbro et al. (2013) found read trimming to be beneficial to *de novo* genome assembly, and so low quality bases were trimmed from genomic sequencing reads, the metrics for which can be seen in Table 2.2. Generally, all reads needed to have the first ~15 bases trimmed due to substantial nucleotide bias at the beginning of sequencing reads (this has also been found for RNA-seq data (Hansen et al. 2010)), most likely indicating the presence of adapter sequence. Trimming sequencing reads prior to genome assembly resulted in 8.97% of total sequencing bases being discarded for *Echis coloratus*; 13.22% for *Echis pyramidum*; and 16.98% for corn snake. This is approximately equivalent to 26 million 100bp paired-end reads being discarded for *E. coloratus*; 3 million 2x250bp paired-end reads for *E. pyramidum*; and 25 million 100bp paired-end reads for corn snake.

Table 2.2. Metrics for sequencing reads before and after trimming of low-quality bases.

Genomic library	Total bases	Total bases post-trim	Total bases discarded
Echis coloratus 300bp	22,812,924,944	20,780,090,048	2,032,834,896
Echis coloratus 600bp	35,389,728,200	32,204,652,662	3,185,075,538
Total	58,202,653,144	52,984,742,710	5,217,910,434
Echis pyramidum 2x150bp	3,294,553,256	2,852,583,956	441,969,300
Echis pyramidum 2x250bp	7,998,447,172	6,947,076,986	1,051,370,186
Total	11,293,000,428	9,799,660,942	1,493,339,486
Corn snake 300bp	20,478,386,704	16,626,016,928	3,852,369,776
Corn snake 600bp	9,150,853,308	7,973,020,704	1,177,832,604
Total	29,629,240,012	24,599,037,632	5,030,202,380

Despite previous reports stating that read trimming improves *de novo* genome assembly (Del Fabbro et al. 2013), results suggest that assemblies using untrimmed reads have a higher number of contigs and scaffolds (Figure 2.17) and in some cases a higher maximum contig or scaffold length (Figure 2.18)

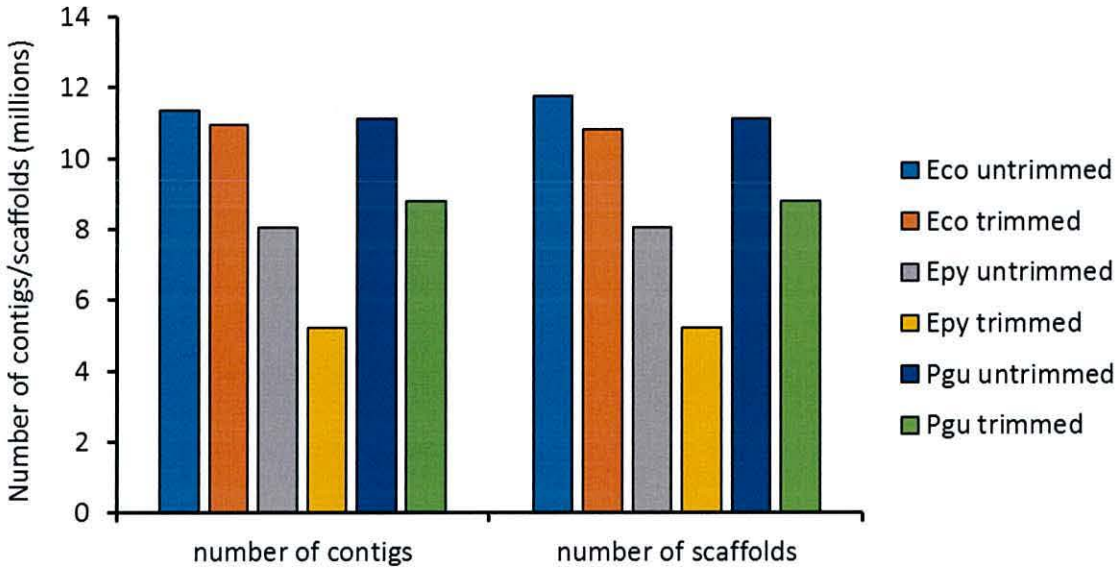


Figure 2.17. Number of contigs and scaffolds in assemblies generated using trimmed or untrimmed sequencing reads. Epy, *Echis pyramidum*; Eco, *Echis coloratus*; Pgu, *Pantherophis guttatus*.

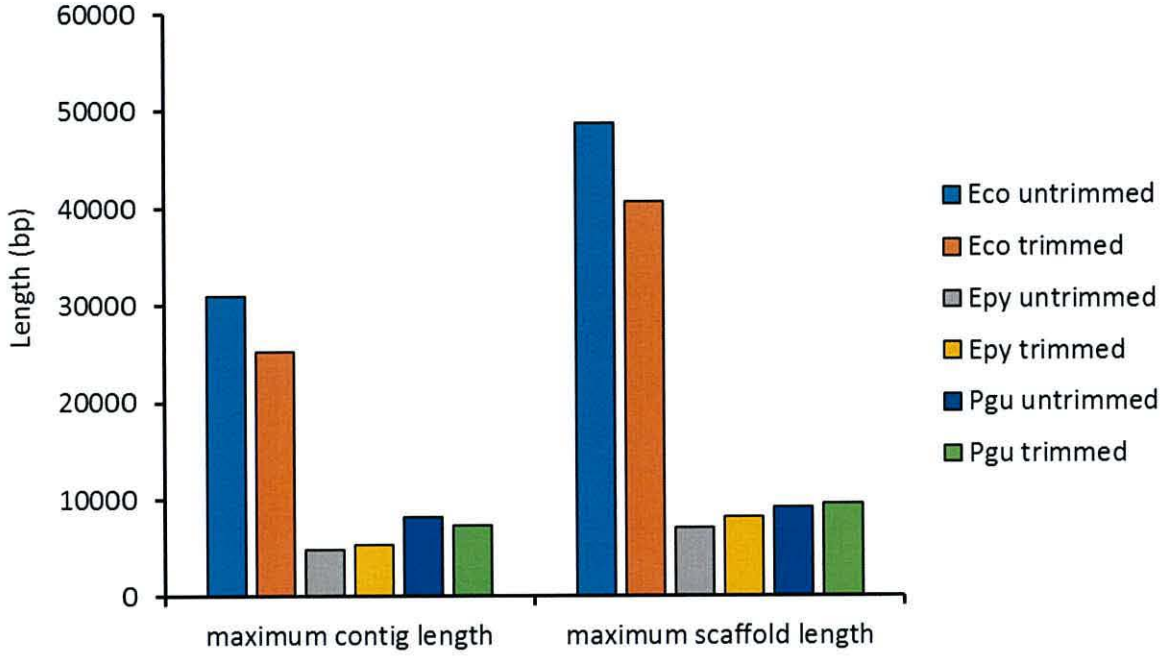


Figure 2.18. Maximum contig and scaffold lengths for assemblies generated using trimmed or untrimmed sequencing reads. Epy, *Echis pyramidum*; Eco, *Echis coloratus*; Pgu, *Pantherophis guttatus*.

Differences in contig and scaffold N50 values were less clear-cut (Figure 2.19) with the scaffold N50 for corn snake and *E. pyramidum* showing little difference between untrimmed and trimmed assemblies, but values for *E. coloratus* being superior when reads were left untrimmed.

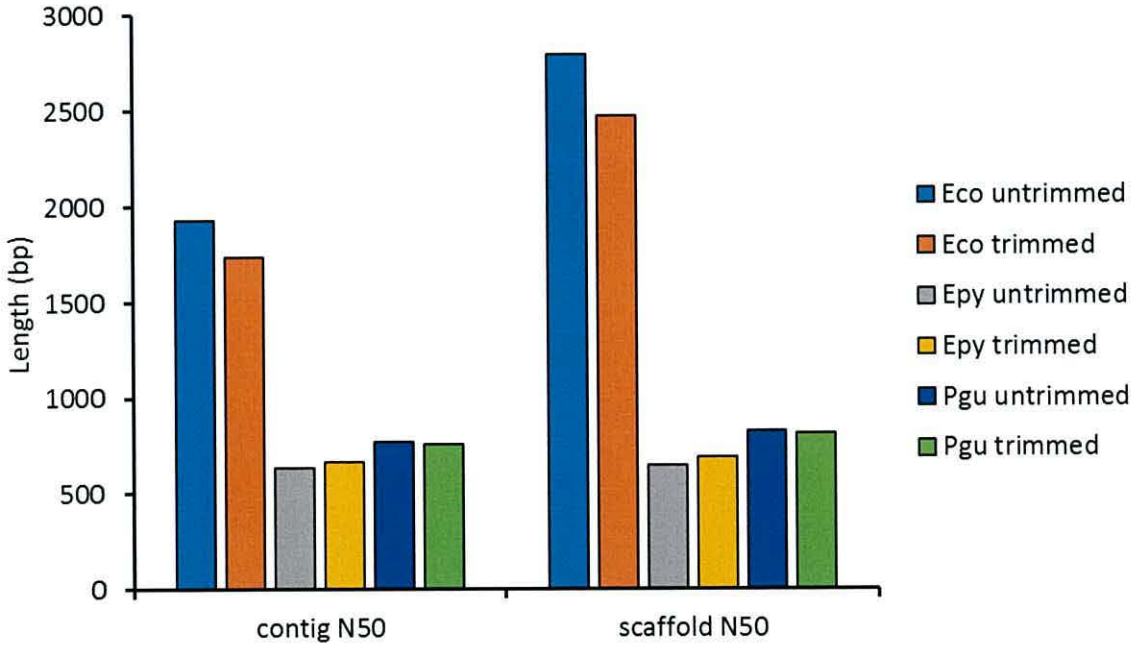


Figure 2.19. Contig and scaffold N50 values for assemblies generated using trimmed or untrimmed sequencing reads. Epy, *Echis pyramidum*; Eco, *Echis coloratus*; Pgu, *Pantherophis guttatus*.

2.5.4 Single-end versus paired-end reads

It is easy to see (Figure 2.20) that using paired-end sequencing data for assembly greatly improves both the mean maximum length of contigs and the mean contig N50.

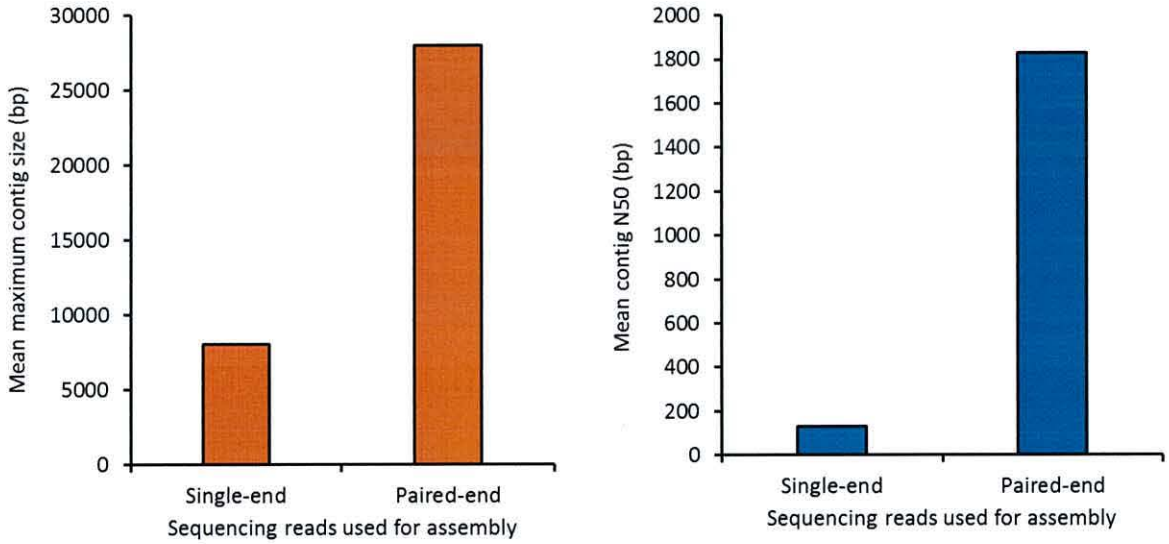


Figure 2.20. Mean maximum contig size and contig N50 for assemblies generated using either single-end or paired-end reads.

To determine whether there was any difference between using "left" or "right" reads (sequence from one end of a DNA fragment or the other), separate assemblies were carried out using each set of reads. Assemblies using left reads appear to be marginally better in terms of their mean maximum contig length and mean contig N50 values (Figures 2.21 and 2.22). Left reads appear to create marginally better assemblies, perhaps due to a difference in error rate between the two reads.

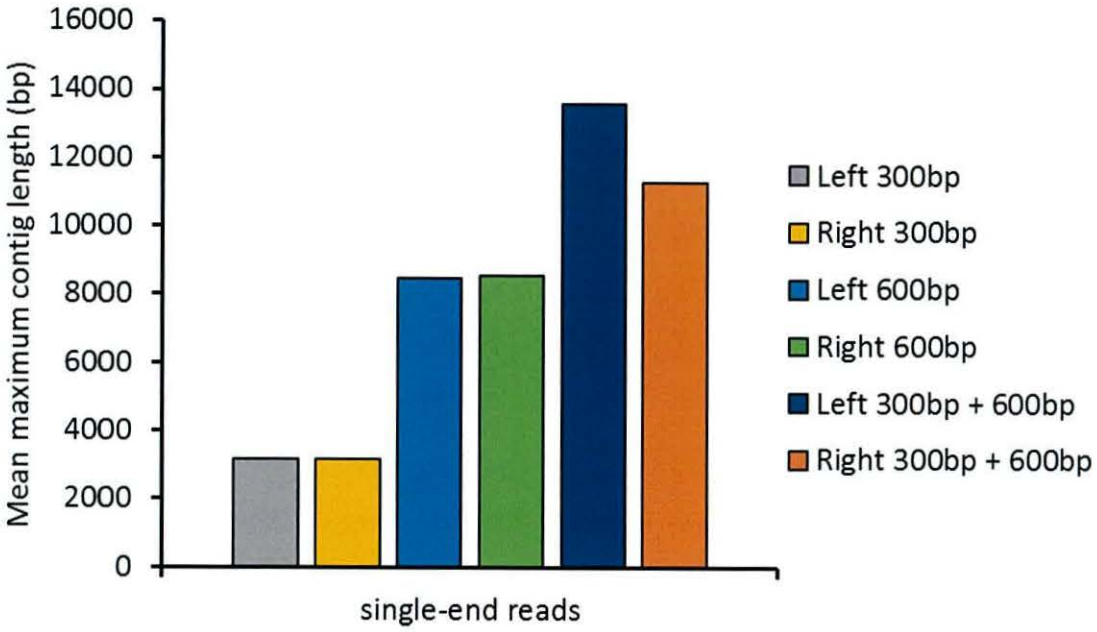


Figure 2.21. Mean maximum contig lengths of assemblies using either left or right single-end reads.

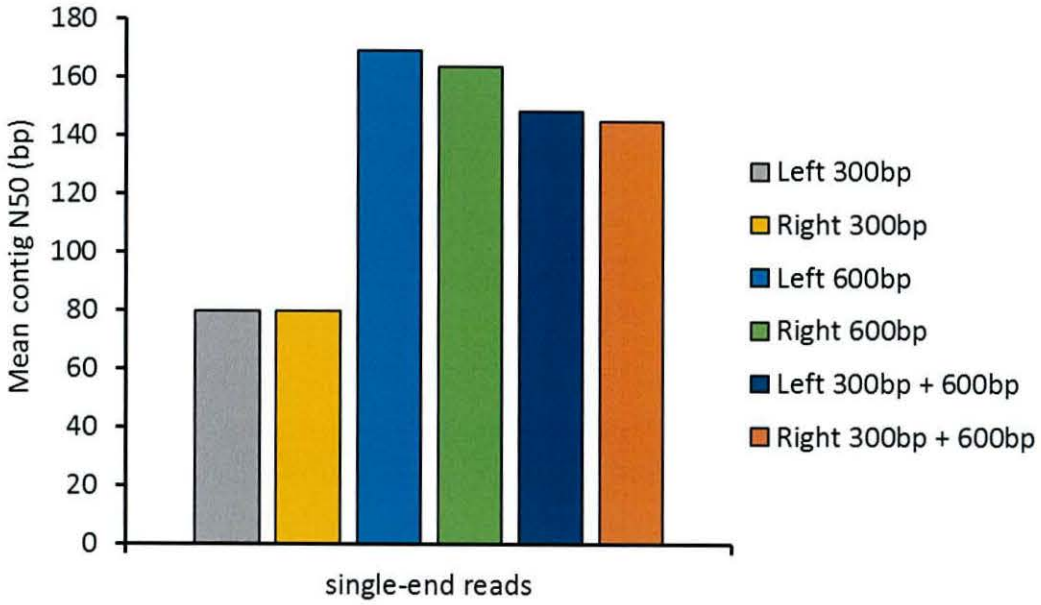


Figure 2.22. Mean contig N50 values of assemblies using either left or right single-end reads.

2.5.5 Sequencing read length

The sequencing data generated by the Illumina HiSeq2000 (giving 2x100bp reads) will not be included in this analysis as the considerably higher sequencing depth of this data may produce misleading results when compared to reads generated by the Illumina MiSeq. Assemblies using reads sequenced using a read length of 2x150bp appear to have produced better assemblies in terms of the maximum contig and scaffold length, but there was little difference in the contig or scaffold N50 values between different read lengths (Figure 2.23). Surprisingly, the combination of both datasets did not lead to any improvement, with assemblies produced using the 2x150bp reads only remaining the best.

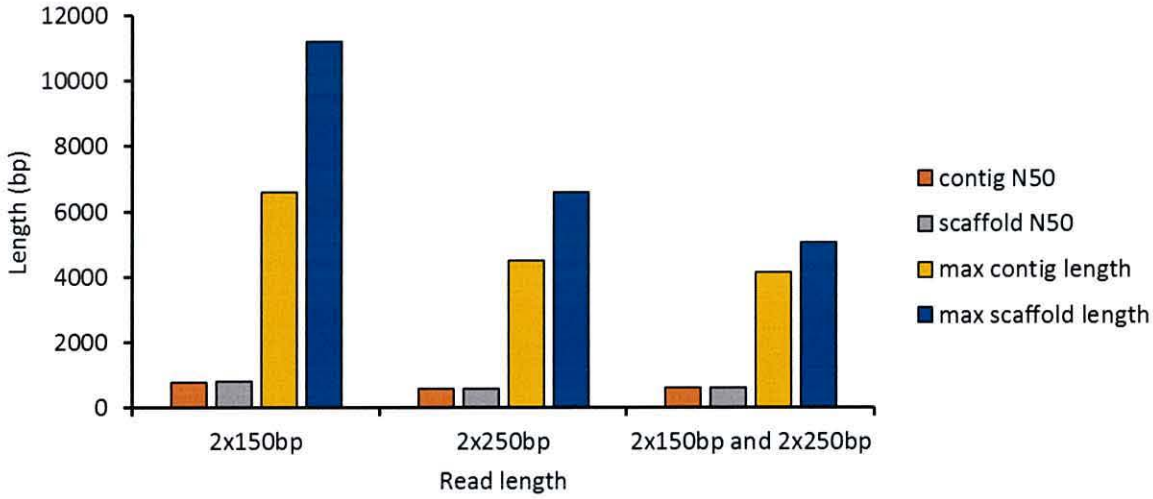


Figure 2.23. Assembly metrics for assemblies generated using either 2x150bp, 2x250bp, or both read lengths sequenced using the Illumina MiSeq.

2.5.6 Library insert size

Sequencing libraries with different insert sizes allowed the evaluation of any effects caused by insert size. The 400bp *E. pyramidum* libraries sequenced on the MiSeq are not considered here due to the difference in sequencing coverage and read lengths between them and those of the other study species, which may be misleading. In general, the larger 600bp insert libraries appear to have produced marginally (in terms of the corn snake) or extremely (in terms of *E. coloratus*) improved assemblies (Figure 2.24). It is most likely that this drastic difference between library insert size in *E. coloratus* is due to the difference in sequencing depth between the two libraries, and not due to the insert size itself.

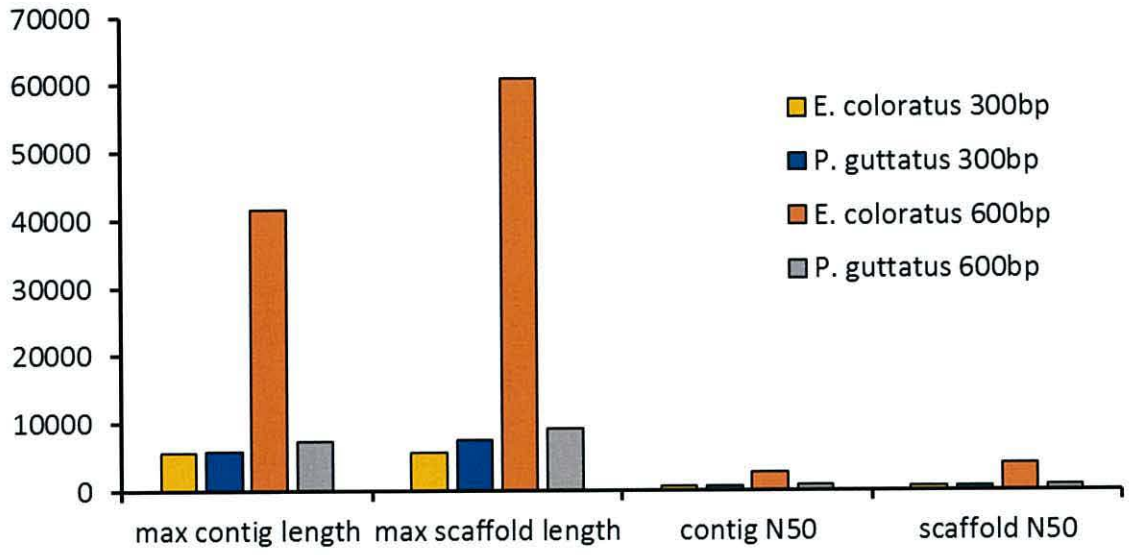


Figure 2.24. Assembly metrics for assemblies generated using read data from libraries with different genomic insert sizes.

2.5.7 k-mer size

An increase in k-mer size results in a commensurate reduction in the number of contigs and scaffolds in an assembly (Figures 2.25 and 2.26).

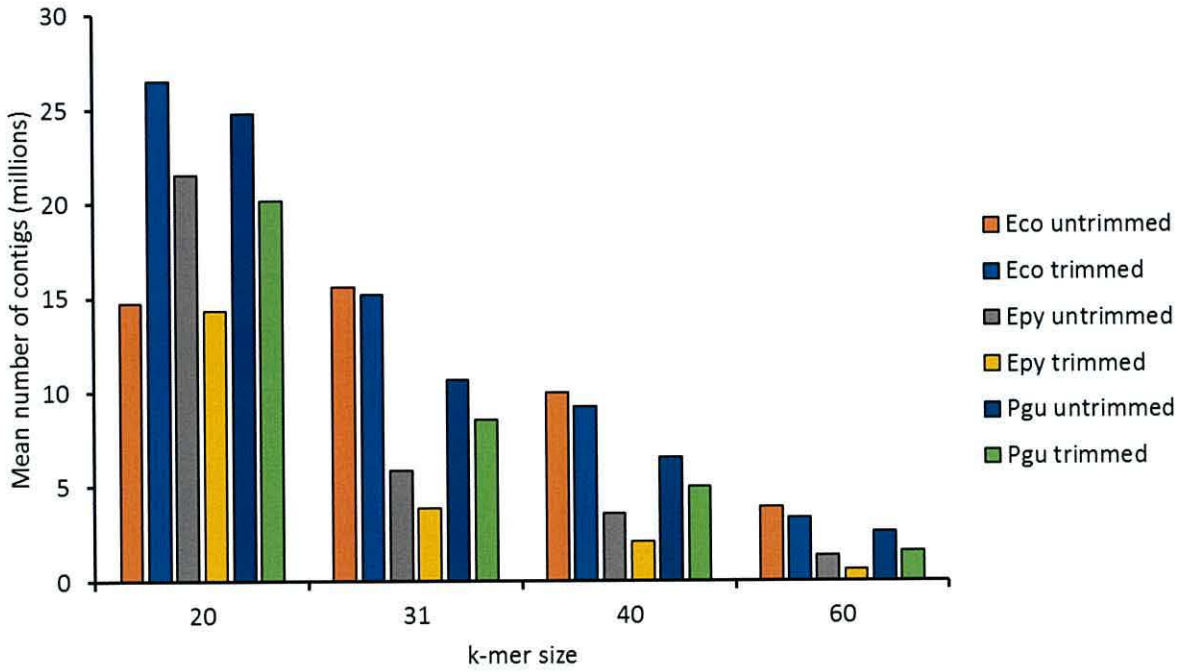


Figure 2.25. Number of contigs (in millions) in assemblies generated using varying k-mer sizes

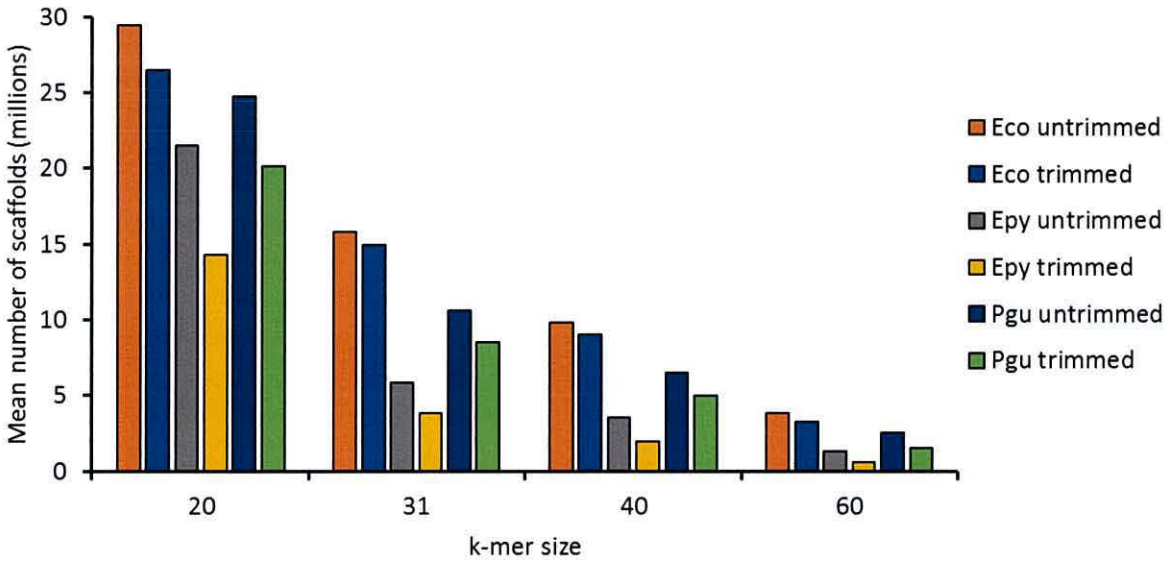


Figure 2.26. Number of scaffolds (in millions) in assemblies generated using varying k-mer sizes.

The effect on the mean maximum scaffold length is less obvious, and appears to be specific to different species. For example, a k-mer of 31 or 40 appears to be optimal for corn snake assemblies, but a k-mer of either 31 or 60 appears to be best for *E. coloratus* assemblies (Figure 2.27). The maximum scaffold length for *E. coloratus* is probably considerably elevated due to the increased coverage of sequencing for the 600bp library, rather than any effect from an increase in k-mer size

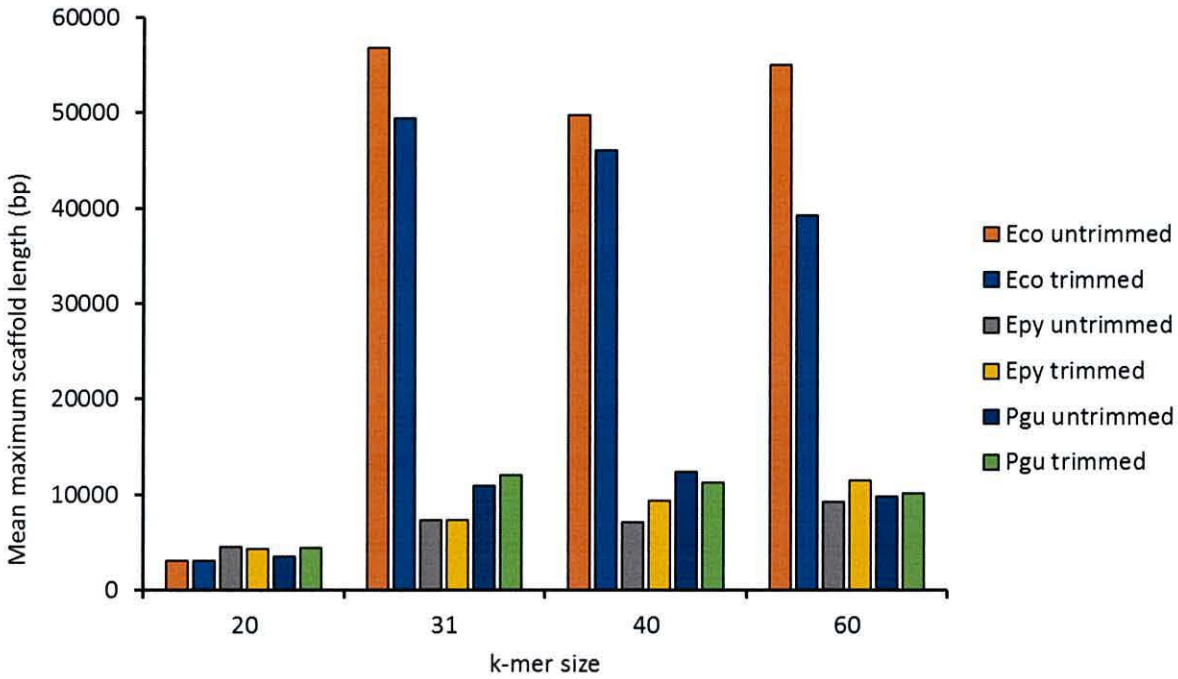


Figure 2.27. Mean maximum scaffold length relative to varying k-mer size.

The contig (Figure 2.28) and scaffold (Figure 2.29) N50 increases with an increase in k-mer size. This increase appears to be slight for corn snake and *E. pyramidum* assemblies but is much more drastic in *E. coloratus*. Again, this could be caused by an increase in sequencing coverage for this species rather than any direct effect caused by an increase in k-mer size on the performance of the assembly.

Also of note is the fact that no paired-end assemblies with a k-mer size of 20 were possible with *E. coloratus* (Figure 2.29). The results displayed for *E. coloratus* at this k-mer size in Figure 2.28 is data based on assemblies carried out using single-end reads which may be due to an inability to construct a De Bruijn graph during assembly due to the k-mer size being too small.

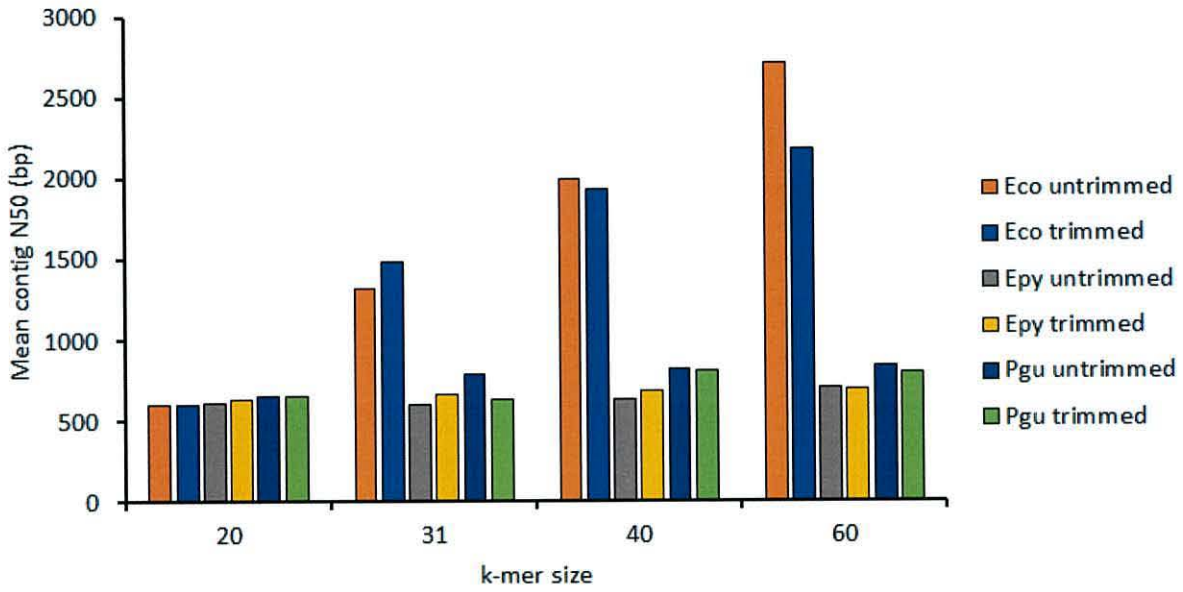


Figure 2.28. Mean contig N50 of assemblies constructed using varying sizes of k-mer.

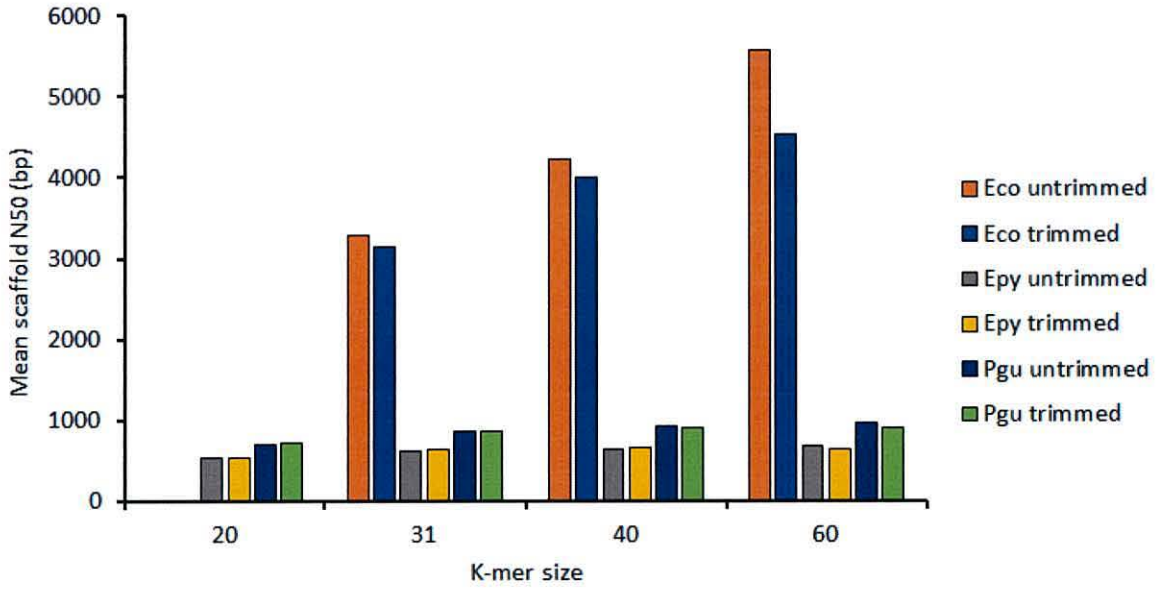


Figure 2.29. Mean scaffold N50 of assemblies constructed using varying sizes of k-mer.

2.5.8 CLC vs. ABySS

In general, the metrics for the CLC genomics workbench assemblies appear to be superior to assemblies generated using ABySS. CLC assemblies have a larger mean maximum contig size compared to the ABySS assemblies for all three species (CLC does not perform scaffolding, and so the maximum scaffold length was not included in this comparison) (Figure 2.30).

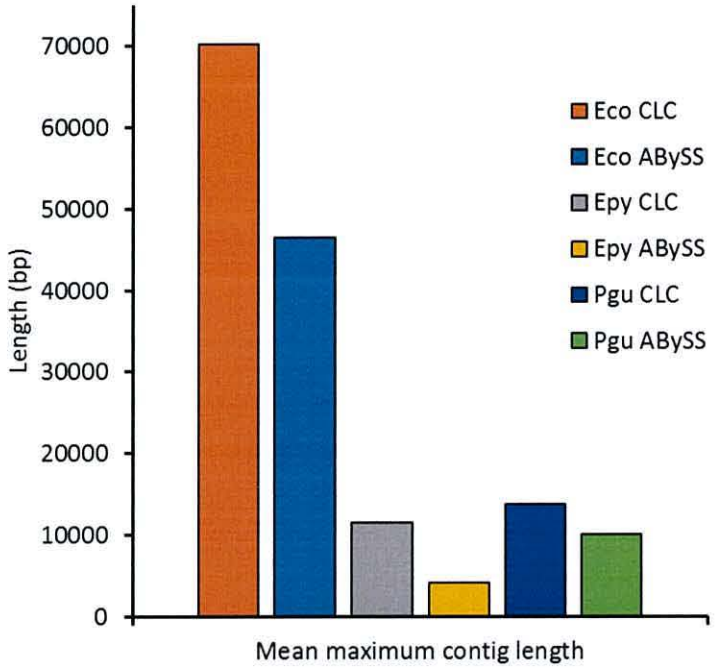


Figure 2.30 Mean maximum contig length of assemblies generated using ABySS or CLC.

Mean contig N50s are slightly more variable, with CLC assemblies for *E. coloratus* being higher, both *E. pyramidum* assemblies being highly similar, and ABySS producing slightly improved contig N50 values for corn snake (Figure 2.31).

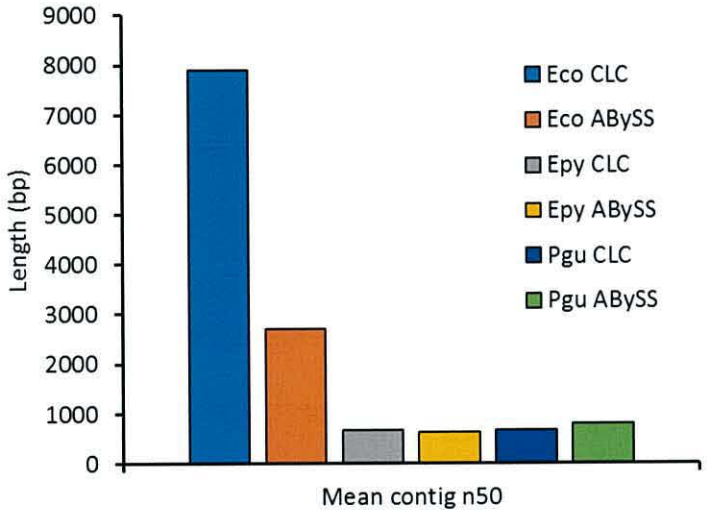


Figure 2.31. Mean contig N50 of assemblies generated using ABySS or CLC.

However, none of the CLC assemblies contain any gene sized contigs/scaffolds and no CEGs (Core Eukaryotic Genes) were found in these assemblies based on analysis with CEGMA (see later sections). Therefore, whilst the metrics for the CLC assemblies may appear to be better, they do not reflect the completeness of the assembly in terms of the number of conserved genes detected or the potential number of contigs/scaffolds containing gene sequences. As such they are less useful for genomic analyses despite the impression given by their metrics.

2.5.9 Gene-sized scaffolds

One metric used to assess the overall quality of an assembly by the Assemblathon 2 (Bradnam et al. 2013) is the number of gene-sized scaffolds (here gene-sized is defined as being ≥25kb in length (Bradnam et al. 2013) based on the average size of a Eukaryotic gene). No assemblies for the corn snake or *Echis pyramidum* had any scaffolds of 25kb or above in length. Likewise, there were no gene sized scaffolds in any *Echis coloratus* assembly assembled using only the reads sequenced from the 300bp insert library.

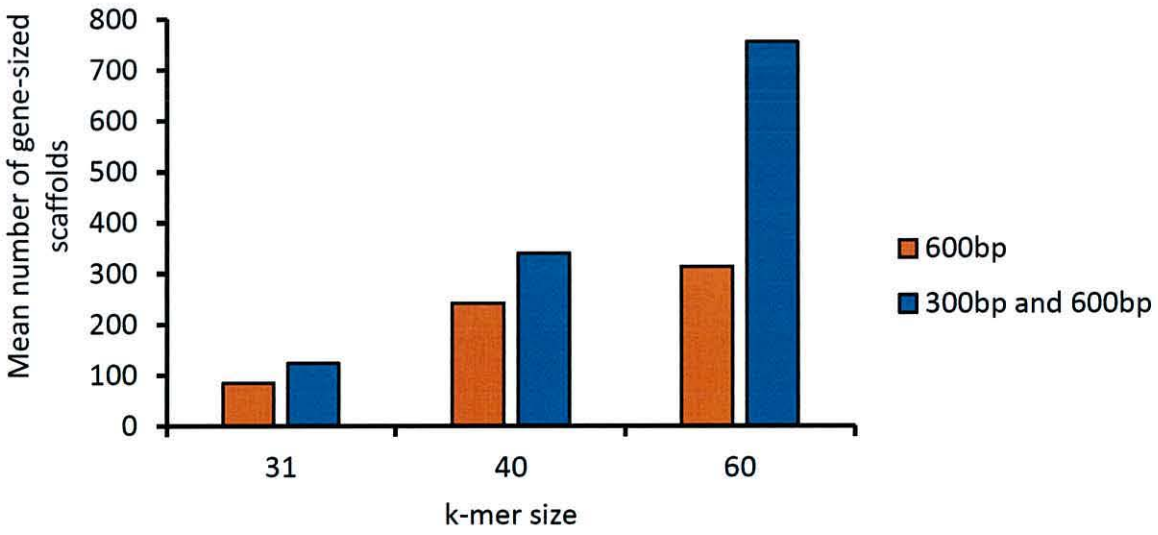


Figure 2.32. Mean number of gene sized scaffolds (≥25Kbp) present in assemblies assembled using varying k-mer size.

It is apparent that an increase in k-mer size led to an increase in the number of gene sized scaffolds in *E. coloratus* assemblies (Figure 2.32). In particular, assemblies using both 300bp and 600bp insert libraries seem to have many more scaffolds of this length, perhaps due to the increased coverage offered by the inclusion of both libraries in the assembly process. The *E. coloratus* ABySS assembly using all untrimmed reads with a k-mer size of 60 appears to have the most gene sized scaffolds with 1,097 detected in this assembly (Table 2.3). Repeating this assembly with trimmed reads appears to have a considerable loss in the number of gene size scaffolds, perhaps due to the reduction in positional information (and overall loss of data) resulting in inhibited scaffold formation.

Table 2.3. Number of gene sized scaffolds contained within genome assemblies of *Echis coloratus*.

Library insert size (bp)	Trimmed (Y/N)	k-mer size	Number of gene-sized scaffolds (bp)
600	N	31	101
600	Y	31	70
600	N	40	279
600	Y	40	205
600	N	60	587
600	Y	60	41
300 and 600	N	31	143
300 and 600	Y	31	104
300 and 600	N	40	382
300 and 600	Y	40	298
300 and 600	N	60	1,097
300 and 600	Y	60	413

2.5.10 CEGMA and genome completeness

CEGMA analysis showed varied results (Appendix 7), with only 5 assemblies for *E. pyramidum* containing any conserved eukaryotic genes (CEGs). Interestingly, all assemblies for the corn snake and all paired-end assemblies for *E. coloratus* contained at least partial CEG sequences, even the assemblies generated using CLC. This would seem to suggest that higher sequencing depth is key to assembling gene sequences, despite having a smaller read length compared to the *E. pyramidum* dataset. Genome completeness estimates ranged from 0%-33.9% in *E. coloratus*, 0-0.4% in *E. pyramidum* and 0-4.8% in corn snake based on the detection of full length CEG sequences. Values based on partial CEG sequences were higher with ranges of 3.6-78.6% in *E. coloratus*, 0-3.2% in *E. pyramidum* and 1.2%-13.7% in corn snake.

2.5.11 Assignment of "best" assemblies

The "best" assemblies for each of the three study species genomes are based primarily on the detection of CEGs in the genome assembly. This justification is based on the fact that an assembly could have an extremely large scaffold N50, but this does not mean that these scaffolds contain any meaningful information. The presence of detectable protein-coding sequences implies that at least some of the sequences in the assembly have assembled correctly. Therefore a trade-off between basic metrics and CEGMA analysis was used to select the best genome assembly.

The best *E. coloratus* assembly was determined to be assembled by ABySS using both 300bp and 600bp library reads which were not trimmed and a k-mer size of 60 was used. This assembly has the highest maximum contig length (but not maximum scaffold length), and contig and scaffold N50 values compared to all other assemblies for this species. It also has the highest number of gene sized scaffolds and, despite not having the highest number of complete CEGs, has the highest number of partial CEG sequences.

The corn snake genome chosen was again assembled by ABySS using reads from both libraries which were not trimmed, only this time using a k-mer size of 40. This assembly did not have the highest N50 values or maximum scaffold length, but it did have the most complete and partial CEGs detected by CEGMA analysis.

The best *E. pyramidum* assembly chosen was like the other two assemblies assembled using ABySS using all reads which were untrimmed, this time using a k-mer size of 31. This assembly was chosen as it contained the most partial CEGs (no complete CEGs were detected in any assembly of this species).

2.5.12 Comparison of published and newly generated snake genome assemblies

Newly generated genome assemblies were compared to those of the three currently available sequenced snake whole genome sequences of the boa constrictor (Bradnam et al. 2013), Burmese python (Castoe et al. 2013) and king cobra (Vonk et al. 2013). All of these assemblies were generated using Illumina paired-end and mate-pair sequencing data (with the exception of the Burmese python which also used some 454 data).

Both the mean and maximum scaffold lengths of the newly generated assemblies are considerably lower than those of the three published snake genome assemblies (Figure 2.33). This is perhaps unsurprising given the lack of mate-pair data from the former three genome assemblies, which would likely greatly aid the assembly and orientation of contigs into much larger scaffold sequences.

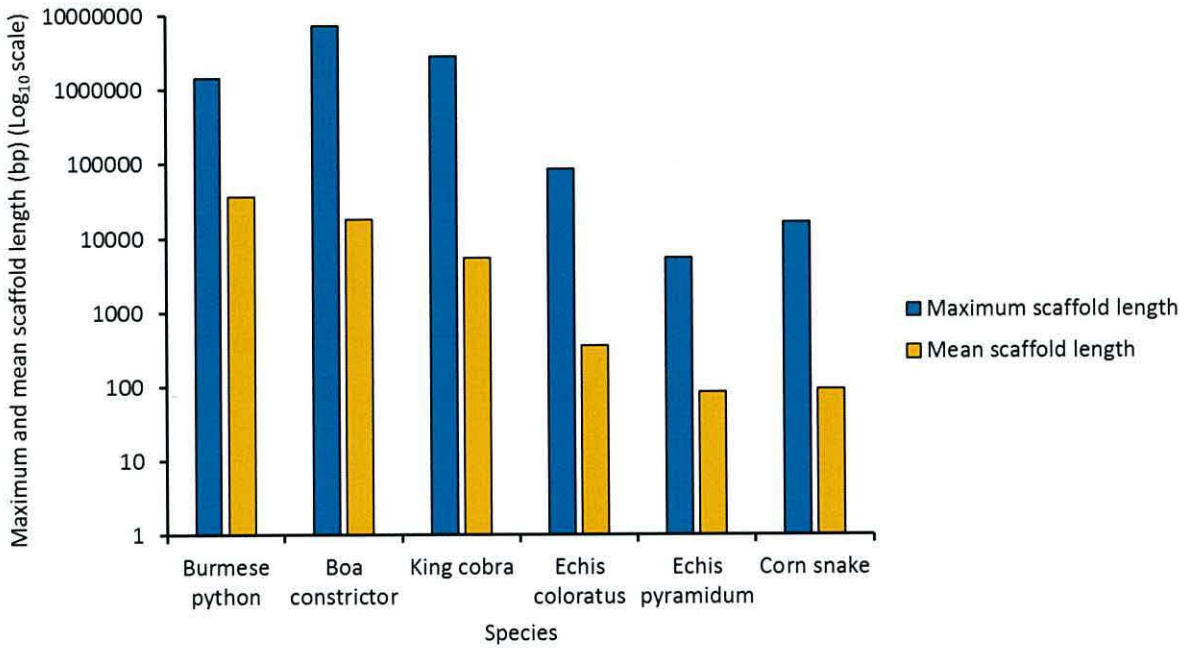


Figure 2.33. Mean and maximum scaffold length of the 6 snake genome assemblies evaluated.

Scaffold N50 and NG50 values are considerably higher in the three published genome assemblies (Figure 2.34), most likely representative of the inclusion of mate-pair library data in these genomes.

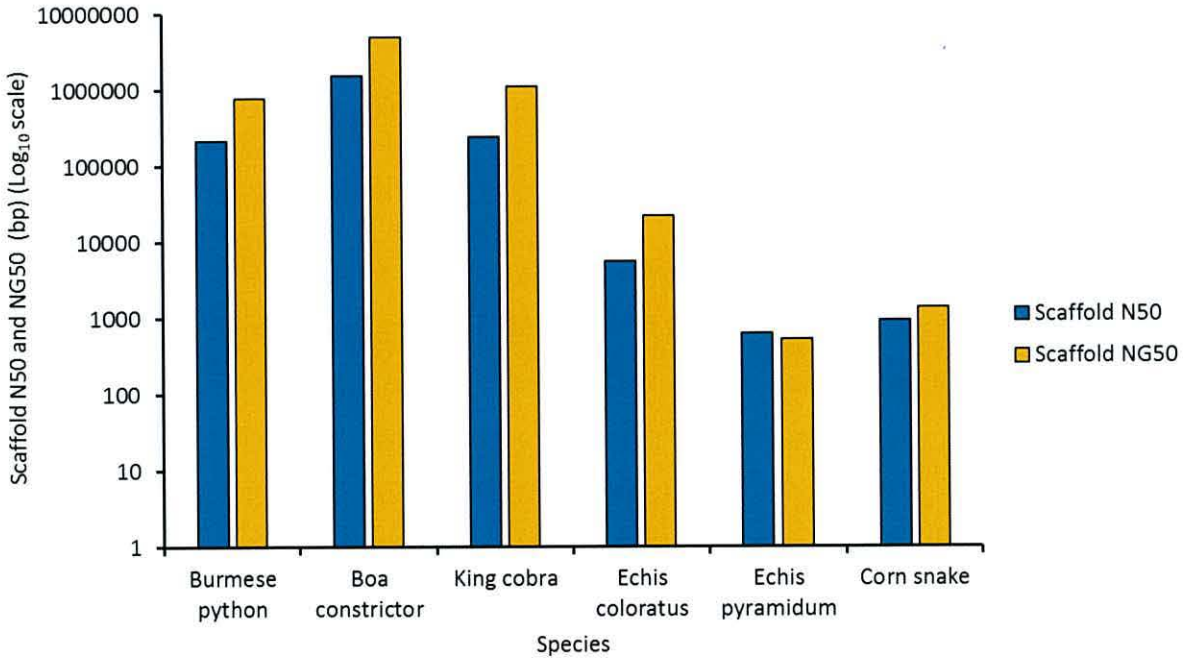


Figure 2.34. Scaffold N50 and NG50 values of the 6 snake genomes assessed.

A surprisingly low amount of gene sized scaffolds were found in the boa constrictor genome assembly (Figure 2.35). Whilst this could initially be interpreted as the assembly containing small scaffolds below 25Kbp, examination of the scaffold N50 and NG50 suggest that this should not be the case. Therefore it is more likely that this is indicative of the boa constrictor assembly containing fewer but longer scaffolds. Conversely, the assemblies for the Burmese python and king cobra appear to have numerous but shorter scaffolds, suggesting that these assemblies contain many gaps. The *E. coloratus* assembly is the only one out of the newly generated genome sequences which contains any gene sized scaffolds.

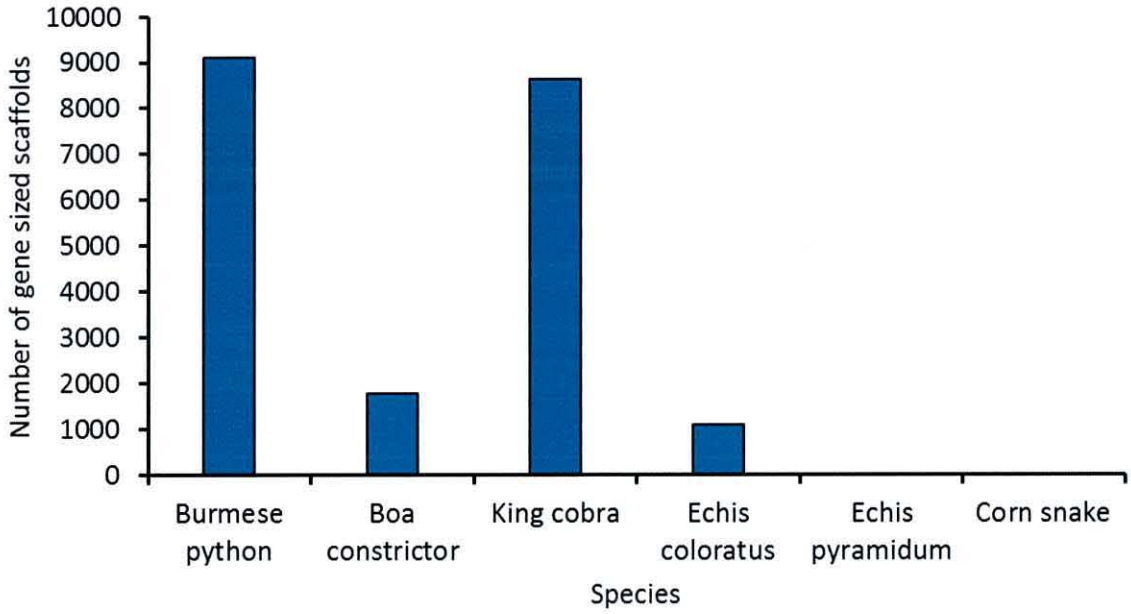


Figure 2.35. Number of gene sized scaffolds (≥25Kb) found in the 6 snake genome assemblies assessed.

Analysis of the conserved eukaryotic gene set using the Core Eukaryotic Genes Mapping Approach (CEGMA (Parra et al. 2007)) identified 81 full length genes out of the 248 highly conserved core eukaryotic genes (CEGs) in the *E. coloratus* assembly, 0 in the *E. pyramidum* assembly, and 12 in the corn snake assembly. This suggests that these genome assemblies are approximately 33%, 0% and 5% complete (compared to 59%, 56% and 49% for the boa constrictor, Burmese python and king cobra respectively). However, 195 partial CEGs were identified in *E. coloratus*, 8 in *E. pyramidum* and 34 in corn snake, giving completeness figures nearer 79%, 3% and 14% (99% for boa constrictor, 96% for Burmese python and 92% for king cobra) (Figure 2.36). Therefore the *E. coloratus* assembly is not too distant in completeness from that of the king cobra, but the mate-pair data included in the king cobra assembly is likely

to be responsible for its increased completeness, especially considering that both assemblies were carried out using ~58Gbp of sequencing data.

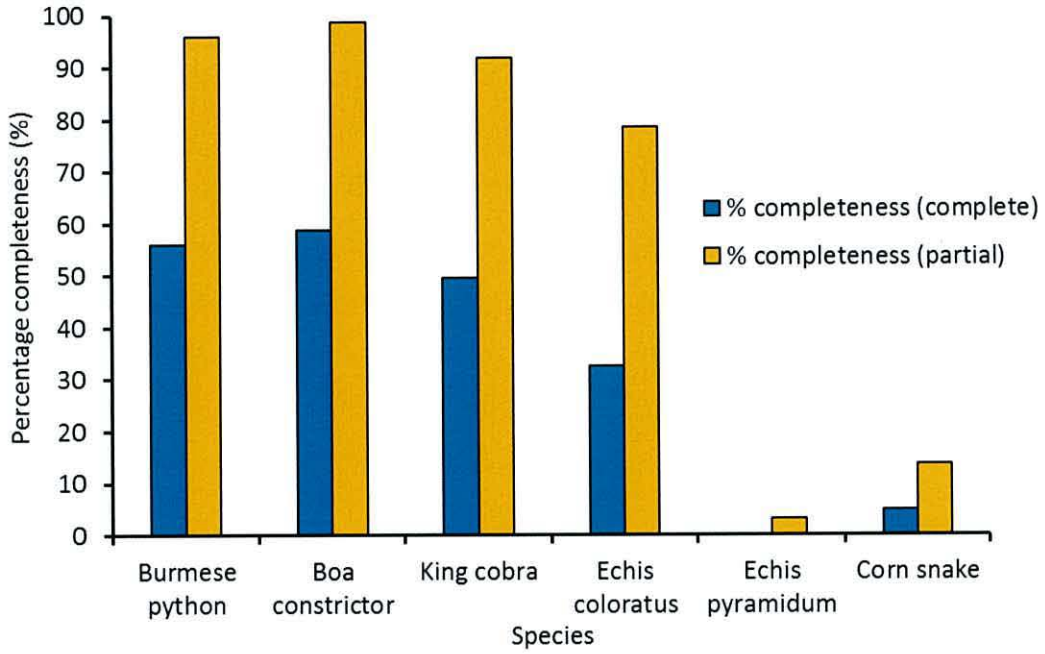


Figure 2.36. Percentage completeness (based on the detection of complete or partial conserved eukaryotic genes) results from CEGMA analysis of 6 snake genome assemblies.

The six snake genome assemblies analysed were ranked according to how they performed in 7 criteria: mean scaffold length, maximum scaffold length, number of gene sized scaffolds, scaffold N50, scaffold NG50, percentage completeness based on complete CEGs, percentage completeness based on partial CEGs (Table 2.4).

Table 2.4. Rankings of six assessed snake genome assemblies in seven assembly evaluation metrics.

Ranking	Mean scaffold length	Maximum scaffold length	No. gene sized scaffolds	Scaffold N50	Scaffold NG50	% completeness (complete)	% completeness (partial)
1	Pmo	Bco	Pmo	Bco	Bco	Bco	Bco
2	Bco	Pmo	Oha	Oha	Oha	Pmo	Pmo
3	Oha	Oha	Bco	Pmo	Pmo	Oha	Oha
4	Eco	Eco	Eco	Eco	Eco	Eco	Eco
5	Pgu	Pgu	Pgu	Pgu	Pgu	Pgu	Pgu
6	Epy	Epy	Epy	Epy	Epy	Epy	Epy

Abbreviations

Bco, *Boa constrictor constrictor*; Eco, *Echis coloratus*; Epy, *Echis pyramidum*; Oha, *Ophiophagus hannah*; Pgu, *Pantherophis guttatus*; Pmo, *Python molurus bivittatus*.

The results of these rankings indicate that the snake assemblies proceed in the following order, with the "best" assembly first- boa constrictor, Burmese python, king cobra, *Echis coloratus*, corn snake and *Echis pyramidum*. The boa constrictor assembly is ranked top in five out of seven categories, and it is worth noting that this assembly is probably only beaten in the "number of gene sized scaffolds" category because this assembly contains fewer but much longer scaffolds than the Burmese python and king cobra assemblies. The Burmese python and king cobra assemblies are roughly similar, with the Burmese python assembly having a slightly higher scaffold lengths and percentage completeness based on CEGMA analysis, but the king cobra assembly having higher scaffold N50 and NG50 values.

The *E. coloratus* assembly is the best of the newly generated genome sequences, having much better results in all categories compared to the much lower coverage genomes of the corn snake and *E. pyramidum*. Surprisingly, the *E. coloratus* assembly is not too dissimilar from the only currently published genome for a venomous snake, the king cobra (Vonk et al. 2013). The addition of mate-pair sequencing data to the *E. coloratus* assembly would likely improve all metrics, potentially raising it to a comparable level with the king cobra and Burmese python assemblies.

2.6 Discussion

122 different genome assemblies were generated for three species of snake, the corn snake (*Pantherophis guttatus*) and the painted (*Echis coloratus*) and Egyptian (*Echis pyramidum*) saw-scaled vipers. The parameters for each assembly for each species were varied in order to test a number of factors which may (or indeed, may not) affect the outcome of a *de novo* genome assembly.

Several factors appear to be very clear based on these results, and should be considered for all snake genome sequencing projects. Firstly, the higher sequencing output of the Illumina HiSeq produced superior genome assemblies compared to the Illumina MiSeq, and therefore this sequencing platform should be used preferentially for initial *de novo* sequencing projects (the MiSeq may be useful for genome re-sequencing when there is a reference sequence to align the

reads to, and so sequencing coverage is less of an issue). Using paired-end reads for assembly was also far superior to single-end reads.

Trimming sequencing reads appears to have had a negative effect on the resulting genome assemblies, with a huge amount of data being discarded before assembly has even been attempted. The Illumina HiSeq25000, using the v3 TruSeq reagents, can currently produce 3-6 billion 100b paired-end reads in high output mode and 600 million 100bp paired-end reads in rapid run mode (www.illumina.com/systems/hiseq_2500_1500/performance_specifications.ilmn). Based on the figures, the amount of sequence discarded from the *E. coloratus* and corn snake read datasets is approximately 0.86% of a HiSeq2500 run in high output mode, and ~4% of a total run in rapid run mode. The Illumina MiSeq is capable of generating 44-50 million paired-end reads in a single run when using the v3 reagent kit (www.illumina.com/systems/miseq/performance_specifications.ilmn). The amount of sequence discarded from the *E. pyramidum* MiSeq reads dataset is approximately 6-7% of an entire MiSeq run. Interestingly, it has previously been noted that ABySS assembles untrimmed datasets better in terms of the resulting basic metrics such as maximum scaffold length and scaffold N50 due to its in-built ability to correct sequencing errors, and that trimming sequencing reads can decrease assembly quality (Del Fabbro et al. 2013)

Sequencing read length does not appear to be an issue at high sequencing depths, but appears to have had an influence at the lower sequencing depth generated by the MiSeq. Assemblies using 2x150bp reads only appear to be better than assemblies with 2x250bp reads. It is possible that this is due to the fact that a maximum k-mer value of 60 was used, whilst the read lengths were now increased by 100bp up to 250bp. The k-mer value may therefore have been too small for these sequencing reads, causing the formation of many lowly supported edges in the De Bruijn graph during assembly, the majority of which would ultimately be collapsed. This is perhaps a major factor to consider when conducting assemblies in future, especially when combining data with different read lengths.

Overall the size of library insert size appears to have had only a minor effect on the resulting assemblies, with metrics for assemblies utilising the 600bp insert size libraries being improved compared to assemblies using only the 300b library. It is likely that the increased positional information given by the 600bp library reads (the increased distance between reads will inform scaffolding much more than reads located close together on a DNA fragment) has improved overall assembly. Based on the assembly metrics, using only the 600bp library for *E. coloratus* produced assemblies which were very similar to those which also included the 300bp library

data. Perhaps this is simply indicative of the high sequencing depth of the 600bp library for this species, and not an improvement based on an increase in library insert size.

Altering k-mer length appears to have a dramatic effect on the resulting assembly, with an increase in k-mer leading to a drastic reduction in the number of contigs and scaffolds in an assembly, and an increase in the contig/scaffold maximum length and N50. As stated in the introduction section, the use of longer k-mers will result in the formation of fewer paths during De Bruijn graph construction, but any overlaps between k-mers will be strict, leading to the assembly of more accurate contig sequences.

Out of the two *de novo* assemblers used, CLC appears to inflate basic assembly metrics such as contig N50 without the resulting assembly containing an increase in meaningful sequence (for example CEG sequences, gene sized scaffolds etc.). This may be why the contig N50 for the king cobra genome is, in terms of the amount of sequencing carried out for this species (41.2Gbp paired-end data and 16.8Gbp mate-pair data), very low but the scaffold N50 is vastly improved following scaffolding with SSPACE (Boetzer et al. 2011; Vonk et al. 2013). This finding highlights the importance of carrying out multiple analyses of a genome assembly such as CEGMA, and not judging the assembly based on basic metrics such as the N50 (as was also found in the Assemblathon 2 (Bradnam et al. 2013)).

Whilst some factors should be applied to all genome sequencing projects, it is also apparent that each assembly project must also be considered on its own merit. For example, a k-mer size of 60 was optimal for *E. coloratus*, but a k-mer of 40 was best for the corn snake. The genome of each species will be different, and hence will present its own individual challenges. Extrapolating from this, the addition of new data for each species may radically alter the "optimal" assembly parameters once more. How well each library sequences on the sequencer also appears to be a key factor, with assemblies using the *E. coloratus* 600bp library only being very similar to assemblies constructed using reads from both sequencing libraries.

In conclusion, there is no widely accepted gold standard for *de novo* genome assembly because each genome is unique, presenting its own unique challenges. The overall quality of an assembly can also be considered to be arbitrary, depending on the purpose of the assembly. For example, the "best" assemblies generated in this study are suitable for gene discovery and even for gene promoter transcription factor binding sites analysis (Chapter 6). However, for analyses requiring large regions of sequence such as investigating genome structure and synteny, these assemblies are not of a sufficient quality.

The limitations of this study and suggested improvements to both the assembly method and approach to evaluating generated assemblies is discussed below. Suggestions for alternative strategies (such as using different sequencing platforms) to improve genome assemblies are discussed in Chapter 7.

Several methods were attempted over the course of this study which were unable to be completed due to technical limitations. Firstly, genome assembly using the assembler SGA (String Graph Assembler) (Simpson and Durbin 2012) was attempted on multiple occasions, as this utilises an overlap-based string graph algorithm rather than the commonly used De Bruijn graph approach. This assembler was also found to be the best for assembling the whole genome sequence of the boa constrictor in the Assemblathon 2 (Bradnam et al. 2013), and so would have been a logical addition to this study. However, after multiple attempts at troubleshooting, this program would not run correctly for unknown reasons.

The remaining issues were all due to limitations of the computing system used. As the size of each file containing one set of single-end reads is considerable (in the region of 44Gb), not enough memory was available on the high performance computing system used in order to carry out these analyses.

The scaffolding of genome assemblies was attempted using the scaffolding software SSPACE (Boetzer et al. 2011), with the aim of comparing the scaffolding results of the built-in scaffolder of ABySS to a stand-alone scaffolder. This program has been shown previously to improve scaffold N50 significantly and to incorporate at least 75% of initial contigs into scaffolds (Boetzer et al. 2011). More specifically, SSPACE was found to out-perform the scaffolding carried out by ABySS (Boetzer et al. 2011), and was used to scaffold the whole genome assembly of the king cobra (Vonk et al. 2013). As the first step of this process involves mapping the sequencing reads onto the genome assembly, and the read files are extremely large, there was not sufficient memory to carry this step out even when using a high powered computing system.

The genome assembly evaluation tool Reapr (Hunt et al. 2013) was trialled as it was again used in the Assemblathon 2 and is a stand-alone tool to evaluate genome assemblies. Most notably it identifies assembly errors (such as incorrect scaffolding) and analyses every single base of the assembly, returning metrics such as a corrected N50 value and the percentage of error-free bases in the assembly. As Reapr is capable of assessing assembly quality without the need for a reference genome (unlike other programs such as Quast (Gurevich et al. 2013)) it would have been perfect for adding a further analysis measure between genome assemblies. As the first step

of the analysis involves mapping sequencing reads onto the genome assembly (and indeed the program is dependent upon the mapping results), it is likely that a limitation of computer storage and memory has prevented the analysis from being carried out.

In future the addition of these aforementioned programs to the analysis would provide different methods to test (in the case of SSPACE and SGA) and an additional metric on which to evaluate the resulting genome assemblies (in the case of Reapr). Whilst some pre-processing was carried out (read trimming), there are additional steps which could be taken and assessed (so much so that they would probably constitute a whole thesis chapter) such as k-mer correction, more stringent sequencing adapter removal and the removal of duplicate sequences. Potentially these could improve, or at least alter, the result of the overall genome assembly.

Finally, the extremely low coverage reads sequenced using the MiSeq for *E. pyramidum* were a limiting factor in the assembly of the genome of this species. It is possible that the assembly could be improved by first mapping the *E. pyramidum* reads onto the better *E. coloratus* genome assembly, and using the result of these read mappings to assemble the *E. pyramidum* genome. This approach has been used previously to close gaps in genomic sequence by mapping reads to a reference genome, isolating the reads which map to gapped regions, carrying out local assembly using only these reads, and then incorporating the resulting contigs back into the genome assembly (Tsai et al. 2010). This approach is also utilised for reconstructing RNA transcripts from reads mapped to the exons of a reference genome sequence (Chapter 3).

Chapter 3

Reptile transcriptome assembly

The transcriptome comprises all of the RNA molecules expressed by a cell or group of cells, and in particular encapsulates all of the protein coding mRNA molecules. As such, the transcriptome can reveal which genes are actively being expressed in a tissue at a given time. Additionally, the quantitative nature of RNA sequencing (RNA-seq) data allows the expression level of transcripts within a transcriptome to be estimated. Here, 48 RNA-seq libraries were sequenced from a range of reptile tissues and species. Three transcriptome assembly software programs were evaluated, including two *de novo* programs and one genome-guided method. It was found that Trinity was a superior *de novo* assembler, but genome-guided assembly greatly improved the assembly of low-coverage short read data. Transcript abundance estimation was carried out across a range of tissue samples and venom gland samples at different timepoints following milking. Results confirmed that β actin and GAPDH are highly variable and are unsuitable as reference genes for qPCR experiments. The results of newly generated venom gland transcriptomes were compared to previous EST-based analyses, and it was found that assemblies generated using RNA-seq contained more lowly expressed transcripts not detected by EST sequencing. This approach was also sensitive enough to detect splice variants of several genes not found previously. Finally, sub-assemblies of an *Echis coloratus* venom gland transcriptome were carried out to assess the minimum required sequencing depth needed to fully characterise a venom gland transcriptome.

3.1 The transcriptome

The transcriptome can be defined as all of the RNA molecules expressed by a cell or population of cells, for example in a particular tissue (McGettigan 2013), and the term was first coined by Charles Auffray in 1996 (McGettigan 2013) and first used in a publication the following year (Velculescu et al. 1997). As this definition also includes all of the expressed mRNA molecules, the transcriptome represents all of the protein coding genes being actively transcribed at the time of sampling, which could vary due to various factors such as developmental stage and tissue type (Rudd 2003). In theory therefore the transcriptome is the precursor to the proteome of a cell or tissue, although post-transcriptional and post-translational modification and regulation are likely to cause some disparity between the two.

Traditionally transcriptomes have been analysed by the cloning and sequencing of expressed sequence tags (ESTs) whereby short fragments of a cDNA library are sequenced and then clustered to give a contiguous sequence. ESTs are ultimately limited due to their short length (typically 200-800bp) (Nagaraj et al. 2007) and the low coverage resulting from this approach, meaning lowly expressed transcripts and splice variants are likely to remain undetected (Rudd 2003).

The utility of next generation sequencing technologies to transcriptome sequencing and analysis is apparent for a number of reasons: the high sequencing depth offered by sequencing millions of reads by RNA sequencing (RNA-seq) means it is more likely to recover full-length transcript sequences (including lowly expressed transcripts), and the higher resolution aids in the identification of alternative splice variants. As the number of reads sequenced from a particular transcript will be representative of the amount of that transcript present in a sample, RNA-seq data is also highly quantitative (Marguerat and Bähler 2010), meaning it can be used both for transcript characterisation and transcript expression analysis in a single experiment.

However, the assembly of RNA-seq reads into a transcriptome assembly also poses several challenges, especially in the absence of a reference genome to aid in the reconstruction of transcripts. Unlike the genome sequence of an organism which remains relatively static, the transcriptome can be highly variable. This means that the number of mRNA transcripts encoding different genes will be present at different abundances within a sample, leading to uneven sequencing coverage (Rudd 2003), particularly in highly transcriptionally active tissues. The short read length of RNA-seq data also means that reads from highly similar transcripts, such as paralogs belonging to the same gene family, may be fused during the assembly process resulting in chimeric sequences. Additionally, alternative transcripts of the same gene may be

omitted altogether if the abundance of one variant in a sample significantly outweighs the other(s). Finally, shared homologous sequences in different genes (such as homeodomains) may be incorporated or omitted erroneously due to uncertainty of which transcript the sequence belongs to. This may also be problematic when mapping RNA-seq reads to transcriptome sequences to carry out transcript abundance estimation (see later section). As alluded to previously, the reads generated from RNA-seq must be assembled into contiguous sequences (contigs) in order to be useful, which can either be done *de novo* or by using a genome sequence as a reference to aid assembly.

3.2 Genome-guided transcriptome assembly

The use of a reference genome sequence in transcriptome assembly means that reads are correctly orientated and assembled into a "real" transcript encoded by the genome, whilst *de novo* assembly may reconstruct artificial transcripts. Additionally, low coverage sequencing may be assembled correctly using this method, whilst lacking the resolution needed to assemble transcripts *de novo* (see results section). The disadvantage of this approach is that splice variant transcripts may be discarded, especially if their expression is lower than that of a full-length transcript and reads which map to multiple regions within the genome can cause ambiguity during transcript reconstruction (Martin and Wang 2011). Indeed, there must also be a sequenced whole genome sequence, ideally of sufficient quality to contain full-length gene sequences on scaffolds in order to assemble full length transcripts.

3.2.1 The Tuxedo suite

The Tuxedo suite (or Tuxedo method) utilises several different programs in order to reconstruct transcripts from short sequencing reads mapped to a reference genome (Trapnell et al. 2012). Firstly the reference genome is indexed with a Burrows-Wheeler index in order to increase the speed and efficiency of searching the genome sequence for the occurrence of short sequences (Langmead et al. 2009). Then short reads from RNA-seq datasets are mapped to the reference genome sequence using the read aligner Bowtie (Langmead and Salzberg 2012) and the splice junction mapper TopHat (Trapnell et al. 2009) uses the results of these read mappings to predict and identify exon splice junctions. The program Cufflinks (Trapnell et al. 2010) then uses this mapping and splicing information to reconstruct transcript sequences from the reference genome sequence.

3.3 *De novo* transcriptome assembly

Several *de novo* assembly programs are available for transcriptome assembly when there is no reference genome available such as Trinity (Grabherr et al. 2011), Oases (Schulz et al. 2012), Trans-ABySS (Robertson et al. 2010), SOAPdenovo-Trans (Xie et al. 2014) and IDBA-tran (Peng et al. 2013). The majority of *de novo* transcriptome assembly programs designed for use with short-read sequencing data utilise De Bruijn graph-based algorithms, as discussed in Chapter 2.

3.3.1 Trinity

Trinity is a *de novo* transcriptome assembly program developed at the Broad Institute and the Hebrew University of Jerusalem. It comprises of three individual modules: Inchworm, Chrysalis and Butterfly (Grabherr et al. 2011) (Figure 3.1). Firstly, Inchworm analyses the short RNA-seq input reads and constructs a k-mer index with a default size of 25 nucelotides. The entire set of reads is then combined into a set of k-mers and any k-mers likely to represent errors are removed. Subsequently, the most commonly occurring k-mer sequence is identified as a "seed" k-mer and used for the construction of a putative transcript contig. The seed k-mer is extended repeatedly by a single base using the most abundant k-mer with a k-1 overlap, first from 5' to 3' and then in reverse, until no more k-mers can be used for extension. This process is then repeated until no more k-mers are left in the k-mer index. The resulting contigs therefore represent the most dominant variant of an expressed transcript, as they are based on the abundance of k-mers within the sample. Only contigs with an average k-mer coverage of 2 and therefore a length of at least 48 nucleotides (the default k-mer value is 25, one k-mer overlap will be k-1, therefore 2x(k-1)=48) proceed to the next software module, Chrysalis.

Chrysalis clusters the contigs generated by Inchworm and constructs a De Bruijn graph for each of them, representing all the transcriptional variations of a single gene or locus. Each edge of the De Bruijn graph is weighted based on the number of k-mers in the entire original set of reads that support the connection.

Finally, Butterfly takes the De Bruijn graph constructed by Chrysalis and first prunes the edges of the graph based on the support from the read data. In this way, incorrect transcript extensions are removed. Paths of the graphs having the most support based upon the original read sequences, paired-end data and the edge support weightings produced by Chrysalis are then selected as final transcript sequences.

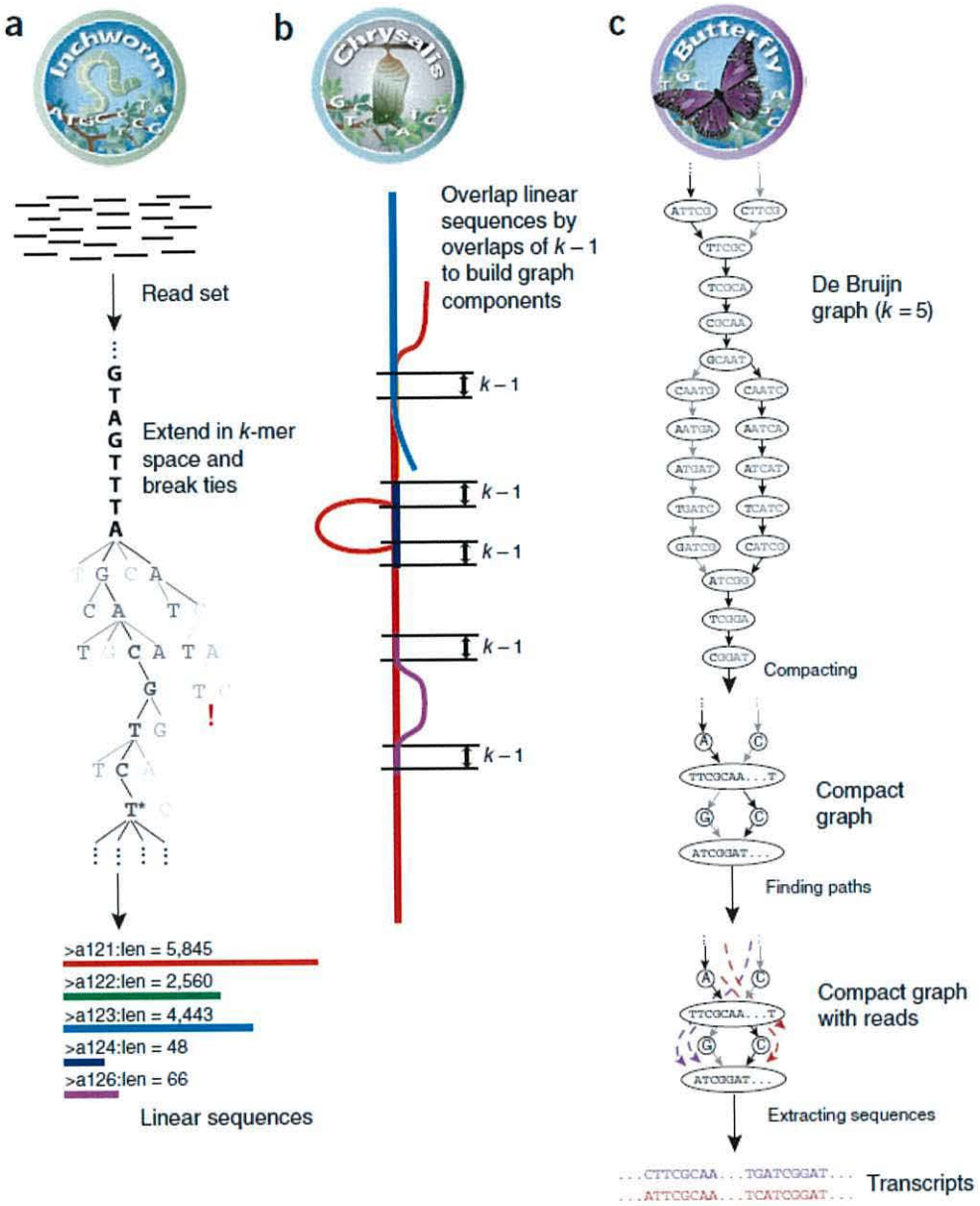


Figure 3.1. Graphical representation of the three modules of Trinity. Figure is taken from (Grabherr et al. 2011).

In summary, Trinity first constructs transcripts based on the extension of the most commonly occurring unique k-mer sequences present in a set of RNA-seq reads, constructs a De Bruijn graph for each of them which is weighted based on the number of k-mers in the original set of reads which support the edges of the graph. Each graph is then pruned to remove erroneous transcripts and the most highly supported paths of the graph are selected as transcript sequences.

3.3.2 SOAPdenovo-Trans

The *de novo* transcriptome assembler SOAPdenovo-Trans (Xie et al. 2014) is a recent assembly program and part of the SOAP (Short Oligonucleotide Analysis Package) *de novo* assembly softwares developed by members of the Beijing Genomics Institute (BGI). This program is similar to Trinity (section 3.3.1) and the majority of other short-read assemblers in that it utilises De Bruijn graph-based algorithms to construct transcript sequences.

Initially contigs are assembled using a De Bruijn graph method also employed by the genome assembler SOAPdenovo2 (Luo et al. 2012) with low-abundance k-mers being removed (as in Trinity). Subsequently the original reads are mapped back onto the constructed contigs and any linkages are used to merge/scaffold contigs together, with any incorrect linkages being removed. Contigs are then clustered into sub-graphs based upon shared exons, and an algorithm used by Oases (Schulz et al. 2012) is then used to generate the most probable transcripts from these graphs (thus constructing sequences for all supported transcript variants). Finally, any gaps between contigs are filled using paired-end information from the original sequencing reads (Xie et al. 2014). A graphical representation of this process is shown in Figure 3.2.

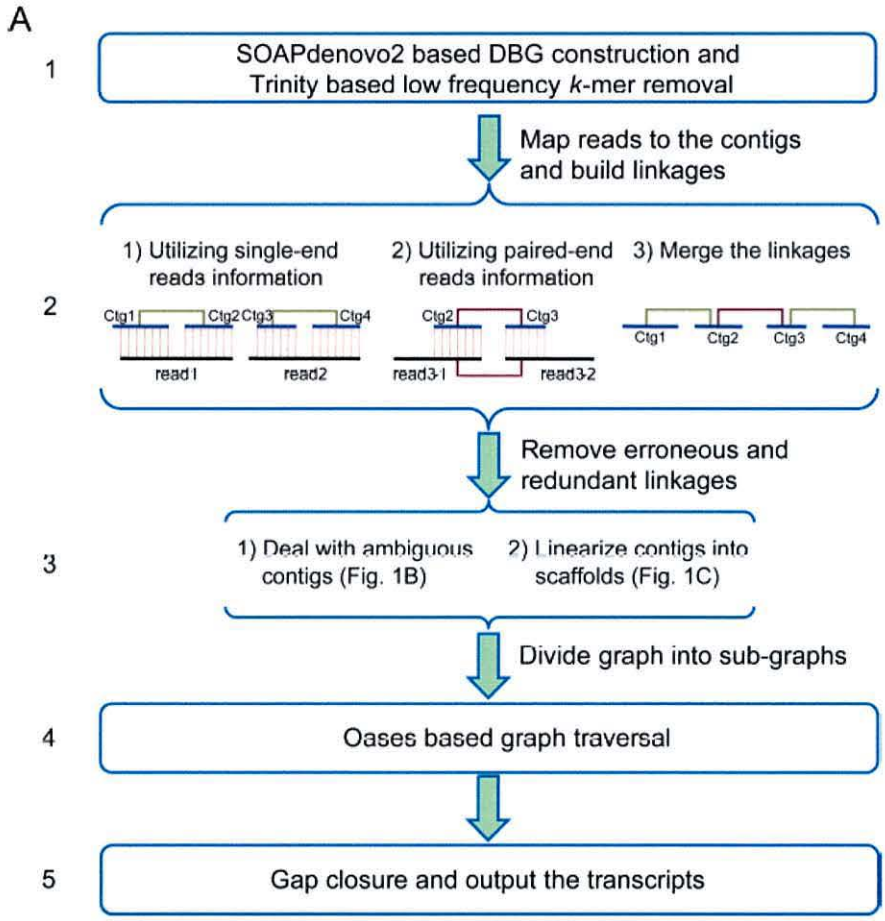


Figure 3.2. Graphical representation of the SOAPdenovo-Trans assembly process, taken from (Xie et al. 2014).

3.4 Transcript abundance estimation

Because of the quantitative nature of RNA-seq data it is possible to infer the relative abundance of a transcript within a sample, or indeed within several samples. Firstly, the original RNA-seq reads are aligned to the assembled transcripts. The assembled transcripts used will depend on the samples for which transcript expression level is to be estimated. For example, to examine the transcript expression level within a venom gland, the reads of that sample would be aligned to the assembled transcripts of that sample. However, using this process it is not statistically valid to compare two venom gland samples to each other (for example), as the assembled transcripts for each sample may differ depending on the read dataset. Therefore, to compare multiple samples it is necessary to first assemble a global transcriptome using all sets of RNA-seq reads belonging to each sample. A program such as RSEM (Li and Dewey 2011) can be used to estimate the expression of each transcript based upon the read alignments to constructed transcripts, followed by statistical inference of the maximum likelihood of transcript abundances. Put simply, the more reads aligned to a transcript, the higher that transcript is expressed within the sample (Figure 3.3).

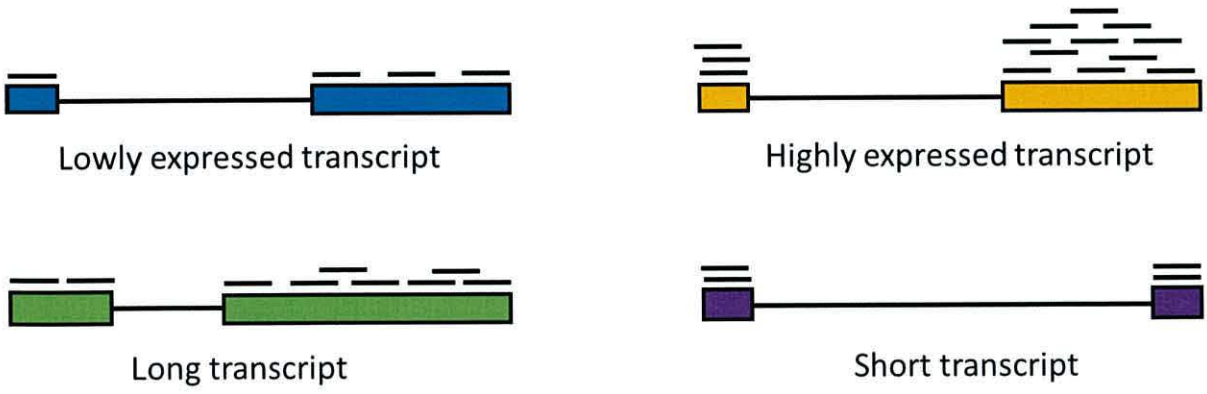


Figure 3.3. Representation of RNA-seq reads mapped to assembled transcripts, with the amount of reads mapped being proportional to the abundance of that transcript expressed in the transcriptome.

All transcript abundance estimation values are given in FPKM (Fragments Per Kilobase of exon per Million mapped reads) (Chandramohan et al. 2013), meaning all values are normalised both to the length of a transcript, and also the sequencing depth of the RNA-seq read dataset.

This process can be informative in identifying differentially expressed transcripts between samples or in identifying the relative expression level of a transcript (or variants thereof) within different tissues (Chapter 4). At face-value, this method can be likened to an *in silico* quantitative PCR (qPCR) experiment, only on the entire transcriptome rather than several target genes of interest. With this in mind, this approach was utilised to assess the expression of 13 different reference genes used for qPCR experiments, both across multiple tissues and in the venom gland at different timepoints following milking. The assessment of venom gene expression by qPCR is a relatively uncommon approach (Jeyaseelan et al. 2001; Currier et al. 2012), the most recent of which used β-actin and GAPDH (glyceraldehyde 3-phosphate dehydrogenase) as reference genes, despite them being shown on numerous occasions to be unsuitable for this purpose due to the variation in their expression across multiple tissues (Selvey et al. 2001; Glare et al. 2002; Radonić et al. 2004). As such, there remains a need to determine suitable housekeeping genes as candidate references to aid future experiments using this approach.

3.5 Minimum required sequencing depth for venom gland transcriptomic analysis

Whilst RNA-Seq is becoming increasingly common in studies seeking to characterise the venom gland transcriptome of different species of snake (Rokyta et al. 2012; Vonk et al. 2013; Margres et al. 2013), the depth of sequencing required in order to fully sequence a snake venom gland transcriptome has not been assessed. An estimated base level of the sequencing depth required will prove useful for planning future transcriptomic experiments, particularly when considering the number of libraries to sequence and the experimental costs this will incur.

This chapter aimed to evaluate several available methods for the assembly of transcriptomes of a range of reptile species and tissues. A total of 38 RNA libraries were prepared and sequenced using the Illumina sequencing platform derived from 6 reptile species including two venomous vipers (the painted saw-scaled viper, *Echis coloratus* and the Egyptian saw-scaled viper, *Echis pyramidum*), two colubrids (the corn snake, *Pantherophis guttatus* and the rough green snake, *Opheodrys aestivus*), a primitive boid (the royal python, *Python regius*) and a member of one of the most basal lineages of squamate reptiles (the leopard gecko, *Eublepharis macularius*).

RNA-seq data for the king cobra (*Ophiophagus hannah*) (Vonk et al. 2013) was also incorporated into analyses. Two *de novo* assembly programs (Trinity and SOAPdenovo-Trans) and one genome-guided assembly method (the Tuxedo suite) were used in order to evaluate approaches to transcriptome assembly. Transcipt abundance estimation was carried out as a downstream analysis on several samples in order to demonstrate its utility, in this case evaluating the expression of several reference genes used in quantitative PCR (qPCR) studies across a range of tissues and in the venom gland of *E. coloratus* at different time points following manual venom extraction. Finally, the newly generated venom gland transcriptomes were compared to pre-existing transcriptomes generating using ESTs, and sub-assemblies of the newly generated data were carried out in order to estimate the minimum required sequencing depth to fully sequence a snake venom gland.

3.6 Methods

3.6.1 Tissue sampling and RNA extraction

All research involving animals was carried out in accordance with institutional and national guidelines and was approved by the Bangor University Ethical Review Committee. All study animals were sacrificed according to Schedule 1 procedures as stipulated in The Animals (scientific procedures) Act 1986. Where possible all body tissues were dissected and preserved for use in future experiments in accordance with the 3Rs (Replacement, Reduction and Refinement) which encourage and offer guidelines to minimise the use of animals in scientific research (Guhad 2005). All tissue samples were snap frozen immediately in either liquid Nitrogen or dry ice and stored at -80°C until required.

Total RNA was extracted from tissues using the RNeasy mini kit (Qiagen) with on-column DNase digestion and eluted into 30µl of RNase-free water. RNA samples were then quantified using a Qubit fluorometer.

The venom glands of four adult painted saw-scaled vipers (*Echis coloratus*) were sampled, each having been "milked" (venom extracted manually, see Figure 3.4) at specific timepoints prior to sacrifice, namely one sample 16 hours post-milking, two samples 24 hours post-milking and one sample 48 hours post-milking. The venom glands of an adult Egyptian saw-scaled viper (*Echis pyramidum*) were also taken 24 hours post-milking.

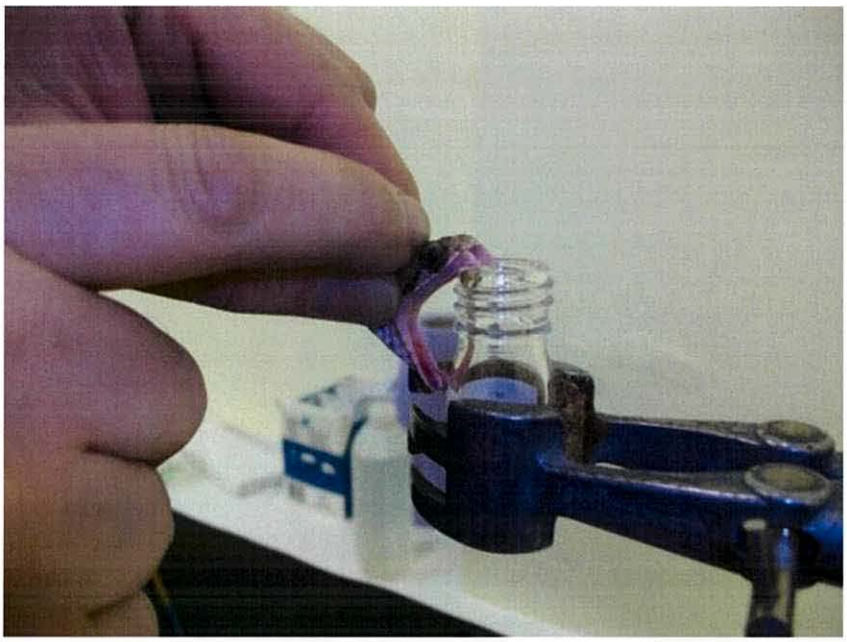


Figure 3.4. Photograph of an adult saw-scaled viper being "milked".

Also sampled were the cloacal scent glands, skin, brain and kidney of two adult *E. coloratus*, and the liver and ovary of a single individual of this species. The salivary glands, cloacal scent glands and skin were also sampled from two individuals of the leopard gecko (*Eublepharis macularius*), royal python (*Python regius*), rough green snake (*Opheodrys aestivus*) and corn snake (*Pantherophis guttatus*).

3.6.2 Library preparation and sequencing

All library preparation was carried out using the Illumina TruSeq sample preparation kit with a selected fragment size of 200-500bp. A graphical representation of sample preparation workflow can be seen in Figure 3.5.

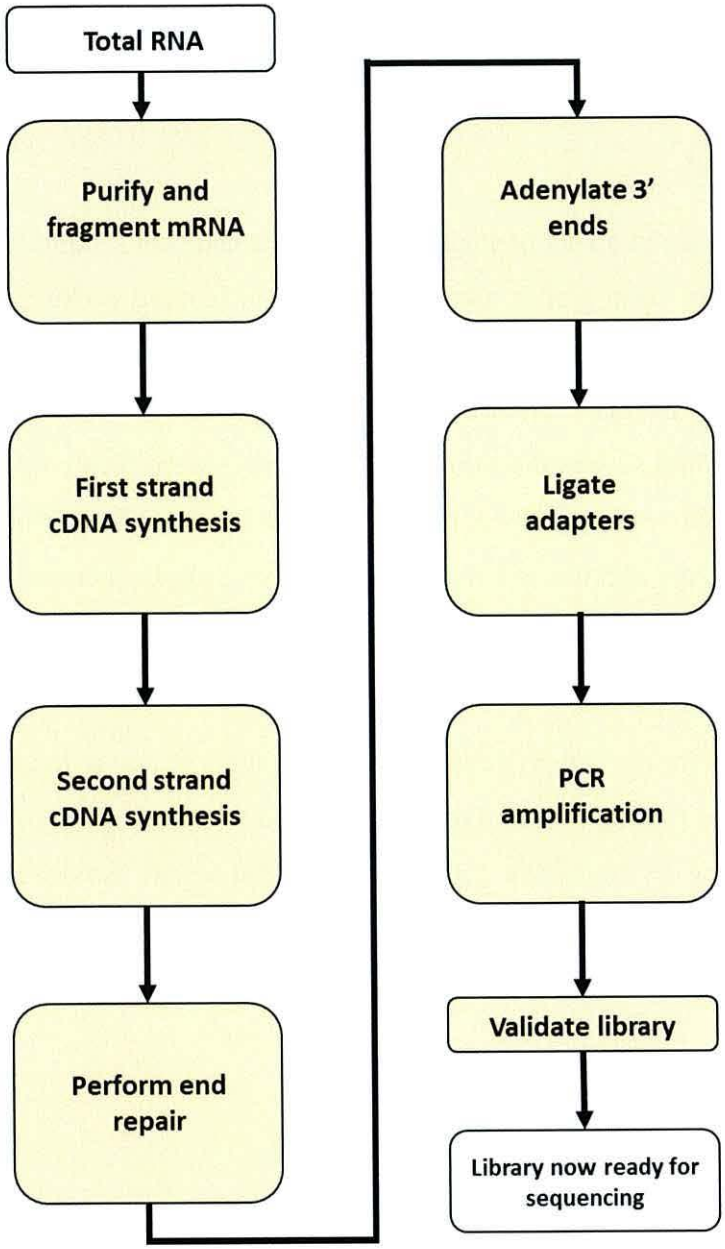


Figure 3.5. Workflow of Illumina TruSeq RNA library preparation, adapted from Figure 8 in the Illumina TruSeq sample preparation kit guide.

Briefly, mRNA is purified from total RNA through two rounds of poly-A purification using Poly-T oligo-attached magnetic beads and then fragmented. First strand and second strand cDNA synthesis is then carried out, with the double-stranded cDNA being subsequently purified using Agencourt AMPure XP magnetic beads (Beckman Coulter). From this point the workflow is very similar to that of the TruSeq DNA library preparation (Chapter 2). The ends of the cDNA molecules are repaired to remove any 5' or 3' overhangs and the 3' ends are adenylated to prevent fragments from ligating to one another forming chimeras during the

adapter ligation step. Adapters are then ligated onto the cDNA fragments and then the library is enriched for fragments which have adapters ligated to them by polymerase chain reaction (PCR). A PCR clean-up is then performed, again using AMPure XP beads and the library is validated and quantified.

The RNA-seq libraries for two venom glands, scent glands and skin samples of Echis coloratus and two salivary glands, scent glands and skin samples of leopard gecko, royal python, corn snake and rough green snake were prepared in collaboration with Dr. Darren Logan at the Wellcome Trust Sanger institute. One skin sample from corn snake was not of sufficient quality for library preparation and so was discarded. These libraries were then pooled and sequenced using 100bp paired-end reads on three lanes of the Illumina HiSeq2000 platform, giving roughly $1/5^{th}$ of a lane per species and $1/10^{th}$ of a lane per sample. Transcriptome assemblies for these sets of sequencing reads were assembled using Trinity (Grabherr et al. 2011) by Dr. Martin Swain at the University of Aberystwyth.

The RNA-seq libraries for the venom gland, brain, and kidney of two individuals and liver and ovary of one individual *E. coloratus* and the venom gland of one *E. pyramidum* were prepared using the same method by myself at the Institute of Biological, Environmental and Rural Sciences (IBERS) phenomics centre at the University of Aberystwyth. These libraries were then loaded by myself and sequenced using 100bp paired-end reads on the Illumina HiSeq2500 platform.

3.6.3 De novo transcriptome assembly using Trinity

All Trinity assemblies were assembled as paired-end datasets using the default k-mer value (k=25).

```
Trinity.pl --seqType fq --JM 400G --left Left_reads.fastq --right
Right_reads.fastq --CPU 32 --output Transcriptome_assembly
```

The full details of the Trinity assembly protocol are now published (Haas et al. 2013).

3.6.4 De novo transcriptome assembly using SOAPdenovo-Trans

Assemblies using SOAPdenovo-Trans were carried out using the following command:

```
./SOAPdenovo-Trans-31mer all -s config_file -o outputGraph
```

Where `config_file` refers to a text file containing the read length, average insert size and file locations of the RNA-seq reads to be assembled.

3.6.5 Genome-guided assembly (Tuxedo suite)

Transcriptomes for the king cobra (*Ophiophagus hannah*), painted saw-scaled viper (*Echis coloratus*) and the royal python (*Python regius*) were assembled using the king cobra genome (Vonk et al. 2013), the *Echis coloratus* genome (Chapter 2) and the Burmese python genome (Castoe et al. 2013) as reference sequences. Due to poor assembly quality, a lack of gene-sized scaffolds, and a low number of CEGs in the *Echis pyramidum* and corn snake draft genome assemblies, the transcriptomes of these species using the aforementioned genome sequences as a reference were not assembled using this method as it is likely that the assembled transcripts would be of poor quality. Genome-guided assembly was carried out using the Tuxedo suite as described in (Trapnell et al. 2012).

Firstly, each reference genome assembly was indexed using the short-read aligner Bowtie (Langmead and Salzberg 2012) using the following command:

```
bowtie-build Genome_assembly.fasta Genome_index
```

RNA-Seq reads were then mapped to each genome assembly using TopHat (Trapnell et al. 2009) to give a .bam file with the default file name of `accepted_hits.bam`.

```
tophat -r 100 Genome_index Left_reads.fastq Right_reads.fastq
```

For the assembly of both *E. coloratus* kidney datasets, there were issues with the process timing-out on the computer server. As a result, the .bam mapping files for each individual kidney dataset were merged using Samtools (Li et al. 2009) using the following command:

```
Samtools merge merged.bam Ec6kidney.bam Eco11kidney.bam
```

Cufflinks (Trapnell et al. 2010) was then used to construct putative transcripts producing a Gene transfer format (.gtf) file.

```
cufflinks accepted_hits.bam
```

As the .gtf file only contains feature information of the transcript (e.g. the positions of exons) and does not contain any nucleotide sequence, the module "gffread" of Cufflinks was then used to construct nucleotide contig files from the reference genome assembly based upon the information contained in the .gtf file using the command:

```
gffread -w Constructed_transcripts.fasta -g Genome_assembly.fasta
transcripts.gtf
```

The .fasta file `Constructed_transcripts.fasta` will now contain assembled contig sequences for the transcriptome.

3.6.6 Transcript abundance estimation using RSEM

In order to allow a comparison of transcript expression between tissues, a global transcriptome assembly was first generated on a per tissue per species basis using Trinity (see above section). That is to say that all samples whose gene expression level is going to be analysed were incorporated into one transcriptome assembly. Transcript abundance estimation was then carried out using RSEM (Li and Dewey 2011) as a downstream analysis of Trinity:

```
run_RSEM_align_n_estimate.pl --transcripts Global_assembly.fasta
--left Left_reads.fastq --right Right_reads.fastq --seqType fq --
prefix RSEM_output
```

This then gives multiple files as output, including the results files `RSEM.genes.results` and `RSEM.isoforms.results`. Both of these files are roughly similar and contain transcript abundance estimation values given in FPKM (Fragments Per Kilobase of exon per Million mapped reads), except the "genes" file contains results where transcript abundance estimation values are given based on the clustering of sequences based on sequence similarity. As Eukaryotic genes can be subject to alternative splicing, the "isoforms" results file was used in all cases in order to gain a full picture of individual transcript expression.

Transcript abundance estimation values were obtained by first identifying contigs of interest in the global transcriptome assembly by local BLAST searches using BLAST+ v2.2.27 (Camacho

et al. 2009). The contig name was the used to cross-reference between the global assembly and the isoforms results file in order to locate the correct transcript abundance estimation value.

3.6.7 Evaluating putative toxin-encoding transcripts in Echis venom gland transcriptomes

Putatitve toxin-encoding transcripts were identified using local BLAST searches carried out with BLAST+ v2.2.27 (Camacho et al. 2009). Contigs were then annotated using the NCBI BLAST server (http://blast.st-va.ncbi.nlm.nih.gov/Blast.cgi) and the Expasy translate tool (http://web.expasy.org/translate/). Transcripts belonging to large gene families (such as snake venom metalloproteinases (SVMPs)) were further identified through phylogenetic analysis to ensure accurate identification of gene paralogs. Putative toxin amino acid sequences were aligned using ClustalW (Larkin et al. 2007) and maximum likelihood trees were constructed using the Jones-Taylor-Thornton (JTT) model (determined as the best-fitting model of protein sequence evolution using ProtTest 3 (Darriba et al. 2011)) with 500 Bootstrap replicates in MEGA5 (Tamura et al. 2011). All nodes below 50% Bootstrap support were collapsed.

3.6.8 Sub-assemblies of the E. coloratus venom gland transcriptome

Paired venom gland reads were first interleaved using the `shuffleSequences.pl` perl script which is a part of the Velvet *de novo* assembly program (Zerbino and Birney 2008). This results in each read pair being maintained during the sub-sampling process. Using the Linux commands `head` and `tail`, 3 sub-sets (designated as "head", "middle" and "tail") of either 2, 4, 8 or 10 million reads were taken from the venom gland RNA-seq reads dataset for Eco7 (Figure 3.6). These reads in particular were used as they were sequenced on the more recent HiSeq2500 platform, and due to the large size of this dataset (44,678,609 paired-end reads) there would not be any overlap in sub-samples taken unlike in the smaller venom gland datasets. The below example samples the top 10 million sequencing reads from the file `file.fasta` and places them in a new file, `headsample10mil.fasta`. As .fastq files (Chapter 1) contain reads placed on 4 individual lines, meaning 8 lines will contain a set of paired reads, multiples of 8 must be used in order to select the correct number of reads.

```
Head -n 80000000 file.fasta >  headsample10mil.fasta
```

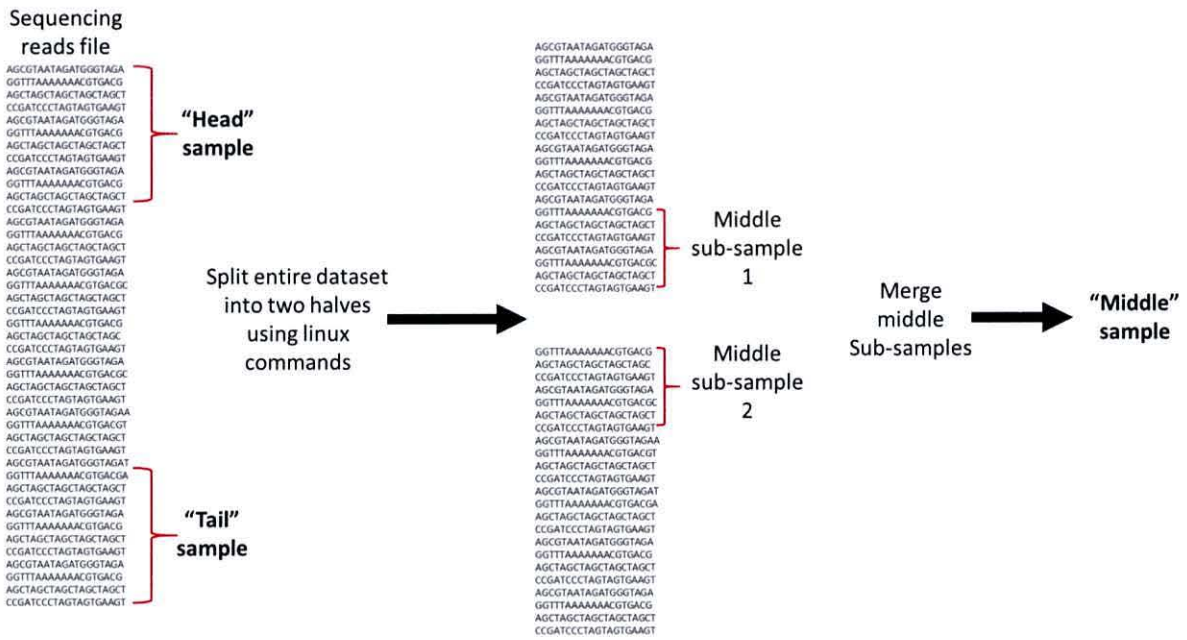


Figure 3.6. Graphical representation of sampling technique used to generate 3 sub-samples of venom gland RNA-Seq reads.

These data were assembled using Trinity (Grabherr et al. 2011; Haas et al. 2013), with parameters set to run as a single-end read dataset (as there is only one .fastq input file), but with the added command-line parameter `--run_as_paired` to indicate that the data contains paired-end data. Local blast surveys were then carried out using BLAST+ version 2.2.27 (Camacho et al. 2009) to identify previously characterised putative toxin genes in *E. coloratus* (see next chapter). Only matches above 75 amino acids were considered (the length of the shortest amino acid query sequence) and presence/absence of gene sequences was recorded. Also recorded was the length of the homologous amino acid sequence and the length of the newly assembled sequence was also converted into the percentage of the query sequence covered (to aid in evaluating the degree of "completeness" of the assembled sequence). The percentage similarity of the assembly sequence to the query sequence was also recorded (to evaluate any occurrences of sequencing errors being including in the assembly and to identify misassembled sequences).

3.7 RNA-seq raw sequencing output

RNA sequencing produced a total of roughly 50.3 Gigabases (Gb) of raw sequence for *Echis coloratus*, 11.6Gb for *Echis pyramidum*, 14Gb for leopard gecko, 13.6Gb for royal python, 13.9Gb for rough green snake and 11.9Gb for corn snake. Full raw sequencing metrics per sample can be seen in Table 3.1. Per tissue, the highest amount of sequencing reads was

generated for the venom gland of *E. coloratus* with approximately 110 million 100bp paired-end reads (Figure 3.7). The lowest amount of sequencing data was generated for the liver of *E. coloratus* with only ~7 million paired end reads. Nevertheless, this is still roughly 1.5x more sequencing than the largest RNA-seq dataset for the king cobra, sequenced from pooled tissues (Table 3.2) (Vonk et al. 2013).

Table 3.1. RNA-seq raw sequence metrics

Species	ID	Tissue	Total Paired-End reads	Total bases
Echis coloratus	Eco6	Venom gland	13,468,544	2,693,708,800
Echis coloratus	Eco7	Venom gland	44,678,609	9,025,079,018
Echis coloratus	Eco8	Venom gland	38,711,180	7,819,658,360
Echis coloratus	Eco215	Venom gland	13,173,683	2,634,736,600
Echis coloratus	Eco2	Scent gland	13,814,547	2,762,909,400
Echis coloratus	Eco3	Scent gland	13,392,440	2,678,488,000
Echis coloratus	Eco1	Skin	7,474,858	1,494,971,600
Echis coloratus	Eco2	Skin	6,691,562	1,338,312,400
Echis coloratus	Eco6	Brain	16,357,991	3,304,314,182
Echis coloratus	Eco11	Brain	15,576,893	3,146,532,386
Echis coloratus	Eco6	Kidney	15,446,045	3,120,101,090
Echis coloratus	Eco11	Kidney	25,702,056	5,191,815,312
Echis coloratus	Eco6	Liver	7,095,517	1,433,294,434
Echis coloratus	Eco6	Ovary	18,155,364	3,667,383,528
Echis pyramidum	Epy3	Venom gland	57,398,601	11,594,517,402
Eublepharis macularius	Ema3	Salivary gland	14,989,388	2,997,877,600
Eublepharis macularius	Ema2	Salivary gland	14,892,722	2,978,544,400
Eublepharis macularius	Ema2	Scent gland	12,955,313	2,591,062,600
Eublepharis macularius	Ema3	Scent gland	12,547,208	2,509,441,600
Eublepharis macularius	Ema1	Skin	7,249,250	1,449,850,000
Eublepharis macularius	Ema3	Skin	7,426,318	1,485,263,600
Python regius	Pre1	Salivary gland	15,009,931	3,001,986,200
Python regius	Pre3	Salivary gland	13,025,114	2,605,022,800
Python regius	Pre1	Scent gland	12,558,278	2,511,655,600
Python regius	Pre2	Scent gland	12,016,725	2,403,345,000
Python regius	Pre1	Skin	7,484,993	1,496,998,600
Python regius	Pre2	Skin	8,158,279	1,631,655,800
Opheodrys aestivus	Oae1	Salivary gland	11,272,761	2,254,552,200
Opheodrys aestivus	Oae2	Salivary gland	13,686,481	2,737,296,200
Opheodrys aestivus	Oae1	Scent gland	15,596,633	3,119,326,600
Opheodrys aestivus	Oae2	Scent gland	12,539,513	2,507,902,600
Opheodrys aestivus	Oae2	Skin	7,484,248	1,496,849,600
Opheodrys aestivus	Oae3	Skin	8,914,677	1,782,935,400
Pantherophis guttatus	Pgu1	Salivary gland	11,903,255	2,380,651,000
Pantherophis guttatus	Pgu2	Salivary gland	13,752,406	2,750,481,200
Pantherophis guttatus	Pgu4	Scent gland	13,141,388	2,628,277,600
Pantherophis guttatus	Pgu1	Scent gland	12,840,821	2,568,164,200
Pantherophis guttatus	Pgu1	Skin	7,862,371	1,572,474,200

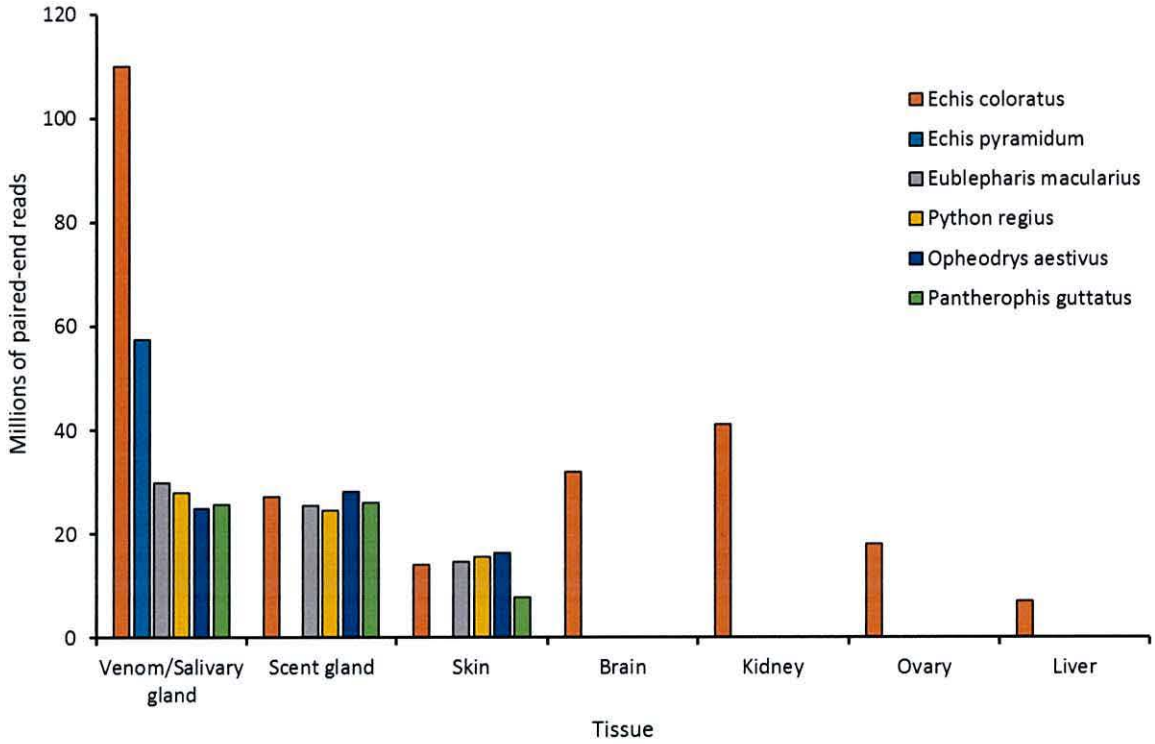


Figure 3.7 Total number of paired-end RNA-Seq reads sequenced per tissue per species.

Table 3.2. Raw RNA sequencing metrics for 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). Metrics are representative of single-end 50bp reads.

Species	Tissue	Total number of SE reads	Total number of bases
Ophiophagus hannah	Venom gland	15,166,590	834,162,450
Ophiophagus hannah	Accessory gland	11,209,677	616,532,235
Ophiophagus hannah	Pooled tissue	17,858,289	910,772,739

3.8 Individual transcriptome assembly metrics

Individual transcriptome assemblies assembled using SOAPdenovo-Trans have a mean contig N50 of 1,152 bp, slightly lower than those assembled using the Tuxedo suite (1,509bp) and Trinity (1,771bp). Both SOAPdenovo-Trans and Trinity appear to have performed poorly when assembling transcriptomes from the king cobra, with assemblies from both programs having lower contig N50 values when compared to assemblies from other species (Tables 3.3 and 3.4), and the majority of contigs for king cobra being less than 300bp in length (86.9% of contigs assembled by SOAPdenovo-Trans for king cobra are less than 300bp). However, king cobra assemblies using the Tuxedo suite are greatly improved, with contig N50 values, maximum contig lengths and number of contigs over 300bp in length being comparable to other species assemblies using this method (Table 3.5) and having similar contig N50 values to other transcriptome assemblies constructed using the *de novo* programs (Tables 3.3 and 3.4).

Table 3.3. SOAPdenovo-Trans individual transcriptome assembly metrics

Assembly	Number of contigs	Number of contigs ≥300nt	Total length	Max contig length	Contig N50
Eco6VG	136,903	33,826	30,011,346	8,331	1,175
Eco7VG	155,646	42,025	41,223,916	13,550	1,371
Eco8VG	169,750	42,258	41,232,390	12,403	2,034
Eco215 VG	151,852	36,037	29,882,067	15,323	1,062
Eco2 SCG	199,985	45,260	41,486,729	27,377	1,266
Eco3 SCG	248,492	53,432	45,190,221	13,742	1,099
Eco1 SK	114,460	27,586	22,503,588	10,559	1,038
Eco2 SK	132,109	32,262	27,296,276	30,044	1,087
Eco6 brain	342,519	66,758	58,712,652	15,506	1,183
Eco11 brain	355,848	67,390	58,328,305	13,956	1,148
Eco6 kidney	255,687	47,925	37,816,286	8,705	987
Eco11 kidney	238,119	49,230	39,563,215	9,096	1,119
Eco liver	87,269	18,486	12,793,748	12,189	799
Eco ovary	266,331	46,382	40,134,187	13,854	1,157
Epy VG	268,209	61,246	61,223,863	23,791	1,475
Pre1 SAL	144,300	34,075	31,041,633	12,880	1,237
Pre3 SAL	89,713	24,886	23,437,955	30,323	1,302
Pre1 SCG	189,582	47,362	47,085,293	10,181	1,471
Pre2 SCG	260,169	59,862	57,128,948	19,357	1,386
Pre1 SK	98,985	27,716	25,243,714	25,860	1,234
Pre2 SK	116,414	31,066	27,066,072	9,133	1,154
Pgu1 SK	129,096	29,940	23,845,304	19,384	991
Oha VG	55,185	7,024	4,405,541	6,659	674
Oha AG	91,359	11,966	7,374,557	6,427	656
Oha PT	139,792	18,663	11,783,875	7,150	686

Abbreviations

Eco, *Echis coloratus*; Epy, *Echis pyramidum*; Pre, *Python regius*; Pgu, *Pantherophis guttatus*; Oha, *Ophiophagus hannah*; VG, venom gland; SCG, scent gland; SK, skin; SAL, salivary gland; AG, accessory gland; PT, pooled tissue.

Table 3.4. Trinity individual transcriptome assembly metrics

Assembly	Number of contigs	Number of contigs ≥300nt	Total length	Max contig length	Contig N50
Eco6VG	59,176	44,470	47,908,742	9,014	1,619
Eco7VG	72,926	51,505	68,062,037	21,683	2,232
Eco8VG	77,119	53,786	72,871,466	16,826	2,338
Eco215 VG	64,576	48,321	50,459,031	15,335	1,515
Eco2 SCG	90,968	68,056	87,303,833	45,270	2,217
Eco3 SCG	115,012	85,888	103,705,361	15,321	2,002
Eco1 SK	51,201	38,009	35,737,402	15,572	1,372
Eco2 SK	62,131	44,967	45,259,422	30,271	1,525
Eco6 brain	85,392	78,074	112,985,184	16,650	2,607
Eco11 brain	86,336	79,495	112,070,564	15,715	2,508
Eco6 kidney	62,576	51,969	53,591,763	15,647	1,600
Eco11 kidney	60,116	54,512	60,726,571	11,043	1,755
Eco liver	31,205	18,169	13,899,146	9,939	950
Eco ovary	81,682	52,264	62,991,109	16,376	2,023
Epy VG	127,519	86,953	134,880,355	29,658	2,898
Ema3 SAL	76,930	57,737	69,820,804	15,697	1,970
Ema2 SAL	87,501	65,506	84,949,737	25,554	2,238
Ema2 SCG	94,871	71,922	88,932,557	25,346	2,113
Ema3 SCG	89,798	67,528	80,371,672	15,226	1,987
Ema1 SK	76,874	56,634	59,240,486	27,092	1,671
Ema3 SK	57,621	44,591	47,491,701	15,482	1,679
Pre1 SAL	63,822	47,081	54,466,186	12,891	1,911
Pre3 SAL	43,260	32,227	34,554,660	33,666	1,693
Pre1 SCG	94,292	71,210	93,946,193	15,309	2,344
Pre3 SCG	125,508	94,451	139,581,641	20,963	2,861
Pre1 SK	47,067	36,885	39,101,411	25,880	1,652
Pre2 SK	53,447	41,398	42,546,841	12,742	1,560
Oae1 SAL	54,867	40,187	40,660,491	14,506	1,506
Oae2 SAL	41,651	30,649	29,951,957	16,883	1,417
Oae1 SCG	89,298	65,779	70,122,643	25,175	1,672
Oae2 SCG	97,282	72,520	89,220,322	25,173	2,065
Oae2 SK	67,511	49,908	53,842,593	26,000	1,679
Oae3 SK	66,641	49,603	50,736,429	35,727	1,559
Pgu1 SAL	52,904	39,304	40,809,074	10,550	1,551
Pgu2 SAL	50,928	38,408	41,516,899	13,241	1,634
Pgu1 SCG	92,325	68,714	88,975,330	18,321	2,243
Pgu4 SCG	68,912	52,002	60,321,648	17,796	1,918

Pgu 1SK	46,783	35,969	33,270,877	19,348	1,329
Oha VG	6,123	2,925	1,690,039	4,585	424
Oha AG	9,046	4,113	2,198,877	3,740	377
Oha PT	8,877	4,135	2,420,103	5,733	413

Abbreviations

Eco, *Echis coloratus*; Epy, *Echis pyramidum*; Ema, *Eublepharis macularius*; Pre, *Python regius*; Oae, *Opheodrys aestivus*; Pgu, *Pantherophis guttatus*; Oha, *Ophiophagus hannah*; VG, venom gland; SCG, scent gland; SK, skin; SAL, salivary gland; AG, accessory gland; PT, pooled tissue.

Table 3.5. Tuxedo suite individual transcriptome assembly metrics

Assembly	% of reads mapped to genome	Number of contigs	Number of contigs ≥300nt	Total length	Max contig length	Contig N50
Eco6VG	77.1	33,917	28,941	36,996,675	14,002	1,625
Eco7VG	65.4	48,484	36,992	46,791,823	29,852	1,651
Eco8VG	68.4	48,912	37,607	48,461,097	14,007	1,683
Eco215 VG	74.8	34,706	30,047	37,025,610	12,374	1,529
Eco2 SCG	78.5	44,791	38,032	50,553,243	28,663	1,810
Eco3 SCG	74.9	48,983	42,123	54,564,734	14,068	1,714
Eco1 SK	77.0	26,773	22,519	25,802,727	14,033	1,442
Eco2 SK	77.2	31,859	27,151	32,411,872	29,866	1,521
Eco6 brain	77.4	63,713	53,857	75,117,795	17,657	1,866
Eco11 brain	74.8	64,178	54,504	72,929,178	15,957	1,758
Eco6 kidney	74.0	44,301	37,525	46,423,284	13,814	1,554
Eco11 kidney	73.3	47,510	39,333	48,063,704	10,747	1,534
Eco liver	64.7	16,711	14,480	15,633,227	8,502	1,243
Eco ovary	62.4	49,138	40,169	49,864,920	19,028	1,570
Pre1 SAL	28.1	27,175	24,029	27,147,810	8,733	1,428
Pre3 SAL	36.1	20,095	18,029	20,252,937	24,614	1,387
Pre1 SCG	55.7	40,677	33,817	37,972,806	8,980	1,470
Pre2 SCG	55.4	45,133	37,701	44,075,685	21,582	1,560
Pre1 SK	61.7	23,813	20,232	21,050,794	18,606	1,282
Pre2 SK	51.9	26,656	22,059	21,768,397	6,575	1,223
Oha VG	73.4	27,097	20,316	19,786,839	16,289	1,155
Oha AG	82.0	41,559	31,931	32,675,866	16,287	1,269
Oha PT	87.5	62,540	46,923	50,849,001	14,774	1,429

Abbreviations

Eco, *Echis coloratus*; Pre, *Python regius*; Oha, *Ophiophagus hannah*; VG, venom gland; SCG, scent gland; SK, skin; SAL, salivary gland; AG, accessory gland; PT, pooled tissue.

3.9 Tissue transcriptome assemblies

Table 3.6. SOAPdenovo-Trans tissue transcriptome assembly metrics

Assembly	Number of contigs	Number of contigs ≥300nt	Total length	Max contig length	Contig N50
Eco6+215 VG	208,027	47,856	40,739,243	15,323	1,114
Eco VG	321,670	68,296	61,290,073	27,004	1,214
Eco SCG	328,080	68,343	60,085,069	28,505	1,174
Eco SK	182,045	41,396	36,605,611	30,057	1,177
Eco brain	519,830	92,660	77,416,630	15,107	1,076
Eco kidney	365,658	67,740	52,916,929	11,895	963
Eco liver	87,269	18,486	12,793,748	12,189	799
Eco ovary	266,331	46,382	40,134,187	13,854	1,157
Epy VG	268,209	61,246	61,223,863	23,791	1,475
Ema SAL	258,148	60,605	58,361,787	16,824	1,355
Ema SCG	308,270	68,336	62,406,643	16,807	1,253
Ema SK	233,201	49,761	46,639,599	18,122	1,320
Pre SAL	166,040	39,982	40,350,565	30,323	1,479
Pre SCG	331,713	78,888	73,916,376	19,367	1,332
Pre SK	150,160	39,338	37,029,241	25,860	1,313
Oae SAL	174,939	36,484	30,588,247	9,276	1,070
Oae SCG	350,966	64,552	51,428,274	13,076	988
Oae SK	256,589	46,040	35,834,796	25,935	955
Pgu SAL	169,037	38,954	36,422,642	13,734	1,277
Pgu SCG	276,353	57,443	51,619,469	15,168	1,212
Pgu SK	129,096	29,940	23,845,304	19,384	991
Oha VG	55,185	7,024	4,405,541	6,659	674
Oha AG	91,359	11,966	7,374,557	6,427	656
Oha PT	139,792	18,663	11,783,875	7,150	686

Abbreviations

Eco, *Echis coloratus*; Epy, *Echis pyramidum*; Pre, *Python regius*; Pgu, *Pantherophis guttatus*; Oha, *Ophiophagus hannah*; VG, venom gland; SCG, scent gland; SK, skin; SAL, salivary gland; AG, accessory gland; PT, pooled tissue.

Table 3.7. Trinity tissue transcriptome assembly metrics

Assembly	Number of contigs	Number of contigs ≥300nt	Total length	Max contig length	Contig N50
Eco6+215VG	84,846	56,805	60,946,763	15,819	1,682
Eco VG	117,125	81,798	121,590,983	30,131	2,623
Eco SCG	138,852	87,389	108,950,322	36,612	2,214
Eco SK	77,402	50,860	53,460,258	30,610	1,693
Eco brain	195,958	134,236	255,562,314	20,155	3,552
Eco kidney	120,728	76,660	93,044,842	12,930	2,044
Eco liver	31,205	18,169	13,899,146	9,939	950
Eco ovary	81,682	52,264	62,991,109	16,376	2,023
Epy VG	127,519	86,953	134,880,355	29,658	2,898
Ema SAL	111,345	73,027	92,083,854	24,285	2,247
Ema SCG	129,951	85,014	102,282,153	29,392	2,173
Ema SK	92,506	61,456	68,881,052	27,091	1,958
Pre SAL	73,492	48,727	59,038,693	33,655	2,122
Pre SCG	163,065	104,329	141,505,606	20,966	2,730
Pre SK	67,200	47,819	52,575,882	25,879	1,796
Oae SAL	65,393	42,558	45,978,497	17,524	1,700
Oae SCG	126,321	77,954	88,074,109	17,135	1,920
Oae SK	92,597	57,242	61,535,012	33,155	1,761
Pgu SAL	64,595	43,565	50,371,460	17,102	1,916
Pgu SCG	110,016	70,265	84,796,474	17,065	2,345
Pgu 1SK	46,783	35,969	33,270,877	19,348	1,329
Oha VG	6,123	2,925	1,690,039	4,585	424
Oha AG	9,046	4,113	2,198,877	3,740	377
Oha PT	8,877	4,135	2,420,103	5,733	413

Abbreviations

Eco, *Echis coloratus*; Epy, *Echis pyramidum*; Ema, *Eublepharis macularius*; Pre, *Python regius*; Oae, *Opheodrys aestivus*; Pgu, *Pantherophis guttatus*; Oha, *Ophiophagus hannah*; VG, venom gland; SCG, scent gland; SK, skin; SAL, salivary gland; AG, accessory gland; PT, pooled tissue.

Table 3.8. Tuxedo suite tissue transcriptome assembly metrics

Assembly	% of reads mapped to genome	Number of contigs	Number of contigs ≥300nt	Total length	Max contig length	Contig N50
Eco6+215VG	76.1	46,547	39,527	51,260,072	14,042	1,649
Eco VG	58.9	75,068	58,431	78,424,107	29,852	1,758
Eco SCG	76.8	64,381	54,672	75,897,775	29,865	1,833
Eco SK	77.1	43,889	36,709	46,275,654	29,866	1,594
Eco brain	76.2	86,972	73,044	102,051,162	18,097	1,861
Eco kidney	73.7	63,331	52,446	64,557,016	14,019	1,615
Eco liver	64.7	16,711	14,480	15,633,227	8,502	1,243
Eco ovary	62.4	49,138	40,169	49,864,920	19,028	1,570
Pre SAL	31.9	34,097	29,575	36,428,378	24,614	1,577
Pre SCG	55.7	56,292	45,183	55,260,875	21,582	1,598
Pre SK	56.7	34,992	28,727	32,182,773	18,606	1,379
Oha VG	73.4	27,097	20,316	19,786,839	16,289	1,155
Oha AG	82.0	41,559	31,931	32,675,866	16,287	1,269
Oha PT	87.5	62,540	46,923	50,849,001	14,774	1,429

Abbreviations

Eco, *Echis coloratus*; Pre, *Python regius*; Oha, *Ophiophagus hannah*; VG, venom gland; SCG, scent gland; SK, skin; SAL, salivary gland; AG, accessory gland; PT, pooled tissue.

3.10 *De novo* assembly methods

Out of the two *de novo* assembly programs, Trinity appears to consistently produce assemblies with a considerably higher contig N50 value when compared to SOAPdenovo-Trans (Figure 3.8).

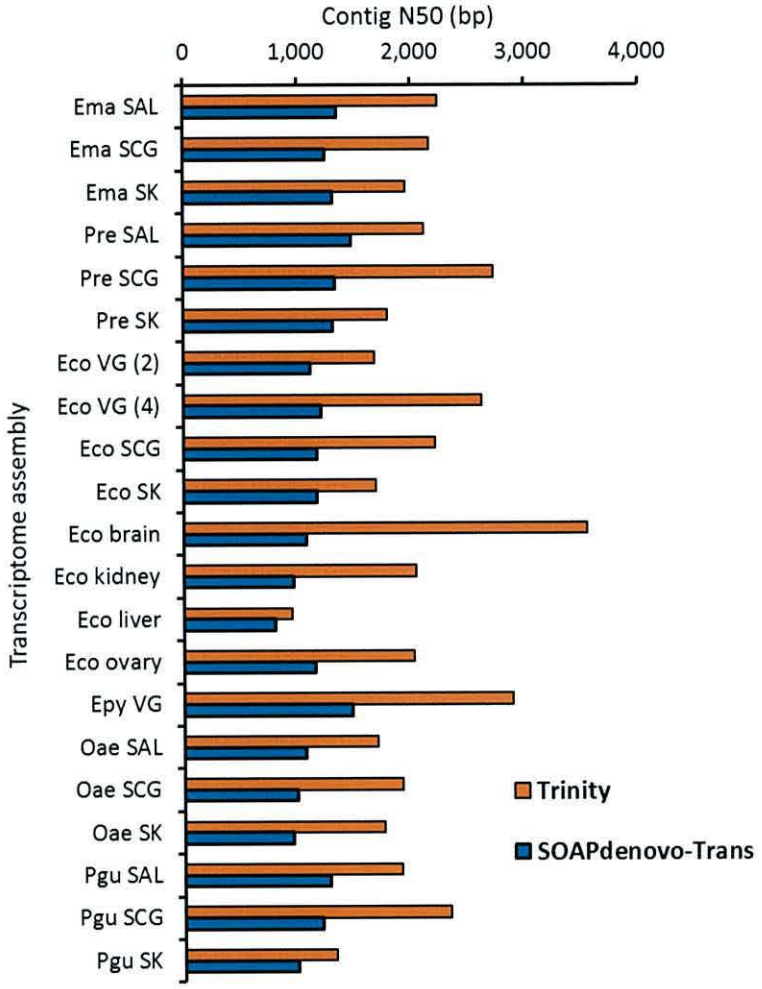


Figure 3.8. Contig N50 values of Trinity and SOAPdenovo-Trans assemblies. N50 values for Trinity assemblies are higher in all cases. Eco VG (2) was assembled using RNA-seq reads from Eco 6 and Eco 215 venom glands whilst Eco VG (4) was assembled using RNA-seq reads from all four venom gland datasets from *Echis coloratus*. Abbreviations: Ema, *Eublepharis macularius* (leopard gecko); Pre, *Python regius* (royal python); Eco, *Echis coloratus* (painted saw-scaled viper); Epy, *Echis pyramidum* (Egyptian saw-scaled viper); Oae, *Opheodrys aestivus* (rough green snake); Pgu, *Pantherophis guttatus* (corn snake).

3.11 Genome-guided assembly

The percentage of RNA sequencing reads mapping to each respective reference genome was variable. An average of 80.97% of reads mapped to the king cobra genome, 71.94% of reads mapped to the *Echis coloratus* genome, and only 48.1% of reads mapped to the Burmese python genome. This is therefore suggestive that RNA-seq reads sequenced from a different species to that of the reference genome (even though the Burmese python and royal python are both members of the genus *Python*) do not map as effectively as when the transcriptome and genome species are the same.

The assembly metrics for each species are also revealing. The *Echis coloratus* assemblies have a mean contig N50 of 1,641bp and a mean maximum contig length of 18,916bp. Royal python assemblies have a mean N50 of 1,434bp but considering individual sample assemblies only, a mean contig N50 of 1,392. Assemblies for the king cobra, despite having considerably less RNA sequencing depth of only 50bp single-end reads have a mean contig N50 of 1,284 and a mean maximum contig length of 15,783. With this in mind, the assemblies for the king cobra data are not too dissimilar to those with a much higher sequencing coverage, and this is more than likely due to the superior completeness of the king cobra genome (more exons will be located on the same scaffold and so transcript reconstruction will be improved) and the overall higher percentage of reads mapping to the reference genome.

3.12 Evaluation of transcriptome assembly methods

It is apparent that for *de novo* assembly Trinity outperforms SOAPdenovo-Trans at assembling both individual and tissue assemblies in terms of contig N50 (Figures 3.9 and 3.10) but not necessarily maximum contig length (Figures 3.11 and 3.12).

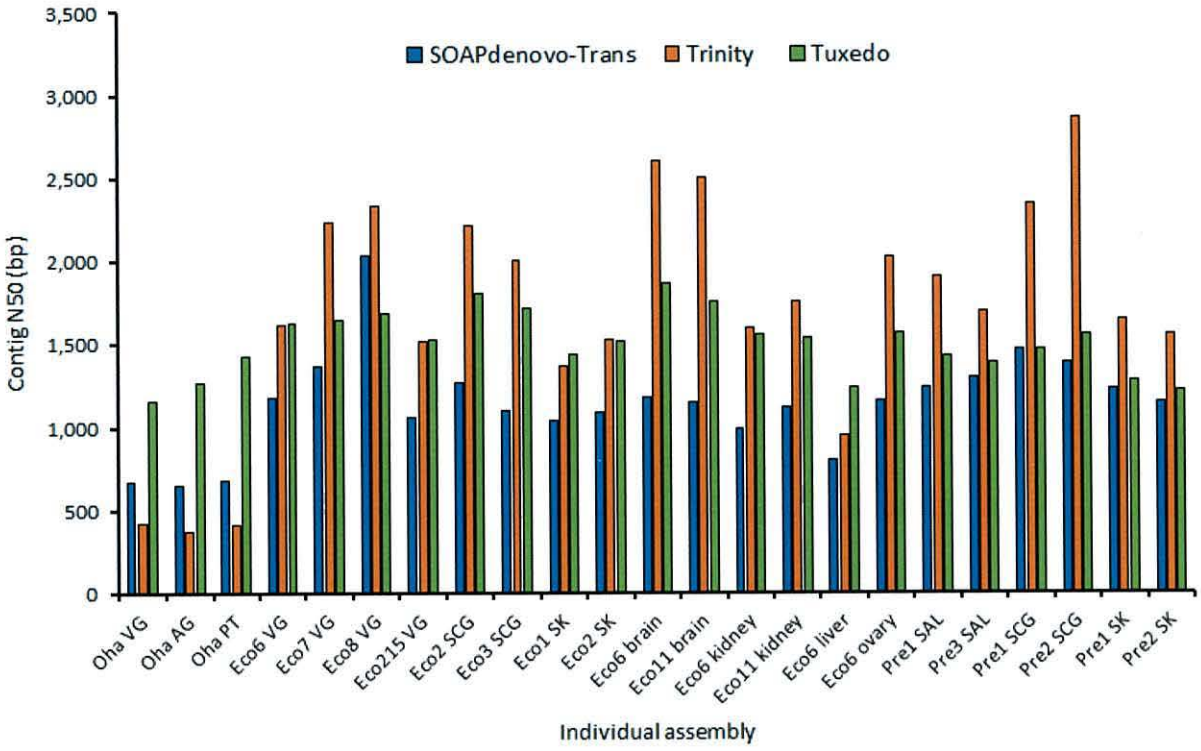


Figure 3.9. Contig N50 values (bp) of individual sample assemblies generated using SOAPdenovo-Trans, Trinity and the Tuxedo suite

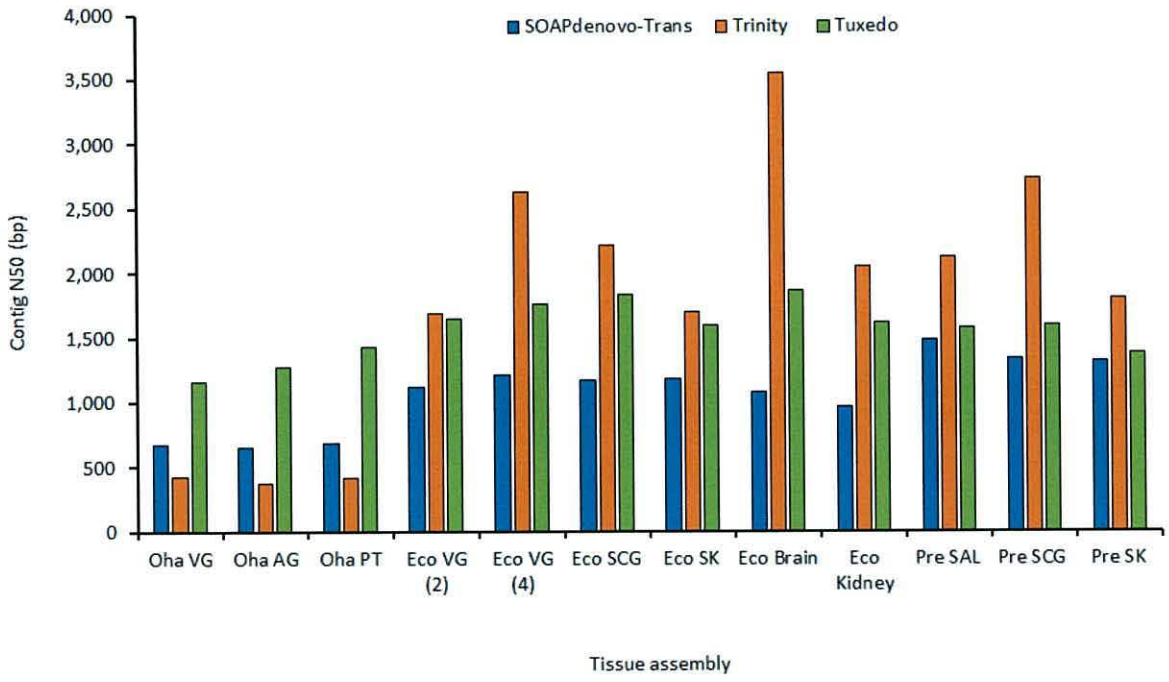


Figure 3.10. Contig N50 values (bp) of tissue assemblies generated using SOAPdenovo-Trans, Trinity and the Tuxedo suite

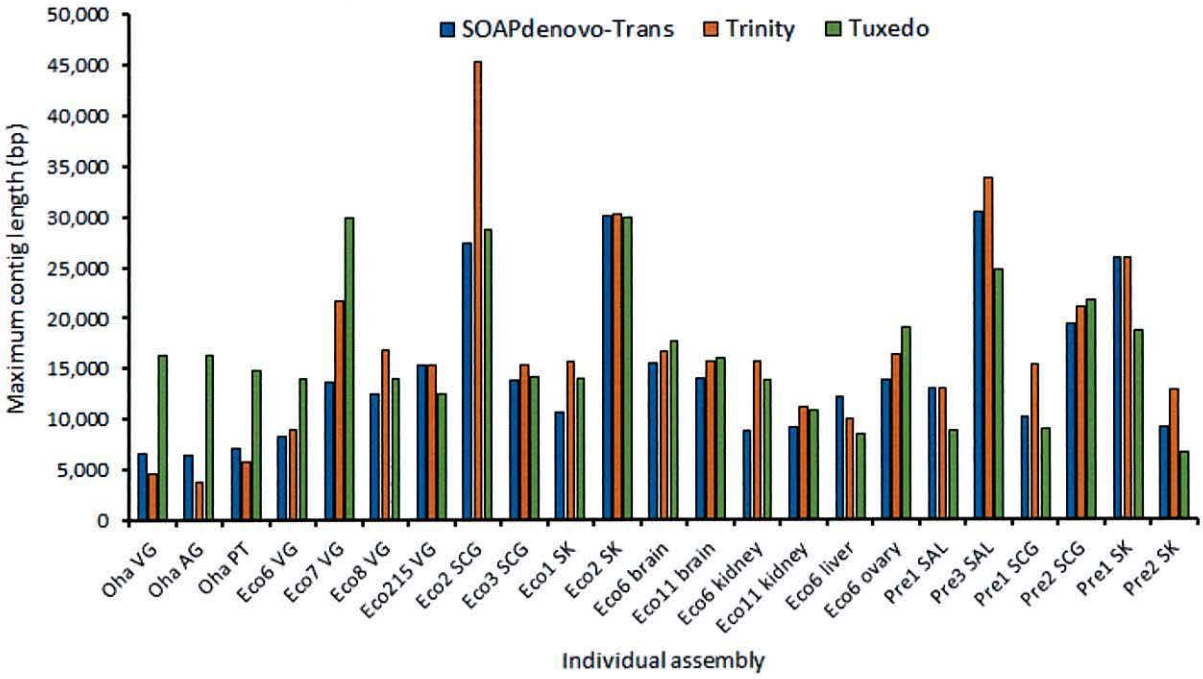


Figure 3.11. Maximum contig length values (bp) of individual sample assemblies generated using SOAPdenovo-Trans, Trinity and the Tuxedo suite

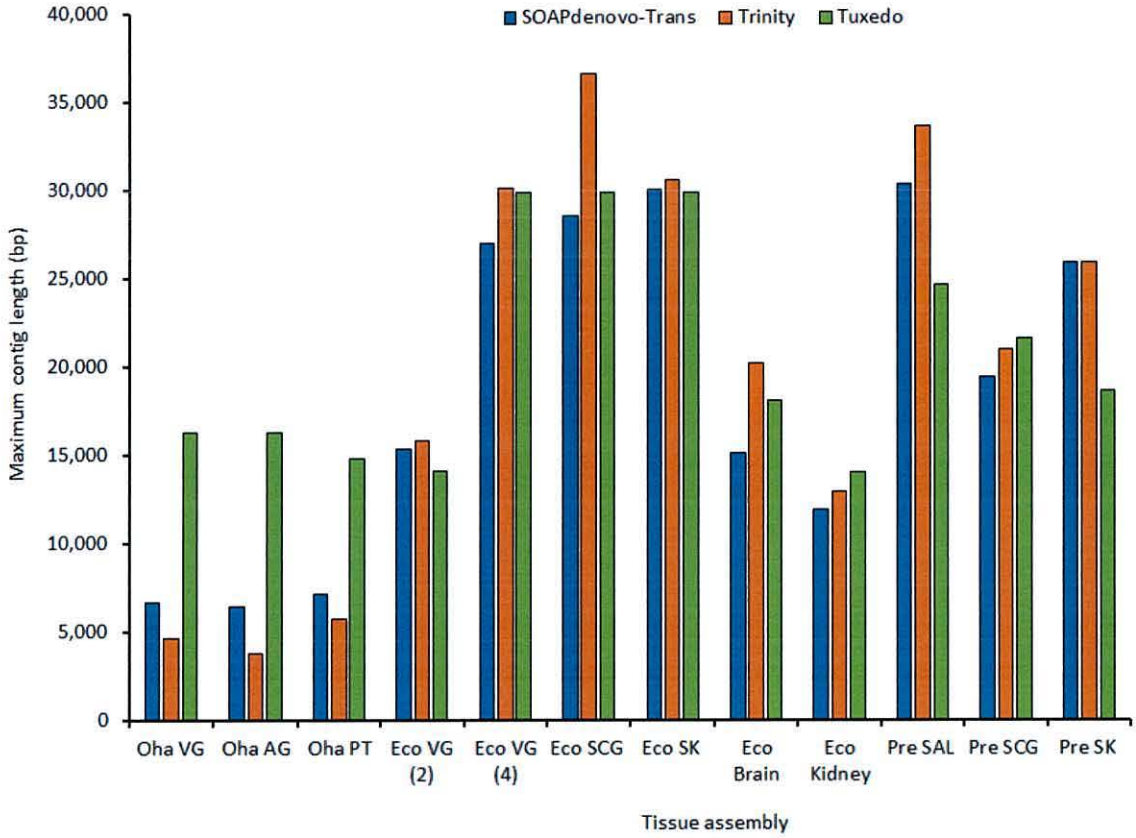


Figure 3.12. Maximum contig length (bp) of tissue assemblies generated using SOAPdenovo-Trans, Trinity and the Tuxedo suite

Overall, Trinity also produces the highest number of contigs ≥300bp in length out of all three assembly methods (Figure 3.13), but appears to be roughly equal to the Tuxedo suite in terms of maximum contig length and contig N50.

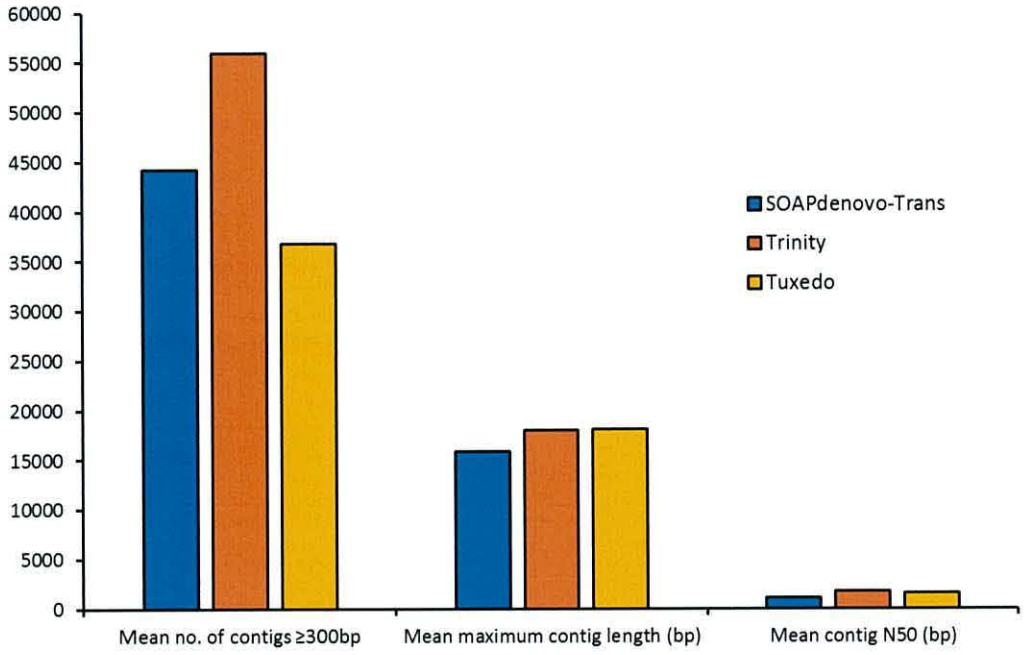


Figure 3.13. Overall comparison between Trinity, SOAPdenovo-Trans and the Tuxedo suite based upon the mean number of contigs >300bp, mean maximum contig length and mean contig N50.

It is obvious from Figures 3.9-3.12 that the Tuxedo suite was vastly superior in assembling the king cobra transcriptomes than either of the *de novo* assembly programs. It is likely that this is due to both Trinity and SOAPdenovo-Trans being unable to construct De Bruijn graphs from such a small amount of short, single-ended sequencing reads, and consequently the majority of contigs constructed will be very short in length. On the other hand, the Tuxedo suite does not rely on a graph-based algorithm, but rather the mapping of reads to positions in the genome, and so will be much more effective in assembling transcripts from the king cobra data.

3.13 Transcript abundance estimation with RSEM

3.13.1 Global assembly metrics

Global transcriptome assemblies comprising of multiple sets of reads were assembled using Trinity in order to allow the comparison of transcript abundance values either between tissues (Tables 3.9 and 3.10) or in the venom gland at different timepoints following milking (Table 3.11).

Table 3.9. Assembly metrics for species-specific global assemblies of salivary/venom gland, scent gland and skin.

Reference assembly	Total number of bases	Number of contigs	Number of contigs >300bp	Contig N50 (bp)	Max contig length (bp)
EcoTissueRef	228,063,624	206,147	149,821	2,445	38,875
PguTissueRef	166,312,211	152,359	112,327	2,315	23,951
OaeTissueRef	210,451,256	204,942	147,597	2,200	52,645
PreTissueRef	301,328,500	219,070	166,578	3,407	34,165
EmaTissueRef	266,096,501	228,645	167,623	2,746	33,255

Table 3.10. Assembly metrics for a global assembly of 7 tissues (venom gland, scent gland, skin, brain, kidney, liver and ovary) from one adult individual specimen of *Echis coloratus* (Eco6)

Reference assembly	Total number of bases	Number of contigs	Number of contigs >300bp	Contig N50 (bp)	Max contig (bp)
Eco6TissuesRSEM	229,385,556	229,535	147,966	3,044	38,449

Table 3.11. Assembly metrics for a global assembly of all four venom gland samples from *Echis coloratus*.

Reference assembly	Total number of bases	Number of contigs	Number of contigs >300bp	Contig N50 (bp)	Max contig (bp)
Eco_7_8_ERR	121,590,983	117,125	81,798	2,623	30,131

3.13.2 qPCR reference genes

To demonstrate the utility of this method, and also to provide additional information to allow the planning of future qPCR experiments to examine the expression of specific genes in an increased number of samples, the transcript abundance of several qPCR reference genes was estimated. This included analysis of transcript expression between 7 different body tissues in *E. coloratus* (for qPCR examining gene expression relative to tissue) (Table 3.12), and also between 4 venom gland samples taken at different time points following milking (for qPCR examining transcript expression at various stages of the venom replenishment cycle following milking) (Table 3.13).

Table 3.12. Transcript abundance estimation values given in FPKM relating to the expression levels of 13 potential reference genes across 7 tissues in *Echis coloratus*.

	Estimated transcript abundance (FPKM)						
Gene	VG	SCG	SK	Brain	Kidney	Liver	Ovary
Alpha tubulin	40.48	4.87	16.58	354.23	5.63	0	3.32
Beta actin	22,110.35	29,101.17	14,124.56	32.43	29.31	7.93	88.13
GAPDH	6,910.7	14,541.36	13,977.9	312.7	952.86	773.16	1,198.18
POLR2A	0.91	1.79	7.15	3.96	1.61	0	3.05
POLR2B	15.47	17.55	25.2	41.48	17.01	4.83	51.0
B2M	503.33	374.73	328.73	139.75	265.26	349.6	368.67
HPRT1	21.12	27.2	26.57	67.28	16.67	0.27	11.24
B glucuronidase	4.94	5.09	4.22	3.09	2.61	1.09	12.86
LDHA	1,919.93	6,338.13	3,631.13	77.93	117.04	643.68	31.76
NONO	70.29	45.53	72.56	81.51	45.87	23.95	108.21
TBP	5.04	3.64	4.27	9.36	7.76	2.03	4.84
TfR1	0.72	0.8	2.45	1.65	1.22	4.84	1.25
G6PD	9.26	5.13	5.95	13.57	2.73	0.05	6.34

Abbreviations

GAPDH, Glyceraldehyde 3-phosphate dehydrogenase; POLR2A, RNA polymerase II polypeptide A; POLR2B, RNA polymerase II polypeptide B; B2M, Beta-2 microglobulin; HPRT1, Hypoxanthine phosphoribosyltransferase 1; LDHA, lactate dehydrogenase A; NONO, non-POU domain containing octamer-binding protein; TBP, TATA-binding protein; TfR1, transferrin receptor 1; G6PD, glucose-6-phosphate dehydrogenase.

Table 3.13. Transcript abundance estimation values given in FPKM relating to the expression levels of 13 potential reference genes across 4 venom gland samples taken at different timepoints following manual venom extraction from *Echis coloratus*.

	Estimated transcript abundance (FPKM)			
Gene	Eco 8	Eco 7	Eco 6	Eco 215
Alpha tubulin	69.76	51.31	24.24	36.83
Beta actin	9,006.95	7,978.58	21,498.12	11,919.12
GAPDH	5,732.58	2,601.72	6,266.49	9,107.47
POLR2A	2.86	0.93	1.53	1.15
POLR2B	7.67	9.01	15.89	23.08
B2M	121.79	157.35	1,040.11	1,702.88
HPRT1	11.85	16.54	23.54	27.77
B glucuronidase	3.95	2.31	5.3	5.47
LDHA	975.49	731.32	1,952.27	1,697.05
NONO	64.41	18.17	74.33	55.93
TBP	2.04	1.46	2.31	1.29
TfR1	1.8	1.85	1.07	0.37
G6PD	5.41	13.51	15.35	14.78

Abbreviations

GAPDH, Glyceraldehyde 3-phosphate dehydrogenase; POLR2A, RNA polymerase II polypeptide A; POLR2B, RNA polymerase II polypeptide B; B2M, Beta-2 microglobulin; HPRT1, Hypoxanthine phosphoribosyltransferase 1; LDHA, lactate dehydrogenase A; NONO, non-POU domain containing octamer-binding protein; TBP, TATA-binding protein; TfR1, transferrin receptor 1; G6PD, glucose-6-phosphate dehydrogenase.

Due to their extremely high expression and high variability both across different tissues and in different venom gland samples taken at varying timepoints following milking, the transcript abundance estimations for GAPDH, β actin and LDHA have been plotted on separate graphs to the remaining candidate reference genes (Figures 3.14 and 3.15).

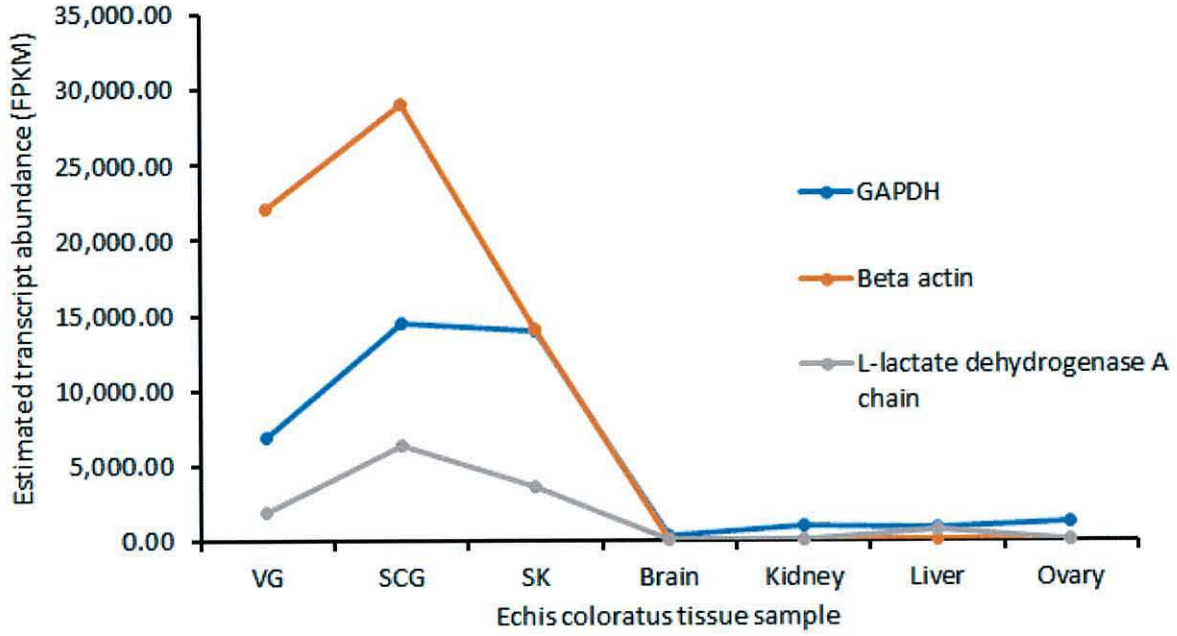


Figure 3.14. Transcript expression levels of GAPDH, β actin and LDHA across 7 *Echis coloratus* tissues.

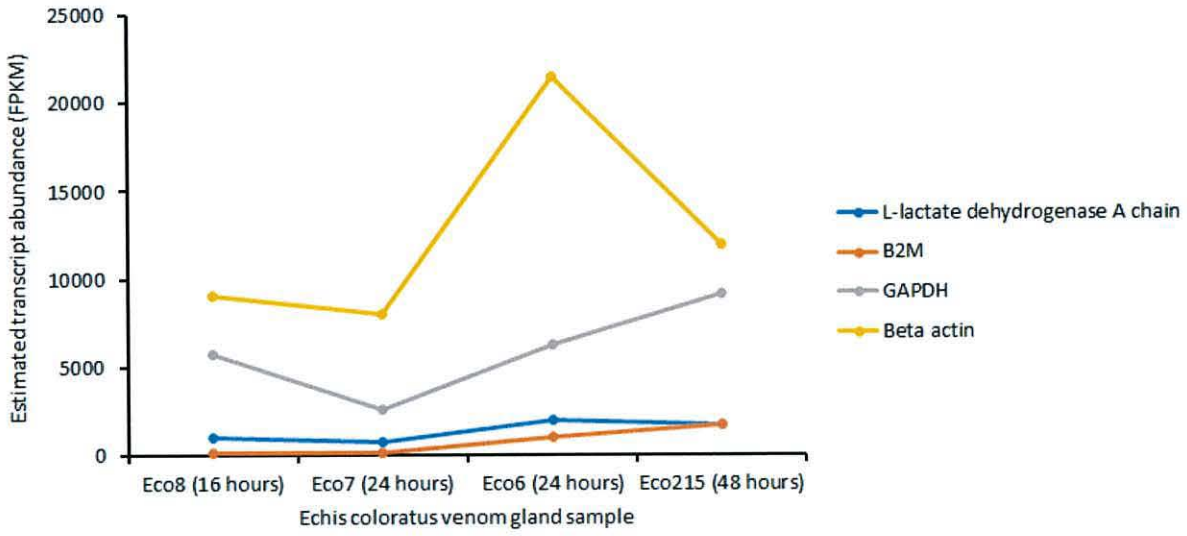


Figure 3.15 Transcript expression levels of GAPDH, β actin and LDHA in the venom gland at different timepoints following milking.

Based on the remaining candidate reference gene expression levels (Figures 3.16 and 3.17), there still appears to be erratic variation across tissues and between venom gland samples at different stages of the venom replenishment cycle. Nevertheless, the genes POLR2A, POLR2B,

B glucuronidase and Transferrin receptor protein 1 stand out as putative candidate reference genes due to showing relatively stable (albeit low) expression levels across all tissues sampled and across all venom gland samples.

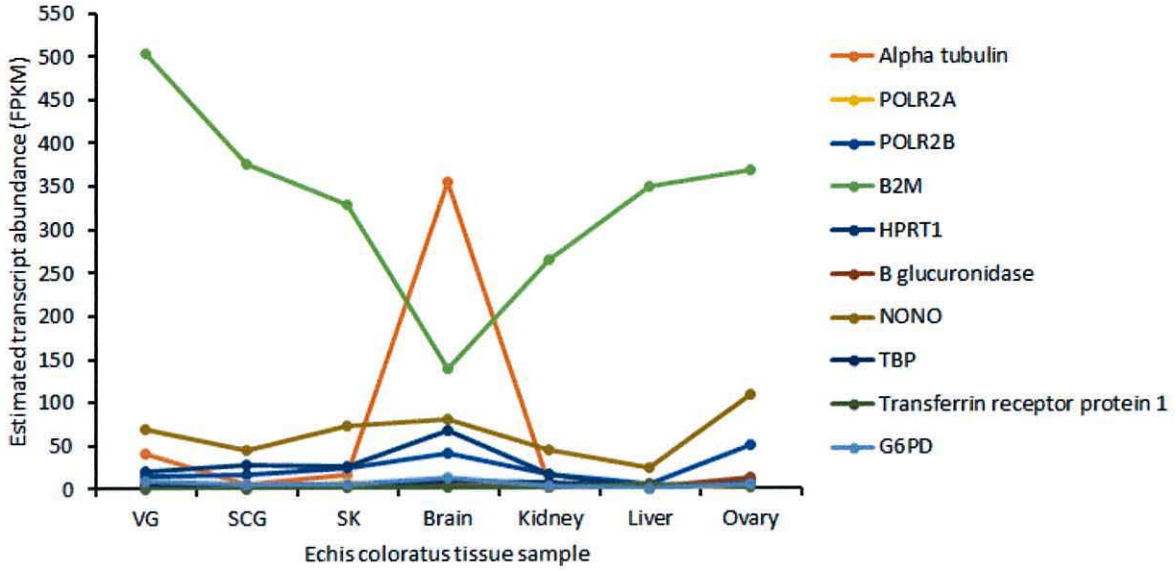


Figure 3.16 Transcript expression levels of putative reference genes across 7 *Echis coloratus* tissues

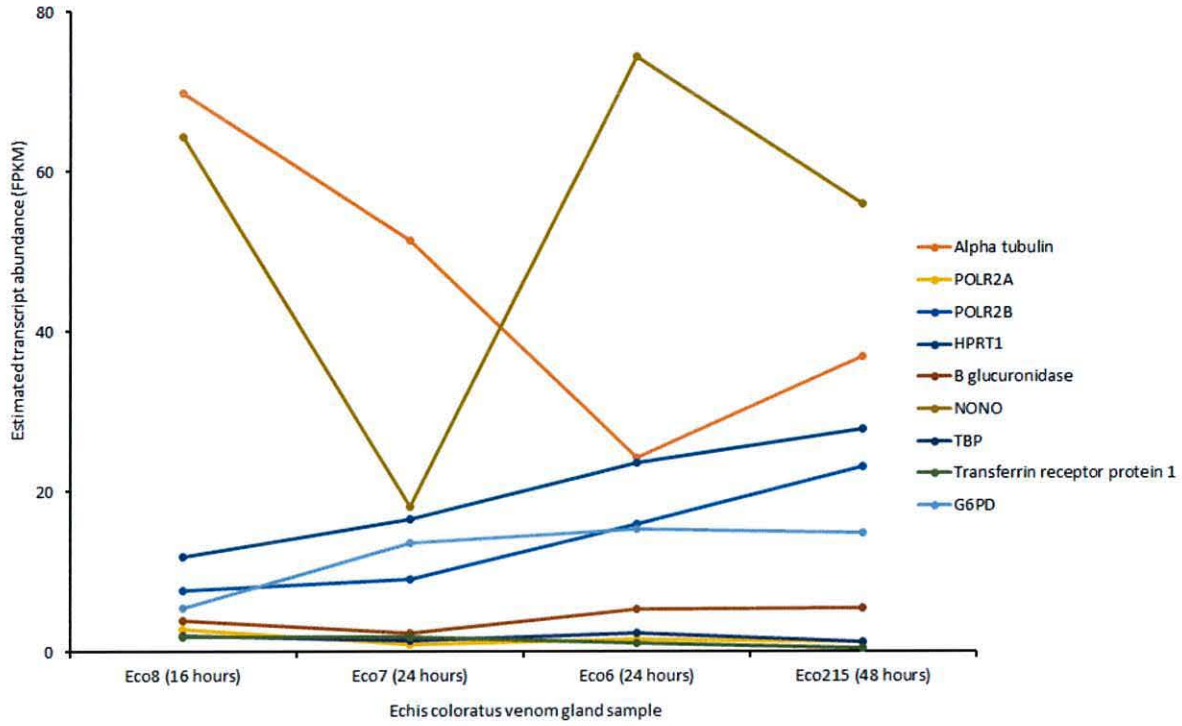


Figure 3.17 Transcript expression levels of putative reference genes in the venom gland at different timepoints following milking.

3.14 Comparison of existing and newly generated *Echis* venom gland transcriptomes

The venom gland transcriptomes of several *Echis* species have previously been characterised by the cloning and sequencing of a small number (~1000) of expressed sequence tags (ESTs) (Wagstaff et al. 2009; Casewell et al. 2009). The venom gland proteome of *E. ocellatus* has also been characterised by reverse-phase HPLC (High Performance Liquid Chromatography), SDS-PAGE, N-terminal sequencing, MALDI-TOF mass spectrometry and CID-MS/MS of tryptic peptides (Wagstaff et al. 2009). The number of putative toxin-encoding EST clusters and identified venom proteins is shown in Table 3.14.

Table 3.14. Numbers of putative venom genes detected in previous EST-based transcriptomic studies of the venom gland and proteomic studies of extracted venom in a range of *Echis* species compared to the results of this study using RNA-seq. Figures for the number of putative toxin-encoding genes in the venom gland trasncriptomes of *E. coloratus*, *E. pyramidum leakyi*, *E. ocellatus* and *E. carinatus sochureki* were obtained from (Casewell et al. 2009; Wagstaff et al. 2009). Similarly, the number of putative toxin encoding peptide sequences from the crude venom of *E. ocellatus* were obtained from (Wagstaff et al. 2009).

	Previously characterised					This study	
Gene	Echis coloratus	Echis pyramidum leakyi	Echis ocellatus	Echis carinatus sochureki	Echis ocellatus proteome	Echis coloratus	Echis pyramidum
SVMP	27	19	20	26	26	13	21
C-type lectin	12	16	6	11	2	8	14
PLA_2 IIA	3	6	2	2	4	3	3
SP	9	3	1	4	1	6	7
LAAO	1	1	2	2	1	2	1
CRISP	1	0	0	2	1	2	1
VEGF	1	1	2	1	0	4	3
NGF	1	1	0	0	0	1	1
LIPA	1	0	0	0	0	2	2
RAP	0	0	2	0	0	1	1
Hyal	0	0	0	0	0	2	4
Kunitz	0	0	0	0	0	3	1
crotamine	0	0	0	0	0	1	0

Abbreviations

SVMP, snake venom metalloproteinase; PLA_2 IIA, phospholipase A_2 type IIA; SP, serine protease; LAAO, L-amino acid oxidase, CRISP, cysteine-rich secretory protein; VEGF, vascular endothelial growth factor; NGF, nerve growth factor; LIPA, lysosomal acid lipase; RAP, renin aspartate protease.

Putative toxin-encoding transcripts were identified in the newly generated venom gland transcriptomes of *E. coloratus* and *E. pyramidum* by local BLAST searches and phylogenetic analysis when transcripts were members of large gene families (Chapter 4 and Appendix 8-13). It is worth noting that these figures may be an overestimation of the number of toxin transcripts in these species as no comparison with a non-venom gland tissue has been made. RNA-seq appears to have detected more lowly-expressed transcripts (such as multiple genes encoding vascular endothelial growth factors) than previous EST analyses, although the numbers of well-known toxin-encoding transcripts such as SVMPs appear to be approximately similar, if not slightly lower. The higher coverage of RNA-seq has also led to the detection of transcript splice variants of several genes including *vascular endothelial growth factor a* and *L-amino acid oxidase b* (see Chapter 4), which were not detected in previous analyses.

3.15 Sub-assemblies of the *Echis coloratus* venom gland transcriptome

In an attempt to determine the minimum required amount of sequencing to fully sequence and assemble the venom gland trancriptome of *Echis coloratus*, sub-sets of RNA-seq reads were extracted and assembled, sequencing metrics for which are displayed in Table 3.15.

Table 3.15. Assembly metrics for sub-assemblies of the venom gland transcriptome of *Echis coloratus*.

Sub-sample	Sample size (million reads)	Total number of contigs	Number of contigs ≥300nt	Total length (nt)	Max. contig size (nt)	Contig N50 (nt)
HEAD	2	24,585	14,744	10,302,850	7,474	808
	4	34,990	22,184	17,605,771	7,860	1,023
	8	45,207	30,121	27,542,537	11,824	1,293
	10	48,349	32,660	31,623,176	11,824	1,420
MIDDLE	2	23,915	14,229	10,036,594	7,840	837
	4	34,383	21,736	17,282,856	8,970	1,027
	8	44,759	29,946	27,116,697	11,738	1,279
	10	47,832	32,451	30,985,872	11,752	1,387
TAIL	2	24,170	14,513	10,059,952	8,547	810
	4	34,735	21,994	17,315,514	8,165	1,004
	8	44,956	29,988	27,283,356	11,803	1,284
	10	48,116	32,535	31,022,314	11,805	1,382

Sub-asemblies were then searched for previously characterised putative toxin sequences from the venom gland transcriptome of *E. coloratus*. The majority of transcripts encoding putative toxin genes (51 out of 64) appear to be present in venom gland transcriptome assemblies generated from only 2 million paired-end reads (here presence is defined as the transcript being found in all three Head/Middle/Tail sub-assemblies) (Table 3.16).

Table 3.16. Presence/absence of putative toxin transcripts in sub-assemblies of the venom gland transcriptome of *Echis coloratus*. H, head; M, middle; T, tail. Detected transcripts are shaded blue, transcripts not found are shaded grey.

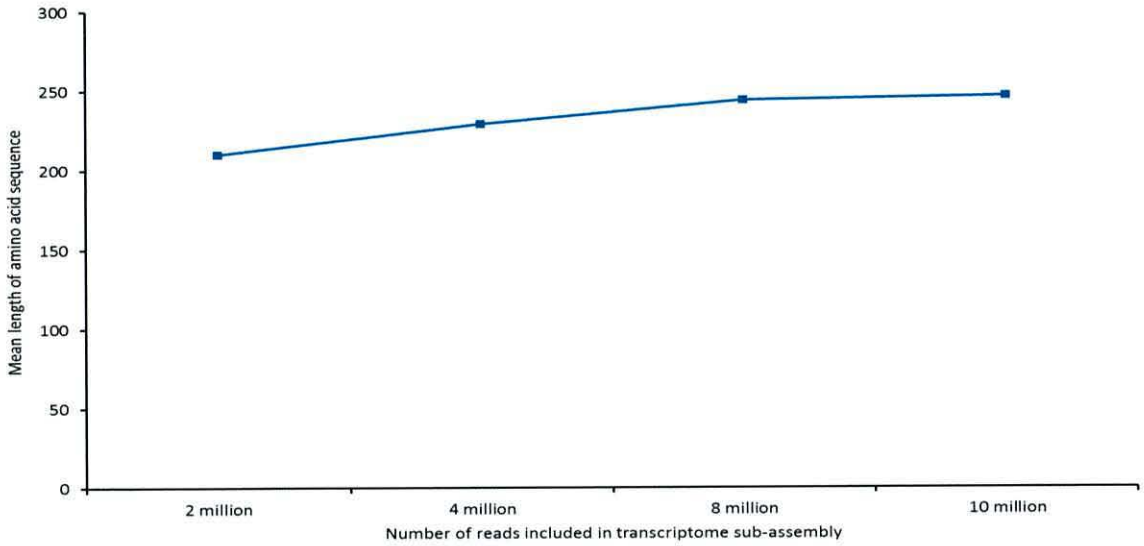
	Sub-sample size and position											
	2 million reads			4 million reads			8 million reads			10 million reads		
Gene	H	M	T	H	M	T	H	M	T	H	M	T
3ftx-a	+	-	+	+	+	+	+	+	+	+	+	+
3ftx-b	+	+	+	+	+	+	+	+	+	+	+	+
ache - transcript 1	-	+	-	+	+	+	+	+	+	+	+	+
complement c3	+	+	+	+	+	+	+	+	+	+	+	+
crisp-b	+	+	+	+	+	+	+	+	+	+	+	+
crotamine-like	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin a	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin b	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin c	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin d	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin e	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin f	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin g	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin h	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin i	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin j	+	+	+	+	+	+	+	+	+	+	+	+
c-type lectin k	+	+	+	+	+	+	+	+	+	+	+	+
cystatin e/m	+	+	+	+	+	+	+	+	+	+	+	+
cystatin f	-	-	-	-	-	-	-	-	-	-	-	-
dpp 3	+	+	+	+	+	+	+	+	+	+	+	+
dpp 4	+	+	+	+	+	+	+	+	+	+	+	+
esp-e1	+	+	+	+	+	+	+	+	+	+	+	+
ficolin	+	+	+	+	+	+	+	+	+	+	+	+

kallikrein	+	+	+	+	+	+	+	+	+	+	+	+
kunitz 1	+	+	+	+	+	+	+	+	+	+	+	+
kunitz 2	+	+	+	+	+	+	+	+	+	+	+	+
laao-a	+	+	+	+	+	+	+	+	+	+	+	+
laao-b1	+	+	+	+	+	+	+	+	+	+	+	+
laao-b2	+	+	+	+	+	+	+	+	+	+	+	+
lipa-a	+	+	+	+	+	+	+	+	+	+	+	+
lipa-b	-	-	-	+	+	+	+	+	+	+	+	+
ngf	+	+	+	+	+	+	+	+	+	+	+	+
PLA_2 IIA-c	+	+	+	+	+	+	+	+	+	+	+	+
PLA_2 IIA-d	+	+	+	+	+	+	+	+	+	+	+	+
PLA_2 IIA-e	+	+	+	+	+	+	+	+	+	+	+	+
PLA_2 IIE	-	-	-	-	+	-	-	+	+	-	-	+
plb	+	+	+	+	+	+	+	+	+	+	+	+
renin	+	+	+	+	+	+	+	+	+	+	+	+
serine protease a	+	+	+	+	+	+	+	+	+	+	+	+
serine protease b	+	+	+	+	+	+	+	+	+	+	+	+
serine protease c	+	+	+	+	+	+	+	+	+	+	+	+
serine protease d	+	+	+	+	+	+	+	+	+	+	+	+
serine protease e	+	+	+	+	+	+	+	+	+	+	+	+
serine protease f	+	+	+	+	+	+	+	+	+	+	+	+
svmp-a	+	+	+	+	+	+	+	+	+	+	+	+
svmp-b	+	+	+	+	+	+	+	+	+	+	+	+
svmp-c	-	-	+	-	+	+	+	+	+	+	+	+
svmp-d	-	-	-	-	+	+	+	+	+	+	+	+
svmp-e	+	+	+	+	+	+	+	+	+	+	+	+
svmp-f	+	+	+	+	+	+	+	+	+	+	+	+
svmp-g	+	+	+	+	+	+	+	+	+	+	+	+
svmp-i	+	+	+	+	+	+	+	+	+	+	+	+
svmp-j	+	+	+	+	+	+	+	+	+	+	+	+
svmp-k	+	+	+	+	+	+	+	+	+	+	+	+

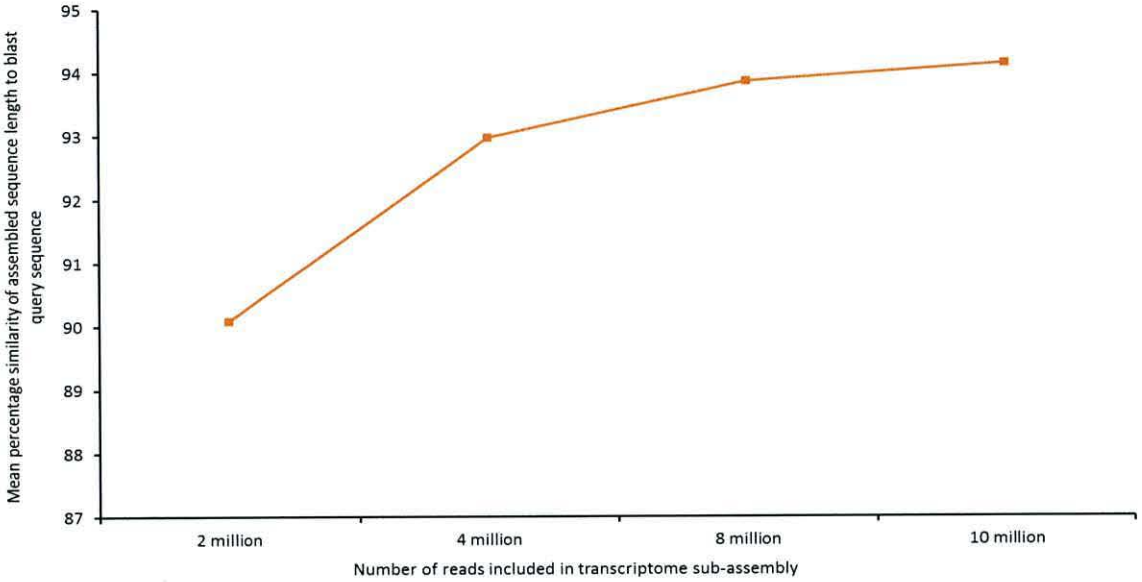
svmp-m	+	-	+	+	+	+	+	+	+	+	+	+
svmp-n	+	+	+	+	+	+	+	+	+	+	+	+
svmp-p	+	-	+	+	+	+	+	+	+	+	+	+
svmp-q	+	+	+	-	-	+	+	+	+	+	+	+
svmp-t	+	+	+	+	+	+	+	+	+	+	+	+
vegf-a	-	-	-	+	-	+	+	+	+	+	+	+
vegf-b	-	-	-	-	-	-	-	-	-	-	-	-
vegf-c	-	-	-	-	-	-	+	-	+	+	+	+
vegf-f	+	+	+	+	+	+	+	+	+	+	+	+
waprin	+	-	+	+	+	+	+	+	+	+	+	+

As the number of reads used for assembly increases the mean length of the amino acid sequence encoded by the assembled transcript also increases, although there is only a 36 amino acid increase between 2 million and 10 million reads (Figure 3.18 panel A). However, the number of contigs ≥300bp roughly doubles (Table 3.15), meaning considerably fewer contigs which are likely to be unplaced paired reads are present in the transcriptome assembly. To gain insight into how this increase in length relates to the quality of the assembled toxin transcript sequences, the percentage of the query sequence covered by the newly assembled sequence was calculated. Again there is only a minor improvement of 4% following an increase from 2 million reads to 10 million (Figure 3.18 panel B). The mean percentage similarity between assembled sequence and query sequence appears to be more variable across the sub-assemblies, with no apparent consistent improvement as the number of reads increases (Figure 3.18 panel C). As the query sequences used for local BLAST searches were obtained from an assembly of multiple *E. coloratus* venom gland datasets (Eco6 and Eco215) in order to represent an over-abundance of sequencing, and the sub-assemblies were assembled from Eco7 venom gland reads (due to the reasons mention in the methods section), it should be expected that not all blast alignments will have a 100% match between query and subject due to minor variation between individuals. However, a lower % identity would indicate that either sequencing errors were incorporated into the assembly or there has been a misassembly, both likely due to a reduced depth of sequencing coverage.

A.


300
250
200
150
100
50
0
Mean length of amino acid sequence
2 million
4 million
8 million
10 million
Number of reads included in transcriptome sub-assembly

B.


95
94
93
92
91
90
89
88
87
Mean percentage similarity of assembled sequence length to blast query sequence
2 million
4 million
8 million
10 million
Number of reads included in transcriptome sub-assembly

C.

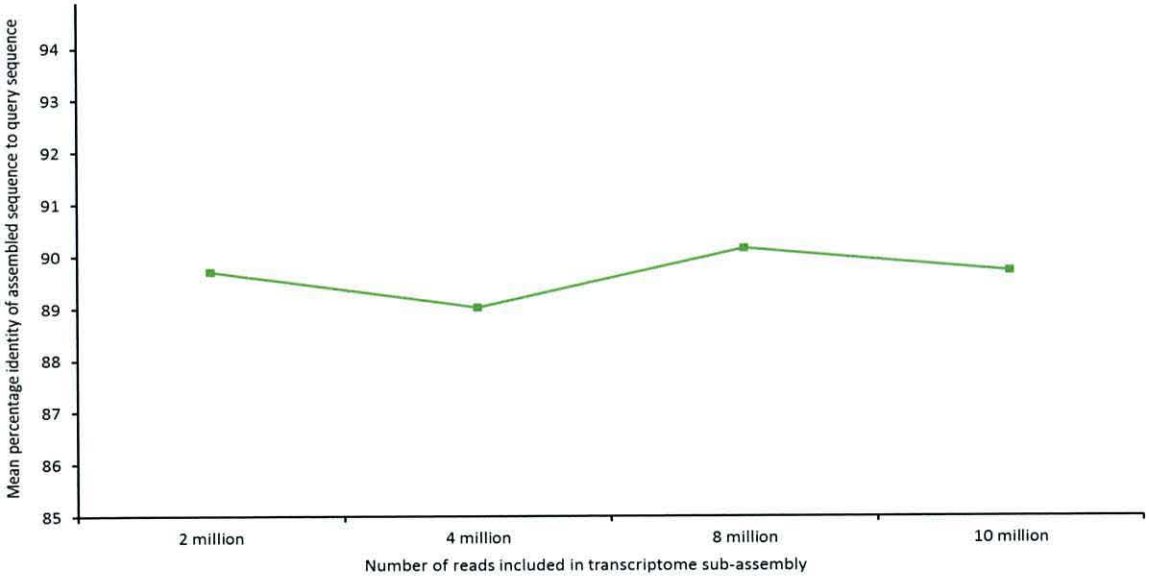

94
93
92
91
90
89
88
87
86
85
Mean percentage identity of assembled sequence to query sequence
2 million
4 million
8 million
10 million
Number of reads included in transcriptome sub-assembly

Figure 3.18. Analysis of sequence assembly quality based on local blast surveys using previously characterised amino acid sequences from *Echis coloratus* venom gland. A. mean length of amino acid sequence matches in sub-assemblies, B. mean percentage length of query sequence covered by assembled sequence and C. mean percentage similarity of assembled sequence to query sequence in sub-assemblies.

3.16 Discussion

Sequencing transcriptomes using next-generation sequencing technology can shed light on both the genes being actively expressed in a particular cell or tissue, and also the levels at which these transcripts are expressed. This is particularly useful for identifying if transcripts are expressed in either multiple or specific tissues (Chapters 4 and 5), or if genes show an elevated expression level in particular tissues (Chapter 4). Alternatively, it allows the investigation of any differences in gene expression in a particular tissue when exposed to varying conditions or after particular treatments, such as gene expression in the venom gland at different times following milking (Chapter 6).

Whilst the utility of this technique cannot be overstated, there is currently no gold standard in how RNA-seq data is assembled, or how much sequencing is required to fully characterise a tissue transcriptome. With the diverse array of assembly software available, it is also hard to determine which is the best assembly program for a particular application. I chose two freely available *de novo* assembly programs, both of which utilise De Bruijn graph-based algorithms to construct contigs from short read data. Using several draft whole genome sequences I also carried out genome-guided assembly to see how this compared to *de novo* assembly. It is worth noting that this analysis is also informative of the quality of the reference genome sequence used as well as the resulting transcriptome assemblies.

For *de novo* sequencing using paired-end data, Trinity outperformed SOAPdenovo-Trans, and the large amount of possible downstream analyses incorporated into the former program make it an appealing tool to use. One very apparent advantage to SOAPdenovo-Trans was its speed, completing most transcriptome assemblies in under 30 minutes, whereas Trinity on the whole took several hours. However, neither *de novo* program managed to assemble the king cobra transcriptomes well, most likely due to the low sequencing depth of this data and being unable to construct De Bruijn graphs from very short singe-end reads.

The Tuxedo suite appears to be sensitive to the species used, as mapping royal python reads to the Burmese python genome appeared to yield a considerably lower percentage of mapped

reads compared to the other two species, despite them belonging to the same genus. With some modification and optimisation of mapping parameters this problem may be abated. As this method is not reliant on graph-based algorithms, and the king cobra genome assembly has large scaffolds, this method was superior for assembling the king cobra transcriptomes compared to either *de novo* method. It is possible that an improved *E. coloratus* genome assembly could lead to an improvement in transcriptome assemblies constructed using this method.

The predicted number of venom toxin sequences belonging to well-known toxin gene families appears to be broadly similar across EST and RNA-seq based analyses. The newly generated transcriptomes did however contain contigs encoding lowly expressed transcripts not detected by previous analyses, and also contained sequences encoding splice variants of several genes which were absent from the EST datasets. It is worth noting that due to the low sequencing coverage of previous transcriptomic analyses (~1000 ESTs were sequenced) it is possible that there has been issues during the clustering of EST sequences and that sequencing errors have been incorporated due to poor resolution, leading to an overestimation of the number of toxin transcripts. Coupled with the ability to estimate the abundance of toxin transcripts within a sample, and the ever decreasing cost of RNA-seq library preparation and sequencing, RNA-seq stands out as an ideal method for snake venom transcriptomic studies. However the use of traditional, but ultimately limited, EST sequencing appears to still be in use (Casewell et al. 2014).

Transcript abundance estimation was used to identify putative candidate reference genes for future qPCR analyses, to demonstrate the utility of this method in determining the expression level of transcripts across multiple samples. It was found that genes often stated as unsuitable references for qPCR (GAPDH and β actin) showed highly variable expression across samples, and so should not be used in qPCR experiments. The genes POLR2A, POLR2B, B glucuronidase and Transferrin receptor protein 1 showed relatively stable expression across multiple tissues and within the same tissue following different periods of time post-milking, and so are potentially reliable reference genes for qPCR analyses in *Echis coloratus*.

8 million reads appears to be sufficient sequencing depth to capture all putative toxin-encoding transcripts to a suitable assembly quality. The Illumina HiSeq2500 sequencing platform can currently produce 300-400 million 100nt paired-end reads in "high output" mode, or 200-300 million 150nt paired-end reads in "rapid run" mode (http://systems.illumina.com/systems/hiseq_2500_1500/performance_specifications.ilmn). With this in mind, and 8 million paired-end reads assumed to be the minimum sequencing depth required to fully capture all putative toxin transcripts, it is possible to sequence ~40 venom

gland libraries on one sequencing lane of the Illumina HiSeq2500 (in "high output" mode). For the four individual venom gland samples sequenced in this study it is now possible to suggest a sequencing coverage level for each of them. The Eco 7 venom gland sample is the highest coverage transcriptome with 44,678,609 paired-end reads and an estimated coverage of 5.6x. Eco 8, also sequenced on the Illumina HiSeq2500, has an estimated coverage of 4.8x. The venom gland transcriptomes of Eco6 and Eco215 sequenced on the Illumina HiSeq2000 are considerably lower coverage, with an estimated 1.7x and 1.6x coverage respectively. Assuming that this is also applicable to the closely related *E. pyramidum*, the venom gland of this species has been sequenced to a coverage of 7.2x.

The assembled sequences produced from this depth of sequencing will be suitable for phylogenetic analysis and gene discovery purposes. It is worth noting that for transcripts which belong to gene families whose members have a high degree of sequence similarity (e.g. SVMPs, 3 finger toxins), an increased sequencing depth may be necessary to avoid the occurrence of chimeric assembled transcripts. Alternatively, the long reads offered by 3rd generation sequencing platforms such as Pacific Biosciences and Oxford nanopore may be the way forward in assembling highly similar transcripts reliably.

Downstream analyses such as transcript abundance estimation may also require an increased sequencing depth in order to determine the expression level of transcripts accurately and avoid false-negative results (i.e. only highly expressed transcripts have been sequenced sufficiently and lower expressed transcripts are considered to be not expressed). The occurrence of incomplete or fragmentary transcripts in a global assembly may also result in an under-representation in the number of reads mapped to the transcript during transcript abundance estimation, and the transcript would therefore be considered to be not expressed. However, the reduced depth of sequencing required does enable an increase in the number of libraries sequenced, meaning more biological replicates are possible which are particularly important in the identification of differentially expressed transcripts (Sims et al. 2014). An alternative solution may be to use RNA-seq to characterise a transcriptome and then use this to design primers for qPCR to analyse the expression of specific genes or transcripts across a wider or increased range of samples.

RNA-seq presents itself as an extremely versatile technique to conduct transcriptomic analyses, both in characterising the full transcriptome with enough resolution to detect alternative splice variants, and also in harnessing the inherent quantitative nature of RNA-seq data to analyse transcript expression levels. Multiple bioinformatics programs are freely available to assemble and analyse the millions of reads produced by RNA sequencing. These bioinformatics processes

are faced with several challenges based on the dynamic variation of gene expression across different tissue samples. Specifically with venom gene transcripts belonging to the same gene family, there is a tendency for sequences to be fused together to form chimeras, resulting in either partial or incomplete sequences. A possible solution to this problem in future would be to utilise the long read sequencing offered by new technologies such as Pacific Biosciences SMRT sequencing and Oxford nanopore sequencing, whilst also carrying out the high coverage short read sequencing offered by Illumina and other sequencing platforms for quantitative analysis. Nevertheless, RNA-seq presents itself as an ideal replacement for the traditionally used EST-based sequencing in venom transcriptomics and in transcriptomic analyses in general.

Chapter 4

Testing the Toxicofera hypothesis

The identification of apparently conserved gene complements in the venom and salivary glands of a diverse set of reptiles led to the development of the Toxicofera hypothesis – the idea that there was a single, early evolution of the venom system in reptiles. However, this hypothesis is based largely on relatively small scale EST-based studies of only venom or salivary glands and toxic effects have been assigned to only some putative Toxicoferan toxins in some species. The distribution of these putative venom toxin transcripts was examined in order to assess to what extent this apparent conservation of gene complements may reflect a bias in previous sampling efforts. This represents the first large-scale test of the Toxicofera hypothesis, and it was found to be unsupported. The majority of genes used to support the establishment and expansion of the Toxicofera are in fact expressed in multiple body tissues and most likely represent general maintenance or "housekeeping" genes. The often claimed conservation and homology of genes across the Toxicofera therefore reflects an artefact of incomplete tissue sampling. In other cases, the identification of a non-toxic paralog of a gene encoding a true venom toxin has led to confusion about the phylogenetic distribution of that venom component. Venom has evolved multiple times in reptiles. In addition, the misunderstanding regarding what constitutes a toxic venom component, together with the misidentification of genes and the classification of identical or near-identical sequences as distinct genes has led to an overestimation of the complexity of reptile venoms in general, and snake venom in particular. These findings have implications for our understanding of (and development of treatments to counter) the molecules responsible for the physiological consequences of snakebite.

4.1 The Toxicofera

Snake venom is frequently cited as being highly complex or diverse (Li et al. 2005b; Wagstaff et al. 2006; Kini and Doley 2010) and a large number of venom toxin genes and gene families have been identified, predominantly from EST-based studies of gene expression during the re-synthesis of venom in the venom glands following manually-induced emptying ("milking") of extracted venom (Pahari et al. 2007; Casewell et al. 2009; Siang et al. 2010; Rokyta et al. 2011; Rokyta et al. 2012).

The apparent widespread distribution of genes known to encode venom toxins in snakes in the salivary glands of a diverse set of reptiles, (including both those that had previously been suggested to have secondarily lost venom in favour of constriction or other predation techniques, and those that had previously been considered to have never been venomous), led to the development of the Toxicofera hypothesis – the single, early evolution of venom in reptiles (Vidal and Hedges 2005; Fry et al. 2006; Fry et al. 2009a; Fry et al. 2012b) (Figure 4.1) (see Chapter 1 for background detail on the Toxicofera hypothesis). Analysis of a wide range of reptiles, including charismatic megafauna such as the Komodo dragon, *Varanus komodoensis* (Fry et al. 2009c), has shown that the ancestral Toxicoferan venom system comprises at least 16 genes, with additional gene families subsequently recruited in different lineages (Fry et al. 2009a; Fry et al. 2012b; Fry et al. 2013). Although toxic effects have been putatively assigned to some Toxicoferan venom proteins in certain species, the problem remains that their identification as venom components is based largely on their expression in the venom gland during venom synthesis and their apparent relatedness to other, known toxins in phylogenetic trees. It has long been known that all tissues express a basic set of "housekeeping" or maintenance genes (Butte et al. 2002) and it is therefore not surprising that similar genes might be found to be expressed in similar tissues in different species of reptiles, and that these genes might group together in phylogenetic trees. However, the identification of transcripts encoding putative venom toxins in other body tissues would cast doubt on the classification of these Toxicoferan toxins as venom components, as it is unlikely that the same gene could fulfil toxic and non-toxic (pleiotropic) roles without evidence for alternative splicing to produce a toxic variant (as has been suggested for *acetylcholinesterase* in the banded krait, *Bungarus fasciatus* (Vonk et al. 2011; Casewell et al. 2013)) or increased expression levels in the venom gland (where toxicity might be dosage dependent).

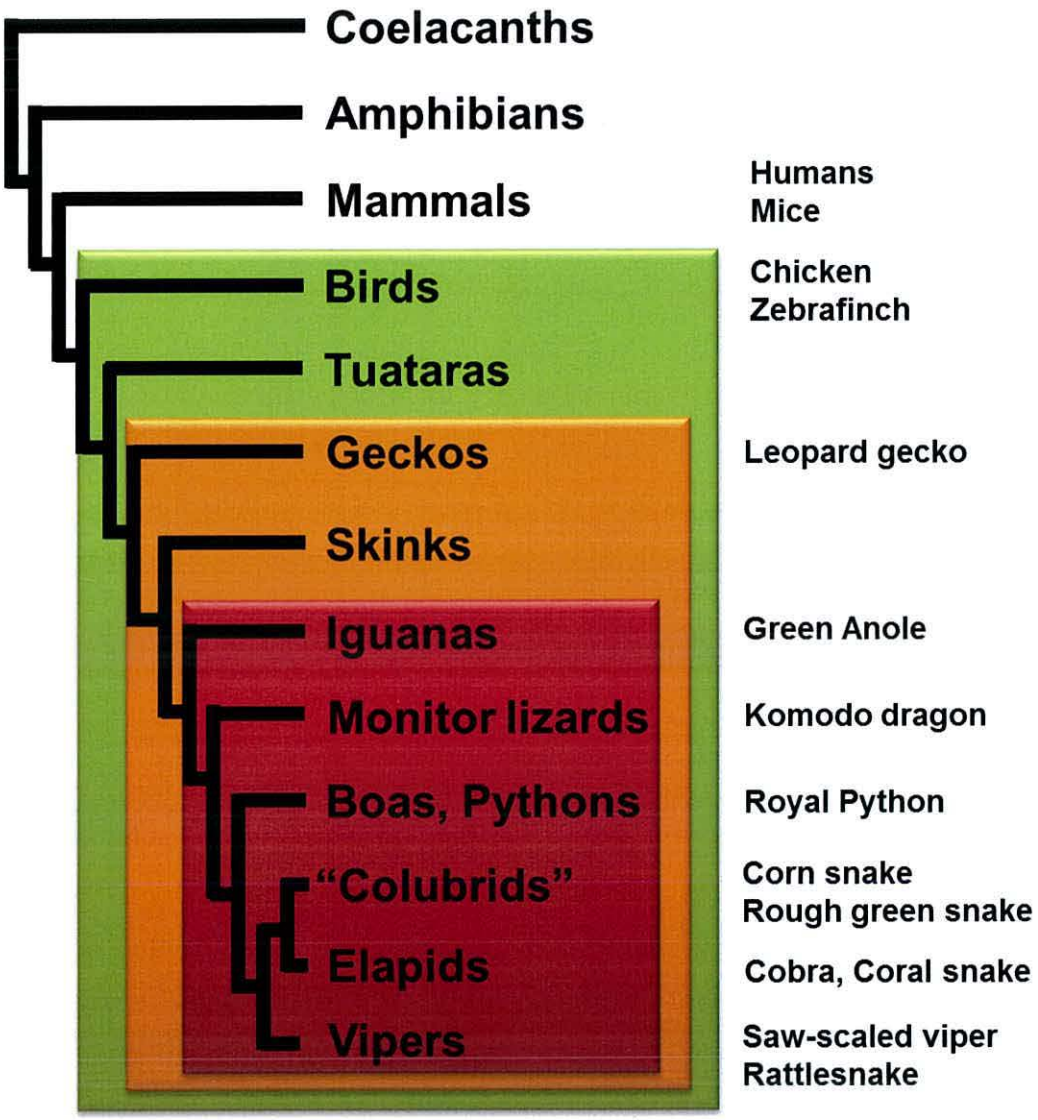


Figure 4.1. Relationships of key vertebrate lineages and the placement of study species. A monophyletic clade of reptiles (which includes birds) is shaded green, the order squamata is shaded orange and the Toxicofera clade (Fry et al. 2013) is shaded red. Modified taxon names are used for simplicity and due to the lack of taxonomic resolution within the Colubridae the term colubrids is placed in inverted commas.

In order to address some of these issues and to test the robustness of the Toxicofera hypothesis, a comparative transcriptomic survey of the venom or salivary glands, skin and cloacal scent glands of five species of reptile was carried out. Unlike the pancreas and other parts of the digestive system (Strydom 1973; Kochva 1987), these latter tissues (which include a secretory glandular tissue (the scent gland) and a relatively inert, non-secretory tissue (skin)) have not

previously been suggested to be the source of duplicated venom toxin genes and therefore only ubiquitous maintenance or "housekeeping" genes should be found to be commonly expressed across these tissues. Here the general term 'salivary gland' is used for simplicity to encompass the oral glands of the leopard gecko and rictal glands and Duvernoy's gland of the royal python, corn snake and rough green snake and no homology to mammalian salivary glands is implied. Study species included the venomous painted saw-scaled viper (*Echis coloratus*); the non-venomous corn snake (*Pantherophis guttatus*) and rough green snake (*Opheodrys aestivus*) and a member of one of the more basal extant snake lineages, the royal python (*Python regius*). As members of the Toxicofera *sensu* Fry et al. (Fry et al. 2013) it should be expected that all of the basic Toxicoferan venom genes are expressed in the venom or salivary glands of all of these species. In addition, corresponding data for the leopard gecko (*Eublepharis macularius*), a member of one of the most basal lineages of squamate reptiles that lies outside of the proposed Toxicofera clade (Figure 4.1) was generated. As an outlier of the Toxicofera clade, none of the basic Toxicoferan genes should be found to be expressed in the salivary gland of the leopard gecko. Available transcriptomes or RNA-Seq data for corn snake vomeronasal organ (Brykczynska et al. 2013) and brain (Tzika et al. 2011), garter snake (*Thamnophis elegans*) liver (Schwartz and Bronikowski 2013) and pooled tissues (brain, gonads, heart, kidney, liver, spleen and blood of males and females (Schwartz et al. 2010)), eastern diamondback rattlesnake (*Crotalus adamanteus*) and eastern coral snake (*Micrurus fulvius*) venom glands (Rokyta et al. 2011; Rokyta et al. 2012; Margres et al. 2013), 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), Burmese python (*Python molurus*) pooled liver and heart (Castoe et al. 2011), green anole (*Anolis carolinensis*) pooled tissue (liver, tongue, gallbladder, spleen, heart, kidney and lung), testis and ovary (Eckalbar et al. 2013) and bearded dragon (*Pogona vitticeps*), Nile crocodile (*Crocodylus niloticus*) and chicken (*Gallus gallus*) brains (Tzika et al. 2011), as well as whole genome sequences for the Burmese python and king cobra (Castoe et al. 2013; Vonk et al. 2013) were also included in analyses.

Assembled transcriptomes were searched for genes previously suggested to be venom toxins in *Echis coloratus* and related species (Wagstaff and Harrison 2006; Casewell et al. 2009; Wagstaff et al. 2009) as well as those that have been used to support the Toxicofera hypothesis, namely *acetylcholinesterase, AVIT peptide* (Fry 2005; Fry et al. 2009a; Vonk et al. 2011; Fry et al. 2012b; Casewell et al. 2013), *complement c3/cobra venom factor, epididymal secretory protein* (Alper and Balavitch 1976; Fry et al. 2012b), *c-type lectins* (Morita 2005; Ogawa et al.

2005), *cysteine-rich secretory protein (crisp)* (Yamazaki et al. 2003a; Yamazaki and Morita 2004), *crotamine* (Rádis-Baptista et al. 2003; Oguiura et al. 2005), *cystatin* (Ritonja et al. 1987; Richards et al. 2011), *dipeptidylpeptidase, lysosomal acid lipase, renin aspartate protease* (Wagstaff and Harrison 2006; Aird 2008; Casewell et al. 2009; Fry et al. 2012b), *hyaluronidase* (Tu and Hendon 1983; Harrison et al. 2007), *kallikrein* (Komori et al. 1988; Komori and Nikai 1998), *kunitz* (Župunski et al. 2003), *l-amino-acid oxidase* (Suhr and Kim 1996; Du and Clemetson 2002), *nerve growth factor* (Angeletti 1970; Kostiza and Meier 1996), *phospholipase A_2* (Lynch 2007), *phospholipase b* (Bernheimer et al. 1987; Chatrath et al. 2011; Rokyta et al. 2011), *ribonuclease* (Aird 2005), *serine protease* (Pirkle 1998; Serrano and Maroun 2005), *snake venom metalloproteinase* (Bjarnason and Fox 1994; Jia et al. 1996), *vascular endothelial growth factor* (*vegf*) (Junqueira de Azevedo et al. 2001; Yamazaki et al. 2003b; Fry 2005; Fry et al. 2006), *veficolin* (OmPraba et al. 2010), *vespryn, waprin* (Torres et al. 2003; Pung et al. 2006; Nair et al. 2007; Fry et al. 2012b) and *3-finger toxins* (Fry et al. 2003a). The abundance of expressed transcripts was also calculated to enable the identification of any instances of pleiotropy (a gene fulfilling a toxic and non-toxic role simultaneously) which would be indicated by a consistently elevated expression level in the venom or salivary gland compared to other tissues.

4.2 Methods

RNA-seq and transcriptome assembly methods are described in detail in chapter 3. Briefly, total RNA was extracted from four venom glands taken from four individual specimens of adult Saw-scaled vipers (*Echis coloratus*) at different time points following venom extraction in order to capture the full diversity of venom genes (16, 24 and 48 hours post-milking). Additionally, total RNA from two scent glands and two skin samples of this species and the salivary, 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*) was also extracted using the RNeasy mini kit (Qiagen) with on-column DNase digestion. Only a single corn snake skin sample provided RNA of high enough quality for sequencing. 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 and species were pooled and assembled using Trinity (Grabherr et al. 2011) (sequence and assembly metrics are provided in Appendix 14-16). Putative venom toxin amino acid sequences were aligned using

ClustalW (Larkin et al. 2007) and maximum likelihood trees constructed using the Jones-Taylor-Thornton (JTT) model with 500 Bootstrap replicates. Transcript abundance estimation was carried out using RSEM (Li and Dewey 2011) as a downstream analysis of Trinity (version trinityrnaseq_r2012-04-27). Sets of reads were mapped to species-specific reference transcriptome assemblies (Appendix 17) to allow comparison between tissues on a per-species basis and all results values shown are in FPKM (Fragments Per Kilobase of exon per Million fragments mapped). Individual and mean FPKM values for each gene per tissue per species are given in Appendix 18-22. All transcript abundance values given within the text are based on the average transcript abundance per tissue per species to account for variation between individual samples. Transcriptome reads were deposited in the European Nucleotide Archive (ENA) database under accession #ERP001222 and GenBank under the run accessions #SRR1287707 and #SRR1287715. Genes used to reconstruct phylogenies are deposited in GenBank under the project accession PRJNA255316.

4.3 Results

Many genes previously claimed to be venom toxins are in fact expressed in multiple tissues (Figure 4.2) and transcripts encoding these genes show no evidence of consistently elevated expression level in venom or salivary glands compared to other tissues (Appendix 18-22). Only two putative venom toxin genes (*l-amino acid oxidase b2* and *PLA_2 IIA-c*) showed evidence of a venom gland-specific splice variant across our multiple tissue data sets. There also appears to have been several cases of mistaken identity, where non-orthologous genes have been used to claim conserved, ancestral expression and instances of identical sequences being annotated as two distinct genes (see later sections). It is therefore suggestive that the putative ancestral Toxicoferan venom toxin genes do not encode toxic venom components in the majority of species and that the apparent venom gland-specificity of these genes is a side-effect of incomplete tissue sampling. These analyses show that neither increased expression in the venom gland nor the production of venom-specific splice variants can be used to support continued claims for the toxicity of these genes.

	Eublepharis macularius			Python regius			Echis coloratus			Pantherophis guttatus			Opheodrys aestivus		
	SAL	SCG	SK	SAL	SCG	SK	VG	SCG	SK	SAL	SCG	SK	SAL	SCG	SK
ache - transcript 1	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-
ache - transcript 2	-	-	-	-	+	-	-	+	+	-	-	-	-	+	-
complement c3/cvf	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+
cystatin-e/m	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
cystatin-f	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+
dipeptidyl peptidase 3	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
dipeptidyl peptidase 4	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
esp-e1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ficolin	+	+	+	+	-	-	+	+	+	+	+	+	+	+	-
kallikrein	+	+	-	+	+	+	+	+	+	-	+	-	-	+	-
kunitz 1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
kunitz 2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
lysosomal acid lipase a	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
lysosomal acid lipase b	-	-	-	-	+	-	+	+	-	-	-	-	-	-	-
nerve growth factor	+	+	+	-	-	+	+	+	+	+	+	-	-	+	+
PLA_2 IIE	+	-	-	+	-	-	+	-	-	+	-	-	+	-	-
phospholipase b	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
renin	-	+	-	-	-	-	+	+	-	-	-	-	-	-	-
3ftx-a	-	-	-	+	-	-	+	+	+	+	+	+	+	-	+
3ftx-b	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-
vespryn	-	+	+	-	+	+	-	-	-	+	+	-	-	+	+
waprin	+	+	+	-	-	-	+	-	+	+	-	-	+	+	+
vegf-a	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
vegf-b	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
vegf-c	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
vegf-f	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
l-amino acid oxidase a	+	+	+	-	-	-	+	+	+	-	-	-	-	-	-
l-amino acid oxidase b1	-	-	-	-	+	-	+	-	-	-	+	-	-	+	-
l-amino acid oxidase b2	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
crotamine-like	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
crisp-a	-	-	-	-	+	-	+	+	+	+	+	+	-	+	+
crisp-b	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-
c-type lectin a	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
c-type lectin b-h,j-k	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
c-type lectin i	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
PLA_2 IIA-a & b	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-
PLA_2 IIA-c,d,e	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
serine proteases a-f	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
Pgu svmp-a & b	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
svmp-a-g, i-k,m,p,q,t	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
svmp-h,r,s,u	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-
svmp-l	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
svmp-n	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
svmp-o	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-

Expressed >1 FPKM + Expressed <1 FPKM + Not found -

Figure 4.2. Tissue distribution of proposed venom toxin transcripts. The majority of transcripts proposed to encode Toxicoferan venom proteins are expressed in multiple body tissues.

Transcript order follows descriptions in the main text and those transcripts found in the assembled transcriptomes but which are assigned transcript abundance of <1 FPKM are shaded orange. VG, venom gland; SAL, salivary gland; SCG, scent gland; SK, skin.

4.3.1 Genes unlikely to represent toxic components of the Toxicofera

Based on quantitative analysis of their expression pattern across multiple species, the following genes are unlikely to represent toxic venom components in the Toxicofera clade (Vidal and Hedges 2005). The identification of these genes as non-venom is more parsimonious than alternative explanations such as the reverse recruitment of a "venom" gene back to a "body" gene (Casewell et al. 2012), which requires a far greater number of steps (duplication, recruitment, selection for increased toxicity, reverse recruitment) to have occurred in each species, whereas a "body" protein remaining a "body" protein is a zero-step process regardless of the number of species involved. The process of reverse recruitment must also be considered doubtful given the rarity of gene duplication in vertebrates (estimated to be between 1 gene per 100 to 1 gene per 1000 per million years (Lynch and Conery 2000; Lynch and Conery 2003; Cotton and Page 2005)) (see next chapter).

<u>Acetylcholinesterase</u>

Identical *acetylcholinesterase* (*ache*) transcripts were found to be expressed in the *E. coloratus* venom gland and scent gland (transcript 1) and an additional splice variant expressed in skin and scent gland (transcript 2). Whilst the previously known splice variants in banded krait (*Bungarus fasciatus*) are differentiated by the inclusion of an alternative exon, analysis of the *E. coloratus ache* genomic sequence (accession number KF114031) reveals that the shorter transcript 2 instead comprises only the first exon of the *ache* gene, with a TAA stop codon that overlaps the 5' GT dinucleotide splice site in intron 1. *ache* transcript 1 is expressed at a low level in the venom gland (6.60 FPKM) and is found in multiple tissues in all study species (Figure 4.2), as well as corn snake vomeronasal organ and garter snake liver. The shorter transcript 2 is found most often in skin and scent glands (Figures 4.2 and 4.3).

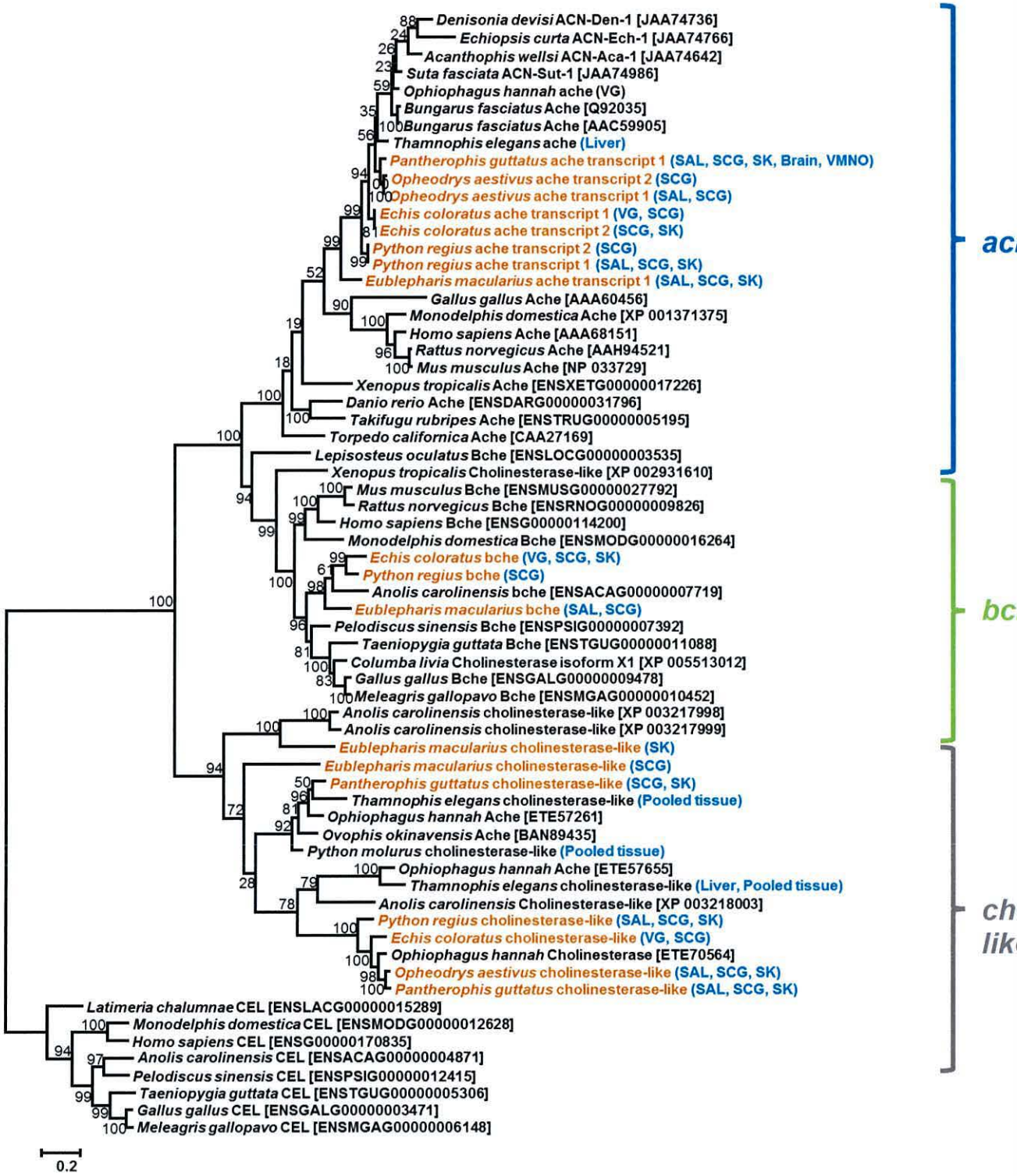


Figure 4.3. Maximum likelihood tree of *acetylcholinesterase* (*ache*), *butyrylcholinesterase* (*bche*) and *cholinesterase-like* sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland; VMNO, vomeronasal organ). Numbers above branches correspond to Bootstrap values for 500 replicates.

The low expression level and diverse tissue distribution of transcripts of this gene suggest that *acetylcholinesterase* does not represent a Toxicoferan venom toxin. Whilst some ACHE activity has been recorded in the oral secretions and venoms of a range of snakes, including opisthoglyphous (rear-fanged) colubrids (Mackessy 2002), experiments with these secretions shows that several hours are needed to achieve complete neuromuscular blockage. It should also be noted that the most frequently cited sources for the generation of a toxic version of *ache* in banded krait via alternative splicing include statements that *ache* "does not appear to contribute to the toxicity of the venom" (Cousin et al. 1998), is "not toxic to mice, even at very high doses" (Cousin et al. 1996a) and is "neither toxic by itself nor acting in a synergistic manner with the toxic components of venom" (Cousin et al. 1996b). Indeed, even purified ACHE from *Bungarus fasciatus*, despite showing a 96-fold increase in *ache* activity compared to crude venom, was non-toxic at doses above 80mg/kg (Cousin et al. 1996a).

AVIT

Only a single transcript encoding an AVIT peptide was detected in this dataset, in the scent gland of the rough green snake (Figure 4.4). The absence of this gene in all of our venom and salivary gland datasets, as well as the venom glands of the king cobra, eastern coral snake and eastern diamondback rattlesnake, and the limited number of sequences available on Genbank (one species of snake, *Dendroaspis polylepis* (accession number P25687) and two species of lizard, *Varanus varius* and *Varanus komodoensis* (accession numbers AAZ75583 and ABY89668 respectively)) despite extensive sampling, would suggest that it is unlikely to represent a conserved Toxicoferan venom toxin.

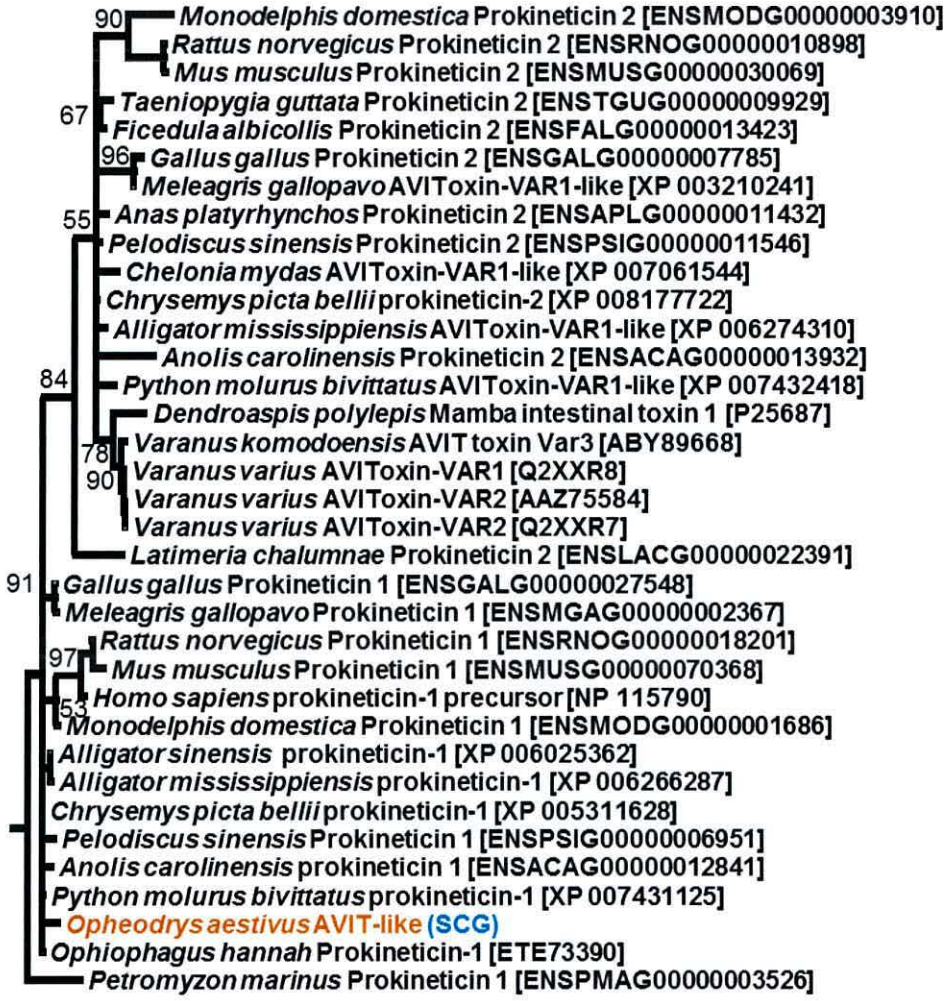


Figure 4.4 Maximum likelihood tree of AVIT peptide sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

Complement C3 ("cobra venom factor")

Identical transcripts encoding *complement c3* were found in all tissues in all species, with the exception of royal python skin (Figures 4.2 and 4.5) and only a single *complement c3* gene is found in the *E. coloratus* genome. These findings, together with the identification of transcripts encoding this gene in the liver, brain, vomeronasal organ and tissue pools of various other reptile species (Figure 3) demonstrate that this gene does not represent a Toxicoferan venom toxin. However, the grouping of additional *complement c3* genes in the king cobra (*Ophiophagus hannah*) and monocled cobra (*Naja kaouthia*) in Figure 4.5 does support a duplication of this gene somewhere in the Elapid lineage. One of these paralogs may therefore represent a venom toxin in at least some of these more derived species and the slightly elevated expression level of this gene in the venom or salivary gland of some of the study species suggests that *complement c3* has been exapted (Gould and Vrba 1982) to become a venom toxin in the Elapids. It seems likely that the identification of the non-toxic paralog in other species (including veiled chameleon (*Chamaeleo calyptratus*), spiny-tailed lizard (*Uromastyx aegyptia*) and Mitchell's water monitor (*Varanus mitchelli*)) has contributed to confusion about the distribution of this "Cobra venom factor" (which should more rightly be called *complement c3b*), to the point where genes in alligator (*Alligator sinensis*), turtles (*Pelodiscus sinensis*) and birds (*Columba livia*) are now being annotated as venom factors (accession numbers XP 006023407-8, XP 006114685, XP 005513793, Figure 4.5).

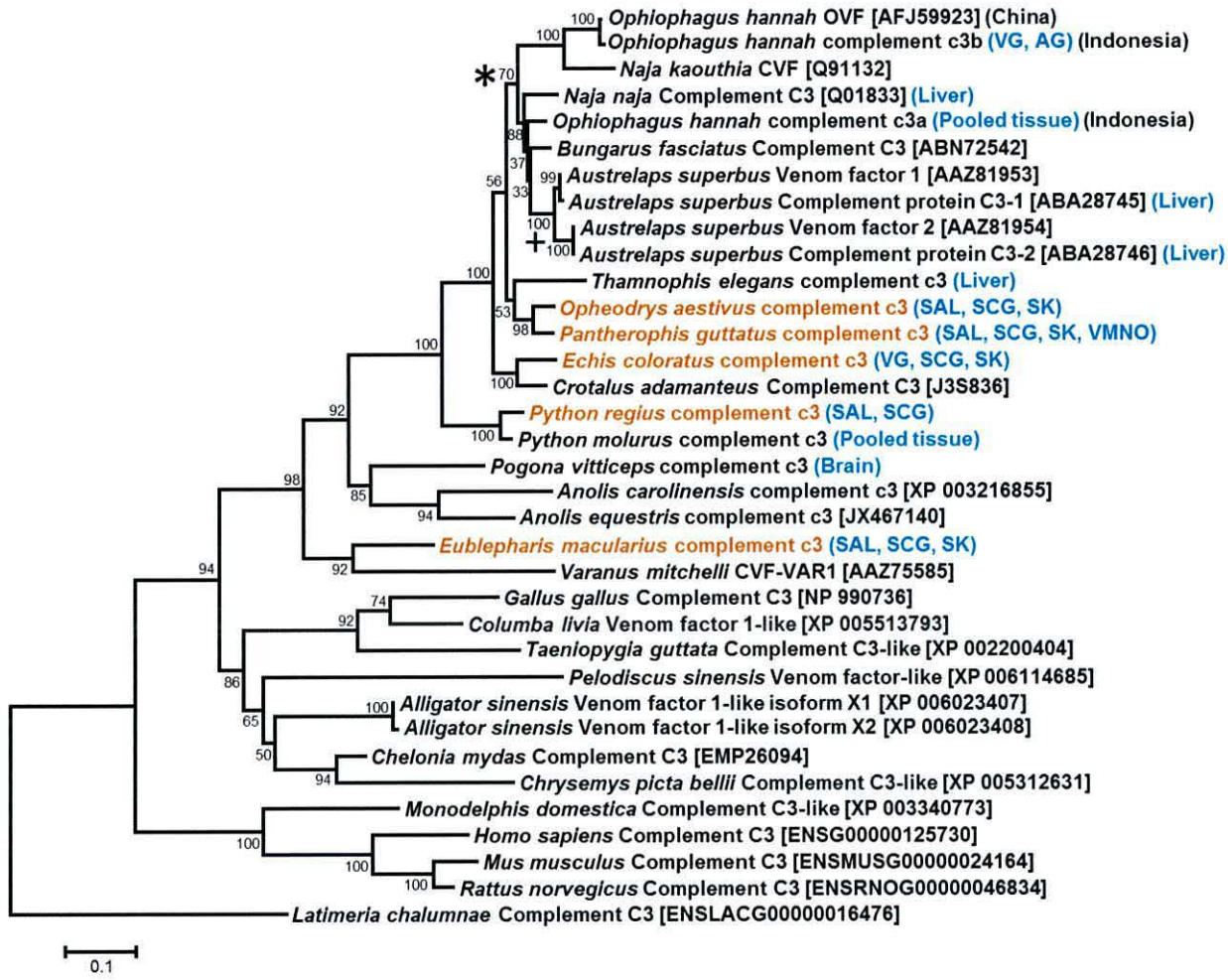


Figure 4.5. Maximum likelihood tree of *complement c3* ("*cobra venom factor*") sequences. Whilst most sequences likely represent housekeeping or maintenance genes, a gene duplication event in the elapid lineage (marked with *) may have produced a venom-specific paralog. An additional duplication (marked with +) may have taken place in *Austrelaps superbus*, although both paralogs appear to be expressed in both liver and venom gland. Geographic separation in king cobras (*Ophiophagus hannah*) from Indonesia and China is reflected in observed sequence variation. Numbers above branches are Bootstrap values for 500 replicates. Tissue distribution of transcripts is indicated using the following abbreviations: VG, venom gland; SK, skin; SCG, scent gland, AG, accessory gland; VMNO, vomeronasal organ and those genes found to be expressed in one or more body tissues are shaded blue.

Cystatin

Two transcripts encoding cystatins were found expressed in the venom gland of *E. coloratus* corresponding to *cystatin-e/m* and *f* (Figures 4.6 and 4.7). *cystatin-e/m* was detected in all tissues from all species used in this study (Figure 4.2), as well as corn snake vomeronasal organ and brain and garter snake liver and pooled tissues. The transcript encoding *cystatin-f* (which has not previously been reported to be expressed in a snake venom gland) is also expressed in the scent gland of *E. coloratus* and in the majority of other tissues of our study species. There is no evidence for a monophyletic clade of Toxicoferan cystatin-derived venom toxins and in accordance with Richards et al. (Richards et al. 2011) it seems likely that low expression level and absence of *in vitro* toxicity represents a "strong case for snake venom cystatins as essential housekeeping or regulatory proteins, rather than specific prey-targeted toxins…" Indeed, it is unclear why cystatins should be considered to be conserved venom toxins, since even the original discovery of cystatin in the venom of the puff adder (*Bitis arietans*) states that there is "…no evidence that it is connected to the toxicity of the venom" (Ritonja et al. 1987).

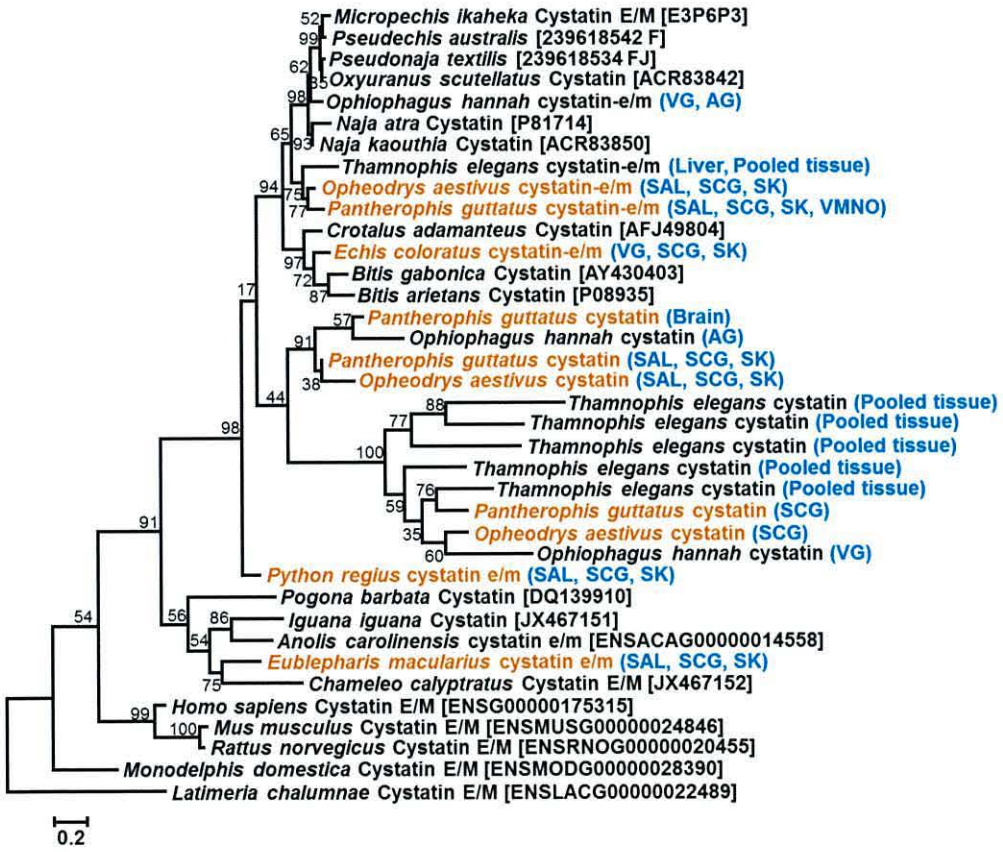


Figure 4.6. Maximum likelihood tree of *cystatin-e/m* sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text

and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland; AG, accessory gland; VMNO, vomeronasal organ). Numbers above branches correspond to Bootstrap values for 500 replicates.

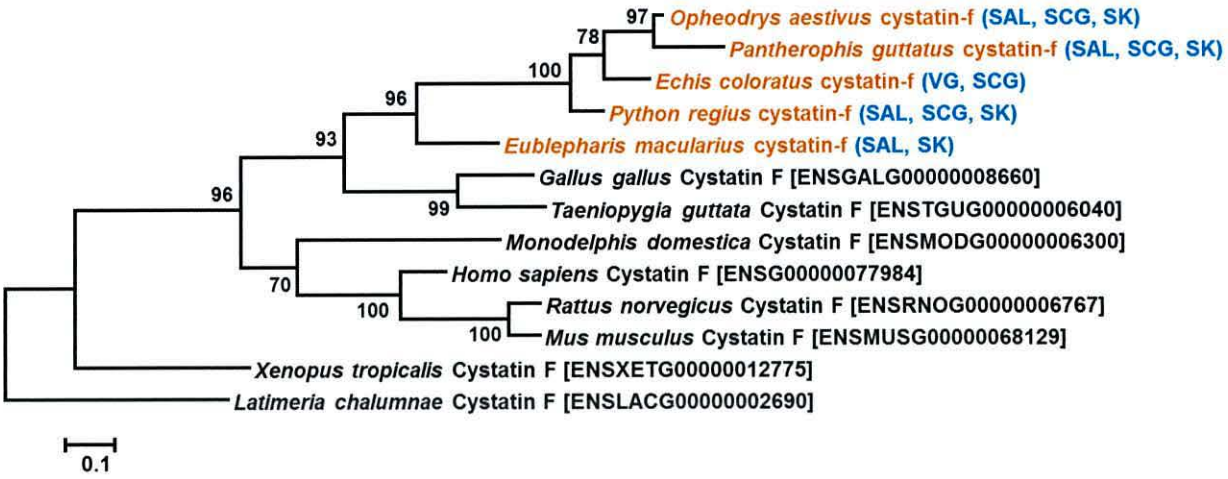


Figure 4.7. Maximum likelihood tree of *cystatin-f* sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

Dipeptidyl peptidases

Identical transcripts encoding *dipeptidyl peptidase 3* and *4* were detected in all tissues in all species except the leopard gecko (Figures 4.2, 4.8 and 4.9), and both of these have a low transcript abundance in the venom gland of *E. coloratus*. *dpp4* is expressed in garter snake liver and Anole testis and ovary and *dpp3* is also expressed in garter snake liver, king cobra pooled tissues and Bearded dragon brain (Figures 4.8 and 4.9). It is therefore unlikely the either *dpp3* or *dpp4* represent venom toxins.

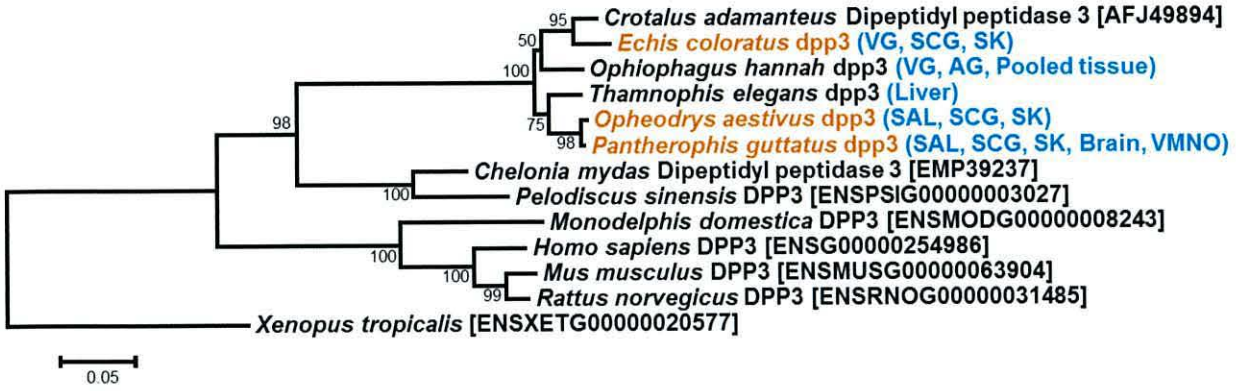


Figure 4.8. Transcripts encoding *dpp3* are found in a wide variety of body tissues, and likely represent housekeeping genes. Numbers above branches are Bootstrap values for 500 replicates. Tissue distribution of transcripts is indicated using the following abbreviations: VG, venom gland; SK, skin; SCG, scent gland, AG, accessory gland; VMNO, vomeronasal organ.

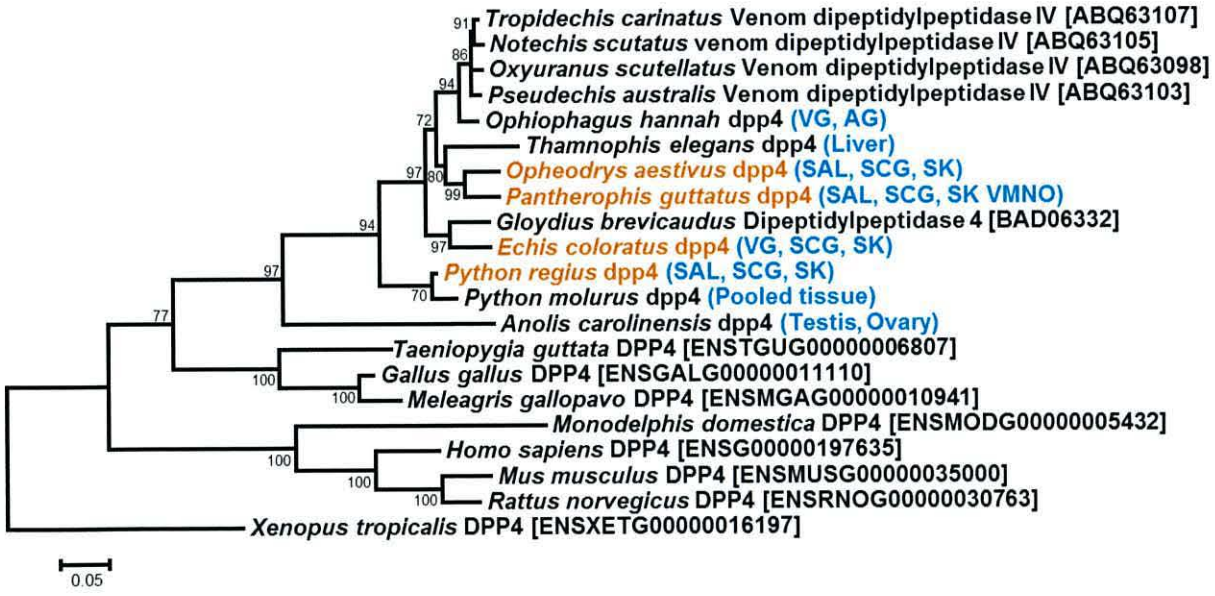


Figure 4.9. Transcripts encoding *dpp4* are found in a wide variety of body tissues, and likely represent housekeeping genes. Numbers above branches are Bootstrap values for 500 replicates. Tissue distribution of transcripts is indicated using the following abbreviations: VG, venom gland; SK, skin; SCG, scent gland, AG, accessory gland; VMNO, vomeronasal organ.

Epididymal secretory protein

One transcript encoding epididymal secretory protein (ESP) is expressed in the venom gland of *Echis coloratus* (9.09 FPKM) corresponding to type E1. This transcript is also found to be expressed at similar levels in the scent gland (13.71 FPKM) and skin (8.64 FPKM) of this species and orthologous transcripts are expressed in all three tissues of all other species used in this study (Figures 4.2 and 4.10), suggesting that this is a ubiquitously expressed gene and not a venom component. Previously described epididymal secretory protein sequences from varanids (Fry et al. 2010a) and the colubrid *Cylindrophis ruffus* (Fry et al. 2013) do not represent *esp-e1* and their true orthology is currently unclear. However, analysis of these and related sequences suggests that they are likely part of a reptile-specific expansion of esp-like genes and that the *Varanus* and *Cylindrophis* sequences do not encode the same gene (Figure 4.11). Therefore there is not, nor was there ever, any evidence that epididymal secretory protein sequences represent venom components in the Toxicofera.

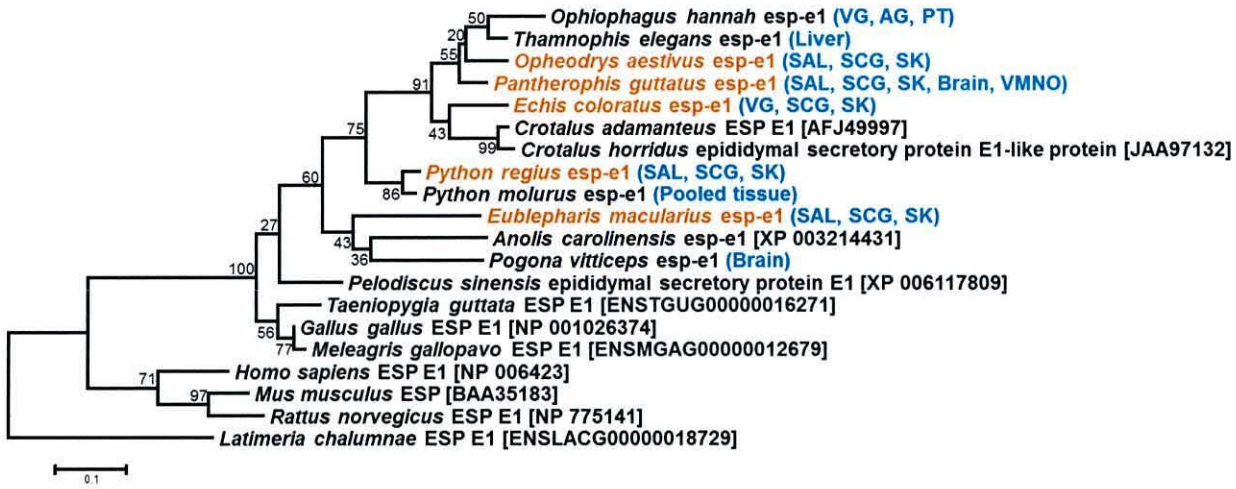


Figure 4.10. Maximum likelihood tree of *epididymal secretory protein e1* (*esp-e1*) sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland; AG, accessory gland; PT, pooled tissue; VMNO, vomeronasal organ). Numbers above branches correspond to Bootstrap values for 500 replicates.

Figure 4.11. Maximum likelihood tree of *epididymal secretory protein* sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

Ficolin ("veficolin")

One transcript encoding *ficolin* is expressed in the *E. coloratus* venom gland and identical transcripts in both scent gland and skin (Figures 4.2 and 4.12) and orthologous transcripts in all corn snake and leopard gecko tissues, as well as rough green snake salivary and scent glands and royal python salivary gland. Paralogous genes expressed in multiple tissues were also found in corn snake and rough green snake (Figure 4.2). These findings, together with additional data from available transcriptomes of pooled garter snake body tissues and bearded dragon and chicken brains show that *ficolin* does not represent a Toxicoferan venom component.

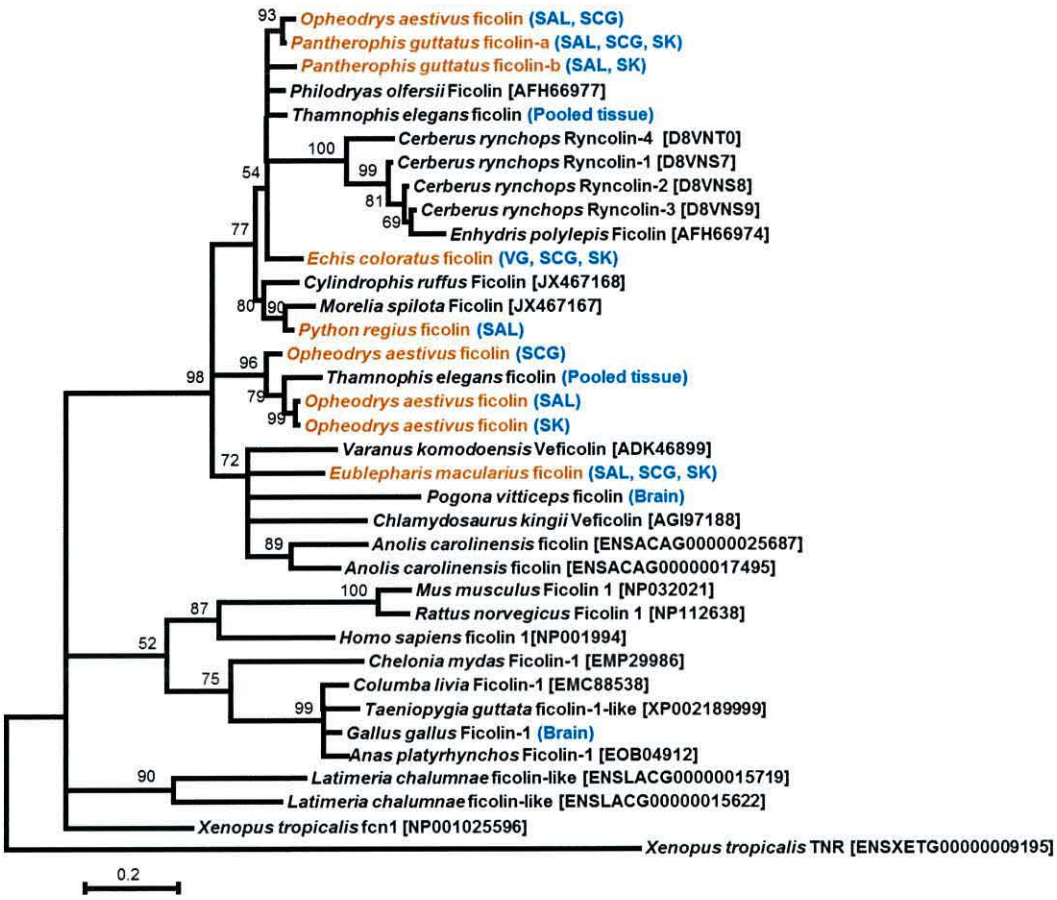


Figure 4.12. Maximum likelihood tree of *ficolin* sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

Hyaluronidase

Hyaluronidase has been suggested to be a "venom spreading factor" to aid the dispersion of venom toxins throughout the body of envenomed prey, and as such it does not represent a venom toxin itself (Kemparaju and Girish 2006). However, two hyaluronidase genes were found to be expressed in the venom gland of *E. coloratus*. The first appears to be venom gland specific (based on available data) and has two splice variants including a truncated variant similar to sequences previously characterised from *Echis carinatus sochureki* (accession number DQ840262) and *Echis pyramidum leakeyi* (accession number DQ840255) venom glands (Harrison et al. 2007). Although hyaluronidase cannot be ruled out in playing an active

(but non-toxic) role in *Echis* venom, it is worth commenting that hyaluronan has been suggested to have a role in wound healing and the protection of the oral mucosa in human saliva (Pogrel et al. 2003). The expression of hyaluronidases involved in hyaluronan metabolism in venom and/or salivary glands is therefore perhaps unsurprising.

Kallikrein

Two Kallikrein-like sequences were found in *E. coloratus*, one of which is expressed in all three tissues in this species (at a low level in the venom gland) and a variety of other tissues in the other study species, and one of which is found only in scent gland and skin (Figures 4.2 and 4.13). These genes do not represent venom toxins in *E. coloratus* and appear to be most closely-related to a group of mammalian Kallikrein (KLK) genes containing *KLK1*, *11*, *14* and *15* and probably represent the outgroup to a mammalian-specific expansion of this gene family. The orthology of previously published Toxicoferan Kallikrein genes is currently unclear and the majority of these sequences can be found in our serine protease tree (see later section).

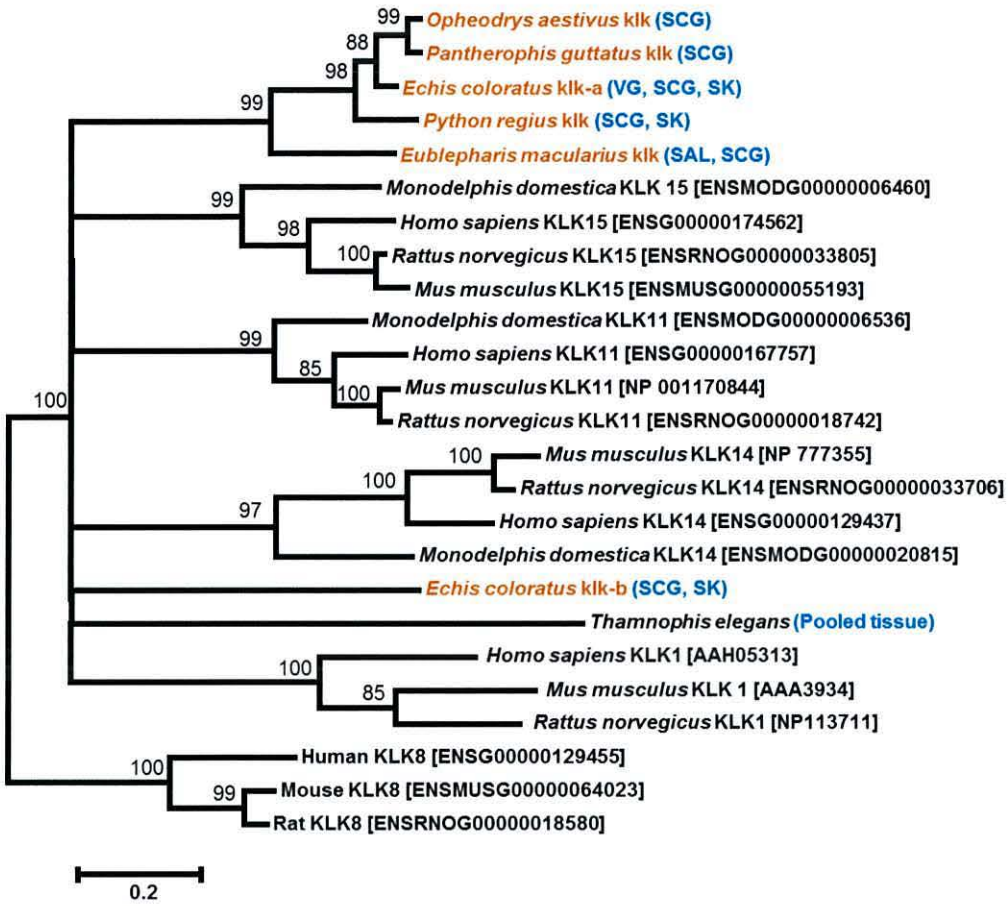


Figure 4.13. Maximum likelihood tree of *kallikrein* (*klk*) sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

Kunitz

A number of transcripts encoding Kunitz-type protease inhibitors were present in the tissue transcriptome data, with the majority of these encoding *kunitz1* and *kunitz2* genes (Figures 4.2 and 4.14). The tissue distribution of these transcripts, together with the phylogenetic position of lizard and venomous snake sequences does not support a monophyletic clade of venom gland-specific Kunitz-type genes in the Toxicofera. The presence of protease inhibitors in reptile venom and salivary glands should perhaps not be too surprising and it again seems likely that the involvement of Kunitz-type inhibitors in venom toxicity in some advanced snake lineages (in this case mamba (*Dendroaspis spp.)* dendrotoxins and krait (*Bungarus*

multicinctus) bungarotoxins (Kwong et al. 1995; Harvey 2001)) has led to confusion when non-toxic orthologs have been identified in other species.

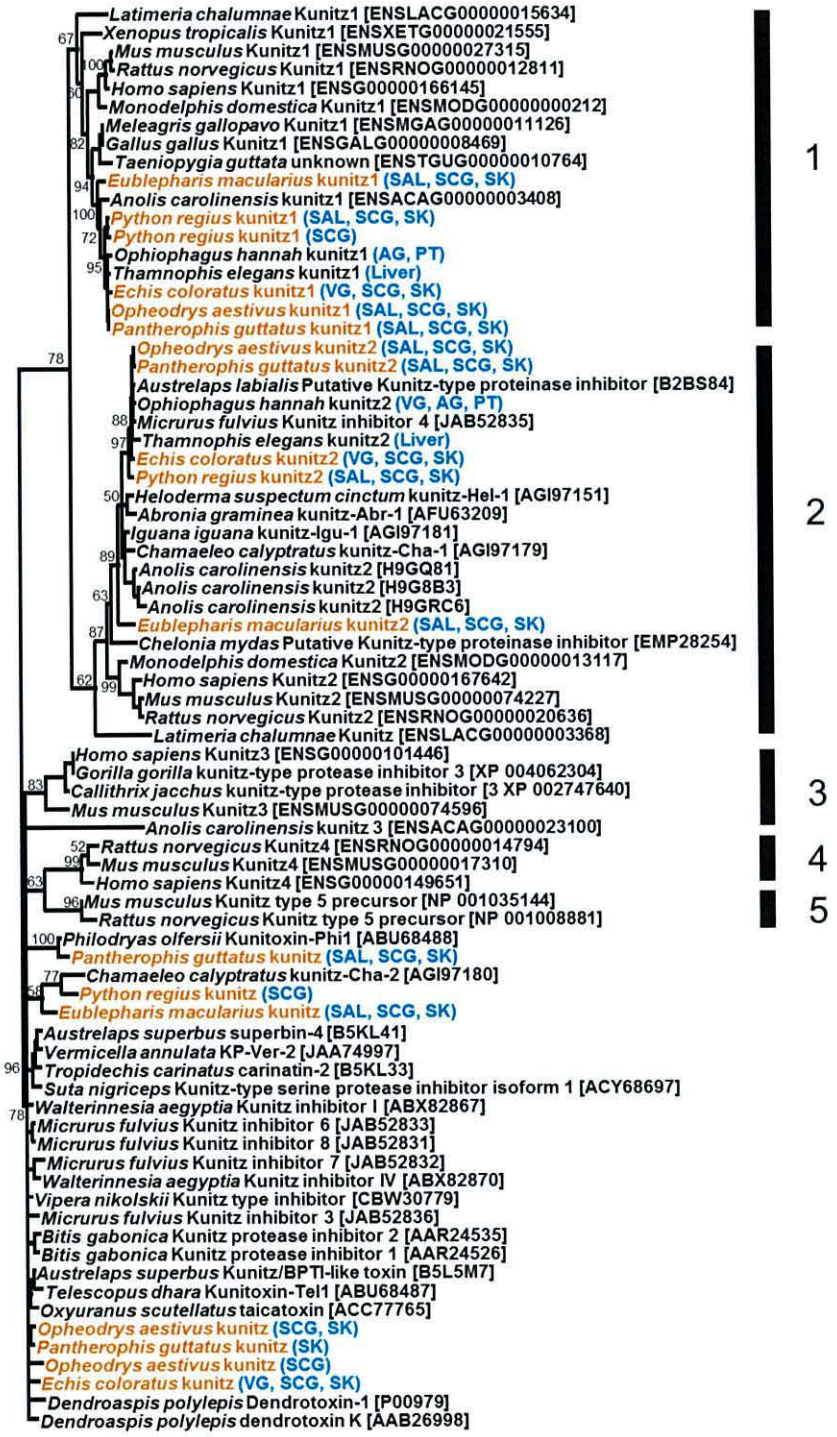


Figure 4.14. Maximum likelihood tree of *kunitz 1* and *2* sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. The three kunitz2 sequences from the Green Anole (*Anolis carolinensis*) are derived from the genome of this species as used

previously by Fry et al. (Fry et al. 2013). Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland; AG, accessory gland; PT, pooled tissue). Numbers above branches correspond to Bootstrap values for 500 replicates.

Lysosomal acid lipase

Two transcripts encoding Lysosomal acid lipase genes were detected in the *E. coloratus* venom gland transcriptome, one of which (*lipa-a*) is also expressed in skin and scent gland in this species and all three tissues of all other study species. *lipa-a*, despite not being venom gland specific, is more highly expressed in the venom gland (3,337.33 FPKM) than in the scent gland (484.49 FPKM) and skin (22.79 FPKM) of *E. coloratus,* although there is no evidence of elevated expression in the salivary glands of our other study species. As this protein is involved in lysosomal lipid hydrolysis (Warner et al. 1981) and the venom gland is a highly active tissue, it is likely that this elevated expression is related to high cell turnover. Transcripts of *lipa-b* are found at a low level in the venom and scent glands of *E. coloratus* and the scent gland of royal python (Figures 4.2 and 4.15). Neither *lipa-a* or *lipa-b* therefore encode venom toxins.

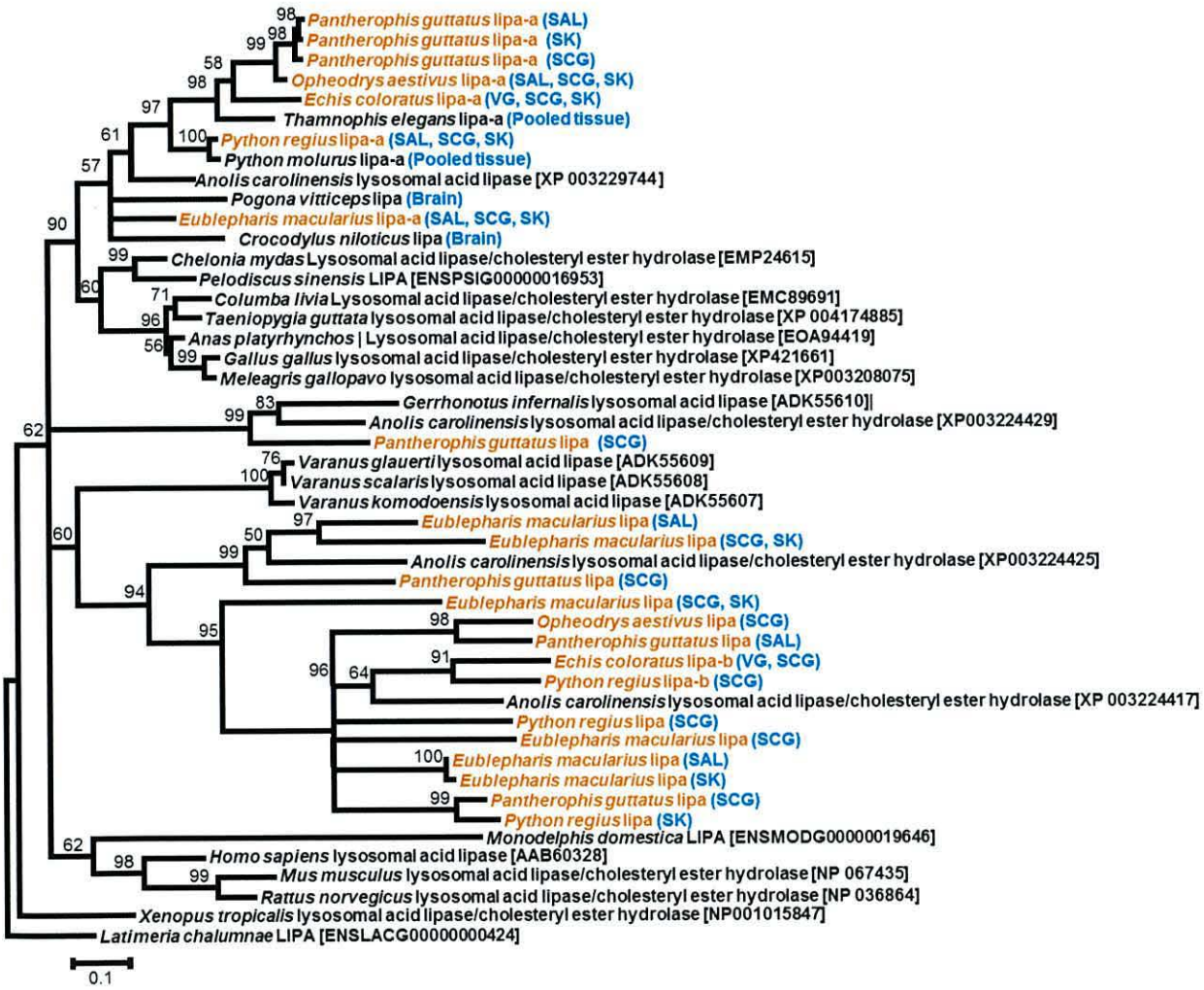


Figure 4.15. Maximum likelihood tree of *lysosomal acid lipase* (*lipa*) sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this

study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

Natriuretic peptide

Only a single natriuretic peptide-like sequence is present in the entire transcriptomic dataset, in the skin of the royal python. The absence of this gene from the rest of all other study species suggests that it is not a highly conserved Toxicoferan toxin.

Nerve growth factor

Identical transcripts encoding *nerve growth factor* (*ngf*) were found expressed in all three *E. coloratus* tissues. Transcripts encoding the orthologous gene are also found in the corn snake salivary gland and scent gland; rough green snake scent gland and skin; royal python skin and leopard gecko salivary gland, scent gland and skin (Figures 4.2 and 4.16). *ngf* is expressed at a higher level in the venom gland (525.82 FPKM) than in the scent gland (0.18 FPKM) and skin (0.58 FPKM) of *E. coloratus*, but not at an elevated level in the salivary gland of other species, again hinting at the potential for exaptation of this gene. Based on these findings, together with the expression of this gene in garter snake pooled tissues, it is likely that *ngf* does not encode a Toxicoferan toxin. However, there is evidence for the duplication of *ngf* in cobras (Figure 4.16), suggesting that it may represent a venom toxin in at least some advanced snakes (Sunagar et al. 2013). As with *complement c3*, it seems likely that the identification of non-toxic orthologs in distantly-related species has led to the conclusion that *ngf* is a widely-distributed venom toxin and confused its true evolutionary history.

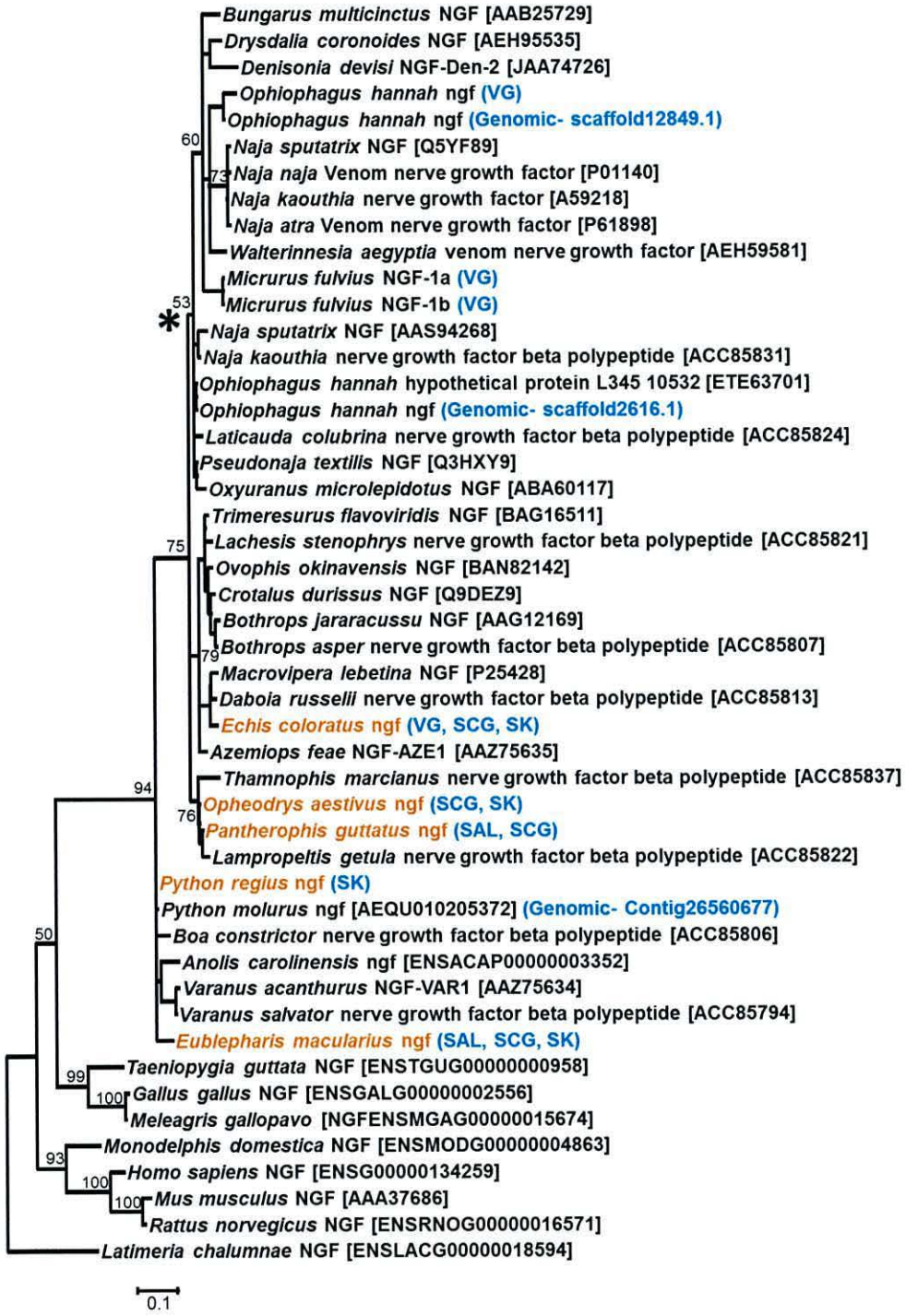


Figure 4.16. Maximum likelihood tree of *nerve growth factor* (*ngf*) sequences. A gene duplication event in the Elapid lineage (marked with *) may have produced a venom-specific paralog. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

Phospholipase A_2 (PLA_2 Group IIE)

Transcripts encoding Group IIE PLA_2 genes were found in the venom gland of *E. coloratus* and the salivary glands of all other species (Figures 4.2 and 4.17). Although this gene appears to be venom and salivary-gland-specific (based on available data), its presence in all species (including the non-Toxicoferan leopard gecko) suggests that it does not represent a toxic venom component.

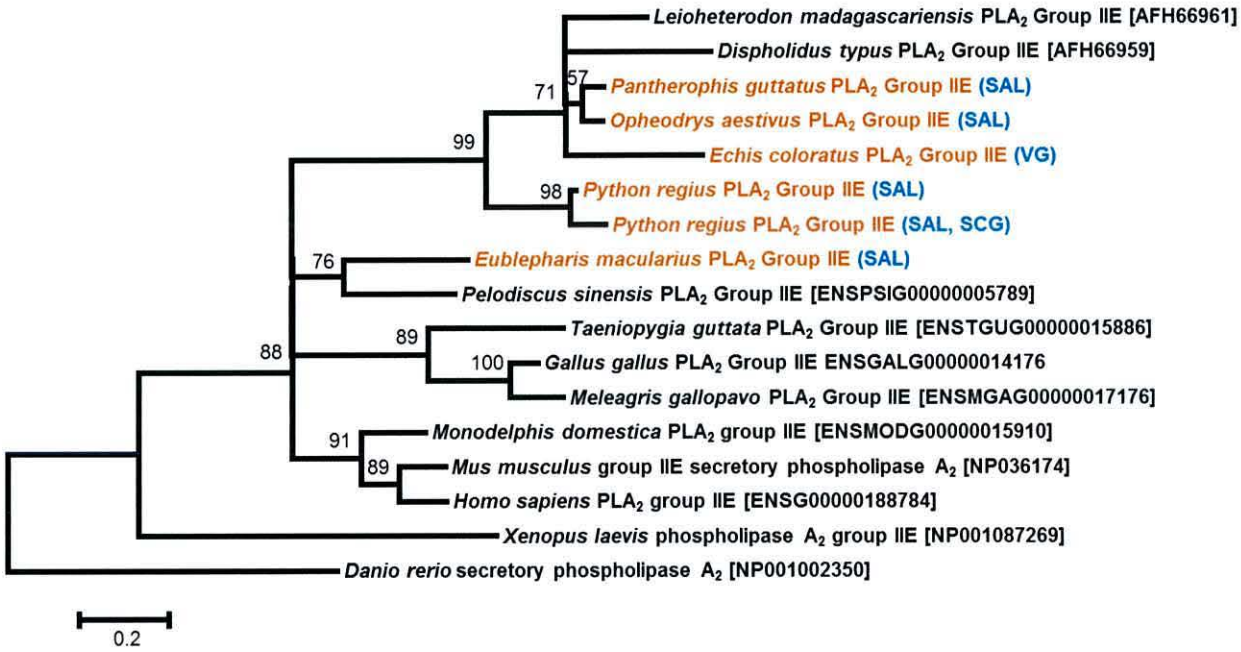


Figure 4.17. Maximum likelihood tree of *Group IIE Phospholipase A_2* (*PLA_2 Group IIE*) sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

Phospholipase B

A single transcript encoding *phospholipase b* was found to be expressed in all three *E. coloratus* tissues (Figures 4.2 and 4.18) and transcripts encoding the orthologous gene are found in all other tissues from all study species, with the exception of rough green snake salivary gland. It was also found in corn snake vomeronasal organ, garter snake liver, Burmese python pooled tissues (liver and heart) and bearded dragon brain (Figure 4.18). The two transcripts in the rough green snake and corn snake are likely alleles or the result of individual variation, and actually represent a single *phospholipase b* gene from each of these species. Transcript abundance analysis shows this gene to be expressed at a low level in all tissues from all study species. Based on the phylogenetic and tissue distribution of this gene it is unlikely to represent a Toxicoferan venom toxin.

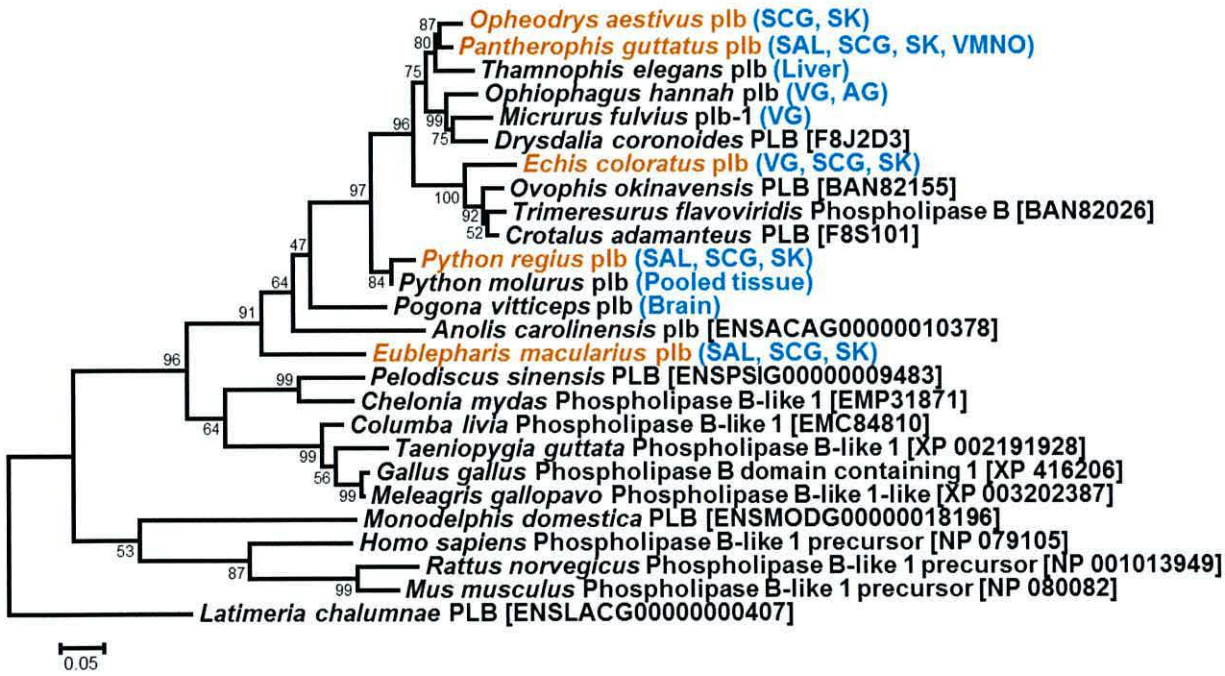


Figure 4.18. Transcripts encoding *plb* are found in a wide variety of body tissues, and likely represent housekeeping genes. Numbers above branches are Bootstrap values for 500 replicates. Tissue distribution of transcripts is indicated using the following abbreviations: VG, venom gland; SK, skin; SCG, scent gland, AG, accessory gland; VMNO, vomeronasal organ and those genes found to be expressed in one or more body tissues are shaded blue.

Renin ("renin aspartate protease")

A number of transcripts encoding renin-like genes were detected in the *E. coloratus* venom gland (Figures 4.2 and 4.19), one of which (encoding the canonical *renin*) is also expressed in the scent gland and is orthologous to a previously described sequence from the venom gland of the ocellated carpet viper (*Echis ocellatus*, accession number CAJ55260). The recently-published *Boa constrictor renin aspartate protease* (*rap*) gene (accession number JX467165 (Fry et al. 2013)) is in fact a *cathepsin d* gene, transcripts of which are found in all three tissues in all five study species. This misidentification may be due to a reliance on BLAST-based classification, most likely using a database restricted to squamate or serpent sequences. It is highly unlikely that either *renin* or *cathepsin d* (or indeed any renin-like aspartate proteases) constitute venom toxins in *E. coloratus* or *E. ocellatus*, nor do they represent basal Toxicoferan toxins.

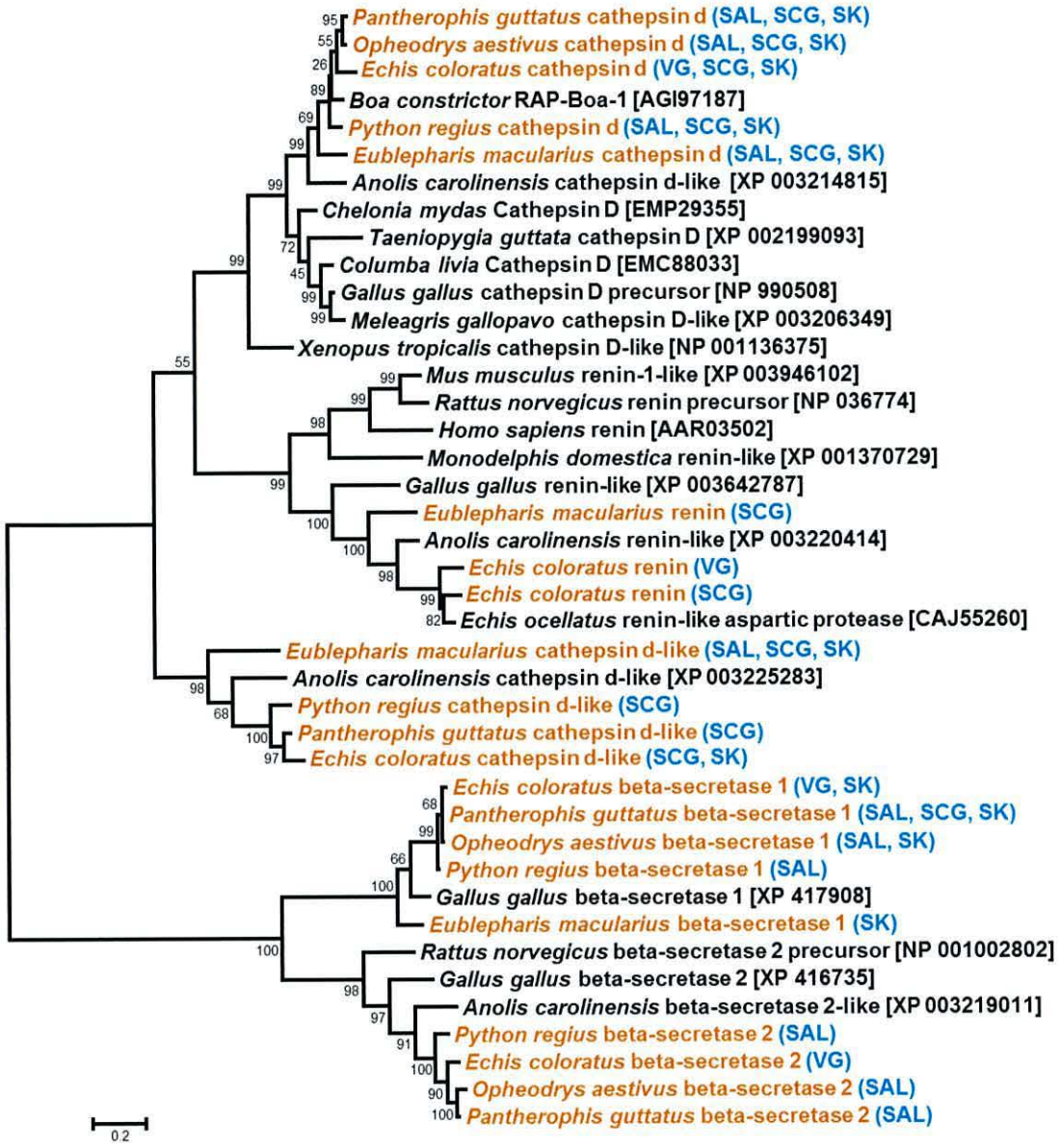


Figure 4.19. Renin-like genes are expressed in a diversity of body tissues. The recently published *Boa constrictor* "RAP-Boa-1" sequence is clearly a *cathepsin d* gene and is therefore not orthologous to the *Echis ocellatus* renin sequence as has been claimed (Fry et al. 2013). Numbers above branches are Bootstrap values for 500 replicates. Tissue distribution of transcripts is indicated using the following abbreviations: VG, venom gland; SK, skin; SCG, scent gland and those genes found to be expressed in one or more body tissues are shaded blue.

Ribonuclease

Ribonucleases have been suggested to have a role in the generation of free purines in snake venoms (Aird 2005) and the presence of these genes in the salivary glands of two species of lizard (*Gerrhonotus infernalis* and *Celestus warreni*) and two colubrid snakes (*Liophis peocilogyrus* and *Psammophis mossambicus*) has been used to support the Toxicofera (Fry et al. 2010a; Fry et al. 2012a). No orthologous ribonuclease genes were identified in any of the salivary or venom gland transcriptomes, nor in the venom gland transcriptomes from the Eastern diamondback rattlesnake, king cobra and eastern coral snake (although a wide variety of other ribonuclease genes were identified). The absence of these genes in seven members of the Toxicofera, coupled with the fact that they were initially described from only 2 out of 11 species of snake (Fry et al. 2012a) and 3 out of 18 species of lizard (Fry et al. 2010a) would cast doubt on their status as conserved Toxicoferan toxins.

Three finger toxins (3ftx)

Two transcripts encoding three finger toxin (3ftx)-like genes were found in the *E. coloratus* venom gland, one of which is expressed in all 3 tissues (*3ftx-a*) whilst the other is expressed in the venom and scent glands (*3ftx-b*). Orthologous transcripts of *3ftx-a* are found to be expressed in all three tissues of corn snake, rough green snake salivary gland and skin, and royal python salivary gland. An ortholog of *3ftx-b* is expressed in rough green snake scent gland. A number of different putative *3ftx* genes are also found in other study species, often expressed in multiple tissues (Figures 4.2 and 4.20). Based on the phylogenetic and tissue distribution of both of these genes it is likely that they do not represent venom toxins in *E. coloratus*. As with other proposed Toxicoferan genes such as *complement c3* and *nerve growth factor*, it seems likely that *3ftx* genes are indeed venom components in some species, especially cobras and other elapids (Fry et al. 2003; Vonk et al. 2013), and that the identification of their non-venom orthologs in other species has led to much confusion regarding the phylogenetic distribution of these toxic variants.

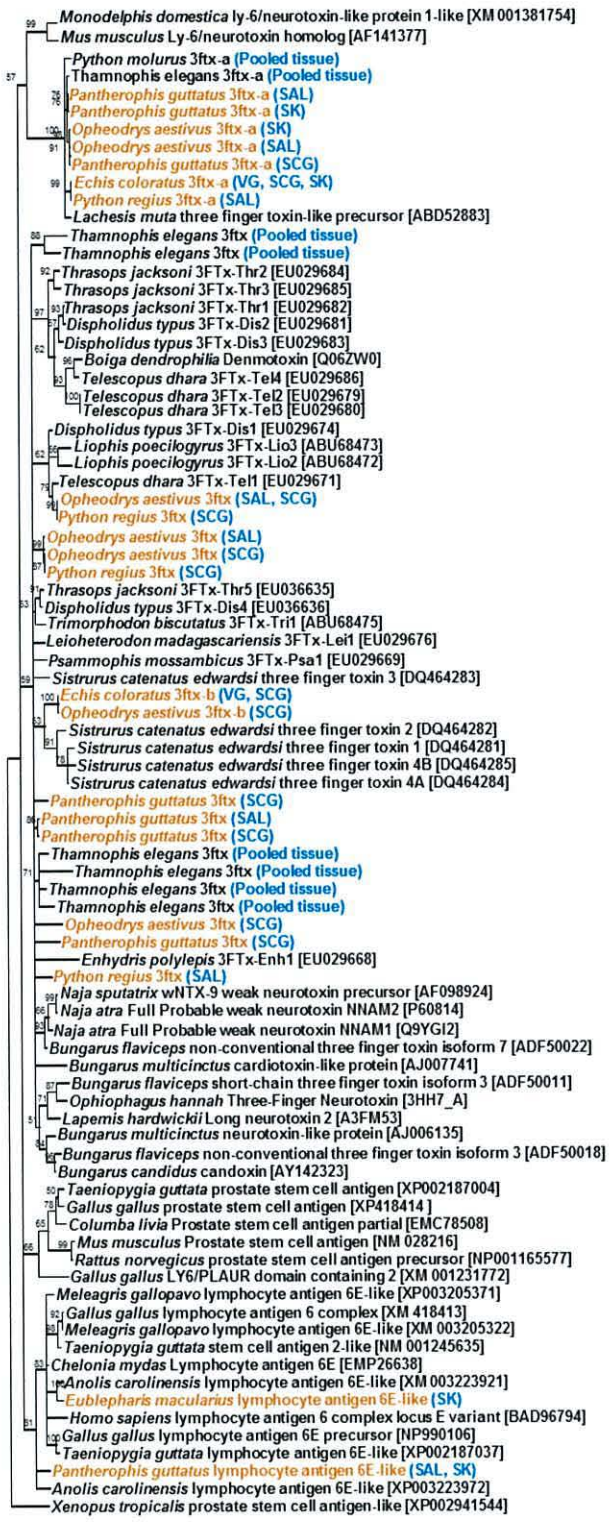


Figure 4.20. Maximum likelihood tree of *three finger toxin (3ftx)* genes. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

Vespryn

No *vespryn* transcripts were found in any *E. coloratus* tissues, although this gene is present in the genome of this species (accession number KF114032). However transcripts encoding this gene were found in the salivary and scent glands of the corn snake, and skin and scent glands of the rough green snake, royal python and leopard gecko (Figures 4.2 and 4.21). The tissue distribution of this gene in these species casts doubt on its role as a venom component in the Toxicofera.

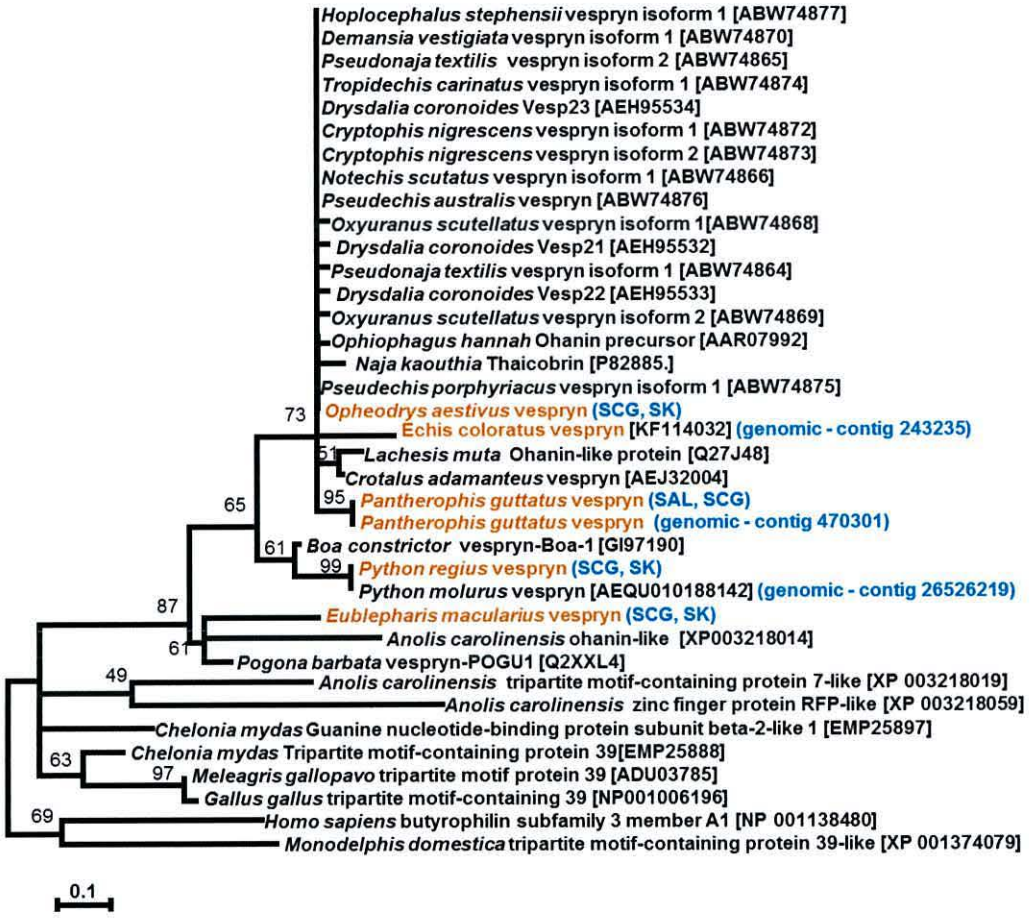


Figure 4.21. Maximum likelihood tree of *vespryn* sequences. Genbank accession numbers are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

Waprin

A number of "waprin"-like genes were found in this dataset, expressed in a diverse array of body tissues. Phylogenetic analysis (Figure 4.22) shows that previously characterised "waprin" genes (Torres et al. 2003; Nair et al. 2007; Fry et al. 2008; Rokyta et al. 2012; Aird et al. 2013) most likely represent *WAP four-disulfide core domain 2* (*wfdc2*) genes, which have undergone a squamate-specific expansion for which there is no evidence for a venom gland-specific paralog. It is unlikely therefore that these genes represent a Toxicoferan venom toxin. Indeed, the inland taipan (*Oxyuranus microlepidotus*) "Omwaprin" has been shown to be "…non-toxic to Swiss albino mice at doses of up to 10 mg/kg when administered intraperitoneally" (Nair et al. 2007) and is more likely to have an antimicrobial function in the venom or salivary gland.

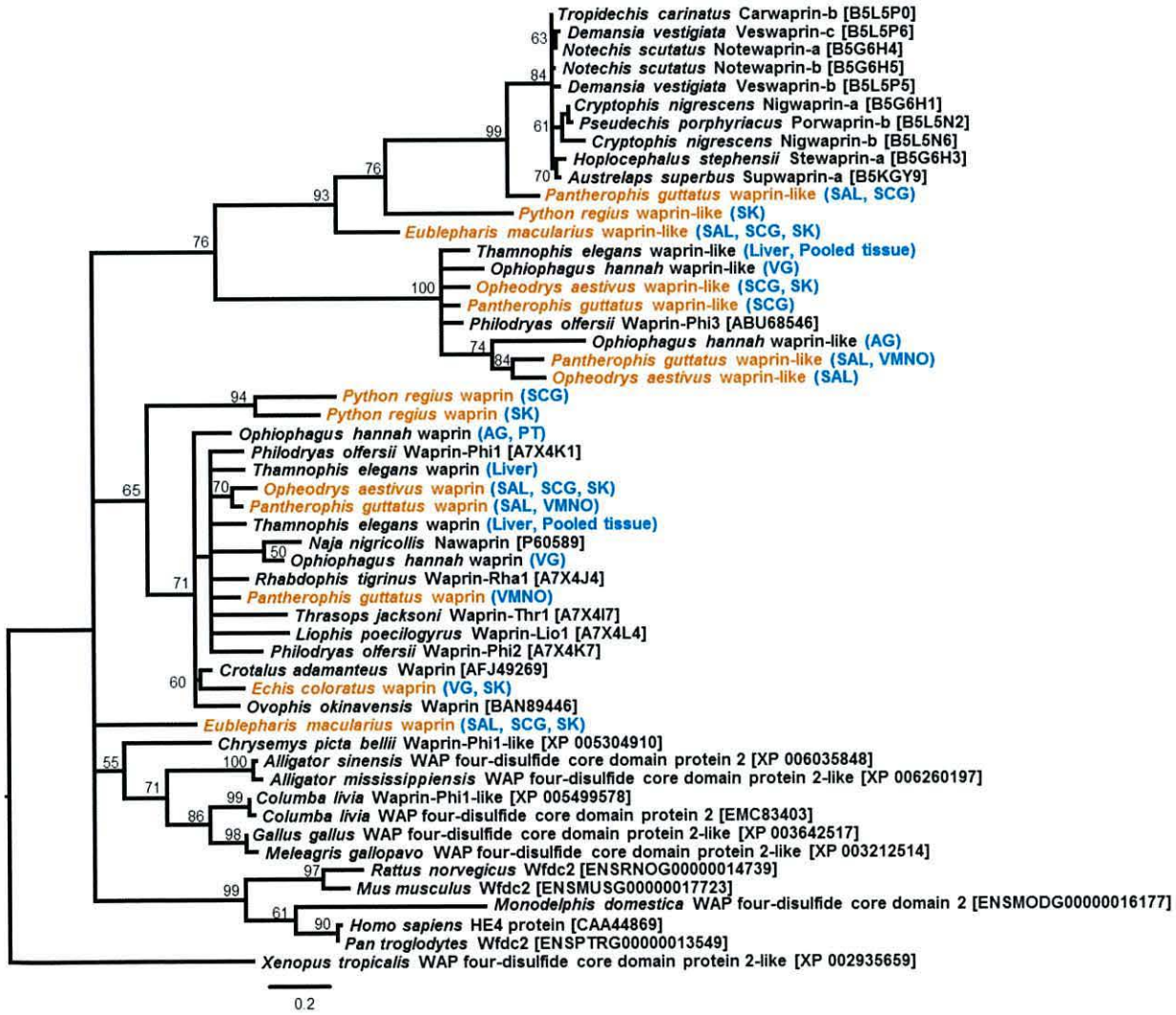


Figure 4.22. Maximum likelihood tree of *waprin* sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text

and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland; AG, accessory gland; PT, pooled tissue; VMNO, vomeronasal organ). Numbers above branches correspond to Bootstrap values for 500 replicates. The two *Echis coloratus* sequences are likely to be alleles or the result of variation in the two individuals used to generate the transcriptomes.

4.3.2 Putative venom toxins of Echis coloratus

The following genes show either a venom gland-specific expression or an elevated expression level in this tissue, but not both. As such they *may* represent venom toxins in *E. coloratus*, but further analysis is needed in order to confirm this.

Vascular endothelial growth factor

Four transcripts encoding vascular endothelial growth factor (VEGF) were found to be expressed in the venom gland of *E. coloratus*. These correspond to *vegf-a*, *vegf-b*, *vegf-c* and *vegf-f* and of these, *vegf-a*, *b* and *c* are also expressed in the skin and scent gland of this species (Figure 4.2). Transcripts encoding orthologs of these genes are expressed in all three tissues of all other species used in this study (with the exception of the absence of *vegf-a* in corn snake skin). In accordance with previous studies (Rokyta et al. 2011), there was evidence of alternative splicing of *vegf-a* transcripts in all species (Figure 4.23) although no variant appears to be tissue-specific. It is likely that a failure to properly recognise and classify alternatively spliced *vegf-a* transcripts (Aird et al. 2013) may have contributed to an overestimation of snake venom complexity.

```
                10        20        30        40        50        60        70        80        90       100
        ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Ema v1  ------------------------LSQAAPTQGDGEKHHSEVILFVDVYDRSVCRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPLEVY
Pre v1  MNFLLTWIHWGLAALLYFHNAKVLQAAPAQGDGDRQQGEVISFMKVFERSACRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPTEVY
Eco v1  ------------------------------PAHGDGDRQQGEVISFLTVYERSACRPVETMVDIFQEYPDEVEYIFKPSCVALMRCGGCCNDEALECVPTEMY
Oae v1  MNFLLTWIHWGLAALLYFHNAKVLQAAPAQGDGDRQQGEVISFMKVFERSACRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPTEVY
Pgu v1  MNFLLTWIHWGLAALLYFHNAKVLQAAPAQGDGDRQQSEVISFMKVFERSACRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPTEVY
Ema v2  MNFLLAWLRWGLAALLYFHNAKLSQAAPTQGDGEKHHSEVILFVDVYDRSVCRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPLEVY
Pre v2  MNFLLTWIHWGLAALLYFHNAKVLQAAPAQGDGDRQQGEVISFMKVFERSACRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPTEVY
Eco v2  MNFLLTWIHWGLAALLYFHNAKVLQAAPAHGDGDRQQGEVISFLTVYERSACRPVETMVDIFQEYPDEVEYIFKPSCVALMRCGGCCNDEALECVPTEMY
Pgu v2  MNFLLTWIHWGLAALLYFHNAKVLQAAPAQGDGDRQQSEVISFMKVFERSACRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPTEVY
Ema v3  MNFLLAWLRWGLAALLYFHNAKLSQAAPTQGDGEKHHSEVILFVDVYDRSVCRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPLEVY
Pre v3  MNFLLTWIHWGLAALLYFHNAKVLQAAPAQGDGDRQQGEVISFMKVFERSACRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPTEVY
Oae v3  MNFLLTWIHWGLAALLYFHNAKVLQAAPAQGDGDRQQGEVISFMKVFERSACRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPTEVY
Pgu v3  MNFLLTWIHWGLAALLYFHNAKVLQAAPAQGDGDRQQSEVISFMKVFERSACRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPTEVY
Pre v4  MNFLLTWIHWGLAALLYFHNAKVLQAAPAQGDGDRQQGEVISFMKVFERSACRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPTEVY
Eco v4  MNFLLTWIHWGLAALLYFHNAKVLQAAPAHGDGDRQQGEVISFLTVYERSACRPVETMVDIFQEYPDEVEYIFKPSCVALMRCGGCCNDEALECVPTEMY
Oae v4  MNFLLTWIHWGLAALLYFHNAKVLQAAPAQGDGDRQQGEVISFMKVFERSACRSIETMVDIFQEYPDEVEYIFKPSCVPLMRCAGCCNDEALECVPTEVY

                110       120       130       140       150       160       170       180       190       200
        ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Ema v1  NVTMEIMRIKHFQSQHINQMSFQHHSKCGCRPKKETRTKQEKKSKRGKGKGQKRKRKRGRYKPPNFHCEPCSERRKHLFVQDPQTCKCSCKFTDSRCKSR
Pre v1  NVTMEIMRLKPFQSQHINAMSFQQHSKCECRQKKEVRIRQEKKSKRGKGKGQKRKRKRGRYKPQNFHCEPCSERRKHLYKQDPLTCKCSCKFTDSRCKSK
Eco v1  NVTMEIMKLKPFQSQHIHPMSFQQHSKCECRPKKEVRIRQEKKSKREKGKGQKRKRKRGRYKPQNFHCEPCSERRKHLYKQDPLTCKCSCKFTDSRCKSK
Oae v1  NVTMEIMKLKHFQSQHIHPMSFQQHSKCECRQKKEVRIRQEKKSKRGKGKGQKRKRKRGRYKPQNFHCEPCSERRKHLYKQDPLTCKCSCKFTDSRCKSK
Pgu v1  NVTMEIMKLKHFQSQHIHPMSFQQHSKCECRQKKEVRIRQEKKSKRGKGKGQKRKRKRGRYKPQNFHCEPCSERRKHLYKQDPLTCKCSCKFTDSRCKSK
Ema v2  NVTMEIMRIKHFQSQHINQMSFQHHSKCGCRPKKETRTKQEN--------------------------HCEPCSERRKHLFVQDPQTCKCSCKFTDSRCKSR
Pre v2  NVTMEIMRLKPFQSQHINAMSFQQHSKCECRQKKEVRIRQEN--------------------------HCEPCSERRKHLYKQDPLTCKCSCKFTDSRCKSK
Eco v2  NVTMEIMKLKPFQSQHIHPMSFQQHSKCECRPKKEVRIRQEN--------------------------HCEPCSERRKHLYKQDPLTCKCSCKFTDSRCKSK
Pgu v2  NVTMEIMKLKHFQSQHIHPMSFQQHSKCECRQKKEVRIRQEN--------------------------HCEPCSERRKHLYKQDPLTCKCSCKFTDSRCKSK
Ema v3  NVTMEIMRIKHFQSQHINQMSFQHHSKCGCRPKKETRTKQEK----------------------------------------------------------
Pre v3  NVTMEIMRLKPFQSQHINAMSFQQHSKCECRQKKEVRIRQEK----------------------------------------------------------
Oae v3  NVTMEIMKLKHFQSQHIHPMSFQQHSKCECRQKKEVRIRQEK----------------------------------------------------------
Pgu v3  NVTMEIMKLKHFQSQHIHPMSFQQHSKCECRQKKEVRIRQEK----------------------------------------------------------
Pre v4  NVTMEIMRLKPFQSQHINAMSFQQHSKCECRQKKEVRIRQEK*
Eco v4  NVTMEIMKLKPFQSQHIHPMSFQQHSKCECRPKKEVRIRQEK*
Oae v4  NVTMEIMKLKHFQSQHIHPMSFQQHSKCECRQKKEVRIRQEK*

                210
        ....|....|....|..
Ema v1  QLELNERTCRCEKPRR*
Pre v1  QLELNERTCRCEKPRR*
Eco v1  QLELNERTCRCEKPRR*
Oae v1  QLELNERTCRCEKPRR*
Pgu v1  QLELNERTCRCEKPRR*
Ema v2  QLELNERTCRCEKPRR*
Pre v2  QLELNERTCRCEKPRR*
Eco v2  QLELNERTCRCEKPRR*
Pgu v2  QLELNERTCRCEKPRR*
Ema v3  ----------CEKPRR*
Pre v3  ----------CEKPRR*
Oae v3  ----------CEKPRR*
Pgu v3  ----------CEKPRR*
Pre v4
Eco v4
Oae v4
```

Figure 4.23. Alignment of conceptual translations of alternative splice variants encoding vascular endothelial growth factor A (VEGF A). Species abbreviations: Ema, leopard gecko (*Eublepharis macularius);* Pre, royal python (*Python regius*); Eco, painted saw-scaled viper (*Echis coloratus*); Oae, rough green snake (*Opheodrys aestivus*); Pgu, corn snake (*Pantherophis guttatus*).

vegf-d was only found to be expressed in royal python salivary gland and scent gland and all three tissues from leopard gecko (Figures 4.2 and 4.24). The transcript encoding VEGF-F is found only in the venom gland of *E. coloratus* and, given the absence of any Elapid *vegf-f* sequences in public databases as well as absence of this transcript in the two species of colubrid in this study, it appears that *vegf-f* is specific to vipers. Whilst *vegf-f* has a higher transcript abundance in *E. coloratus* venom gland (186.73 FPKM) than *vegf-a* (3.24 FPKM), *vegf-b* (1.28 FPKM) and *vegf-c* (1.54 FPKM), compared to other venom genes in this species (see section

4.6) it has a considerably lower transcript abundance suggesting it represents at most a minor venom component in *E. coloratus*.

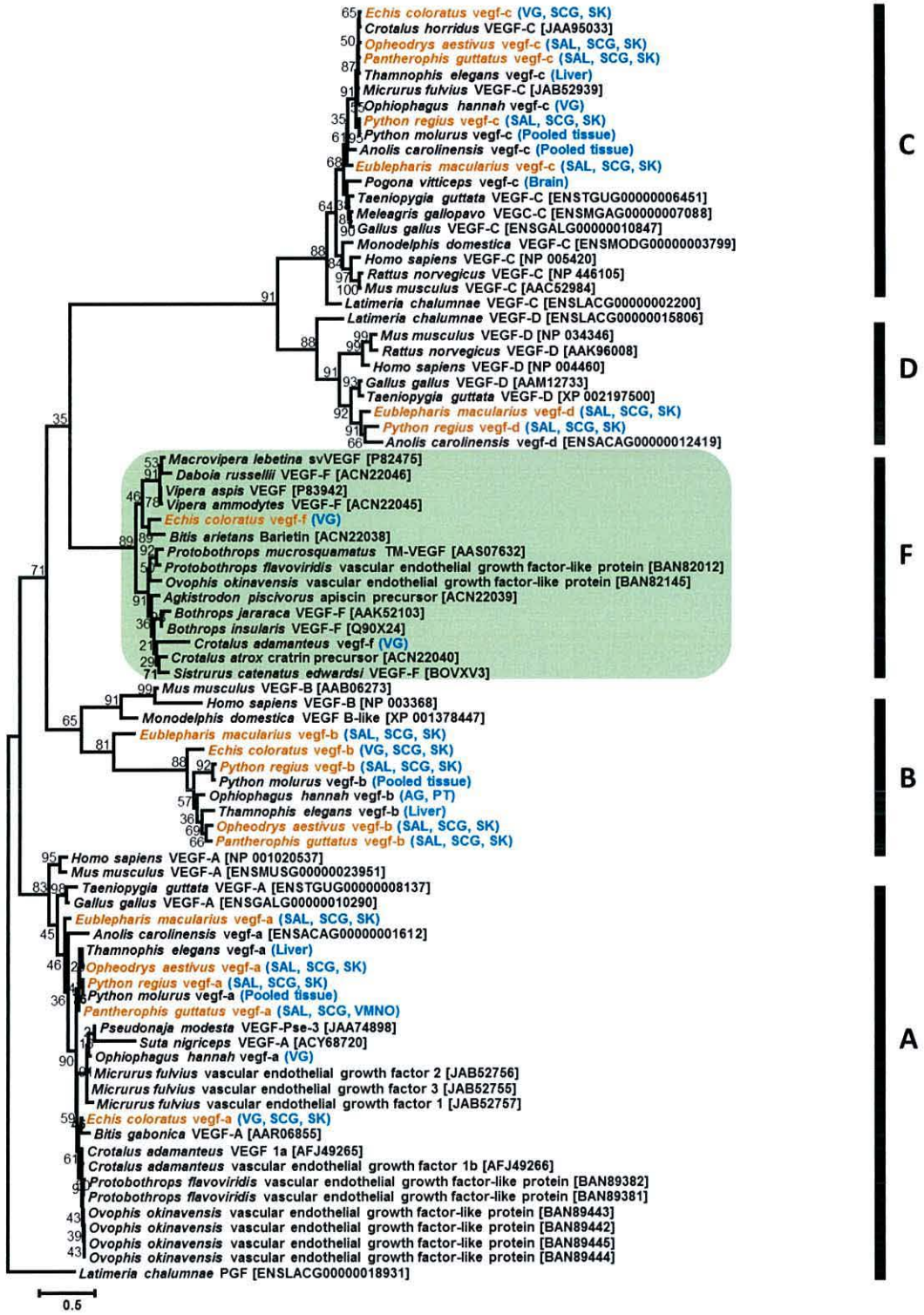


Figure 4.24. Maximum likelihood tree of *vascular endothelial growth factor* (*vegf*) sequences. Monophyletic clades representing the five members of the *vegf* family (*a-f*) are indicated and the putative viper venom-specific *vegf-f* clade is shaded green. Genbank accession numbers or

Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland; AG, accessory gland; PT, pooled tissue; VMNO, vomeronasal organ). Numbers above branches correspond to Bootstrap values for 500 replicates.

L-amino acid oxidase

Transcripts encoding two *l-amino acid oxidase (laao)* genes were detected in *E. coloratus*, one of which (*laao-b*) has two splice variants (Figures 4.2 and 4.25).

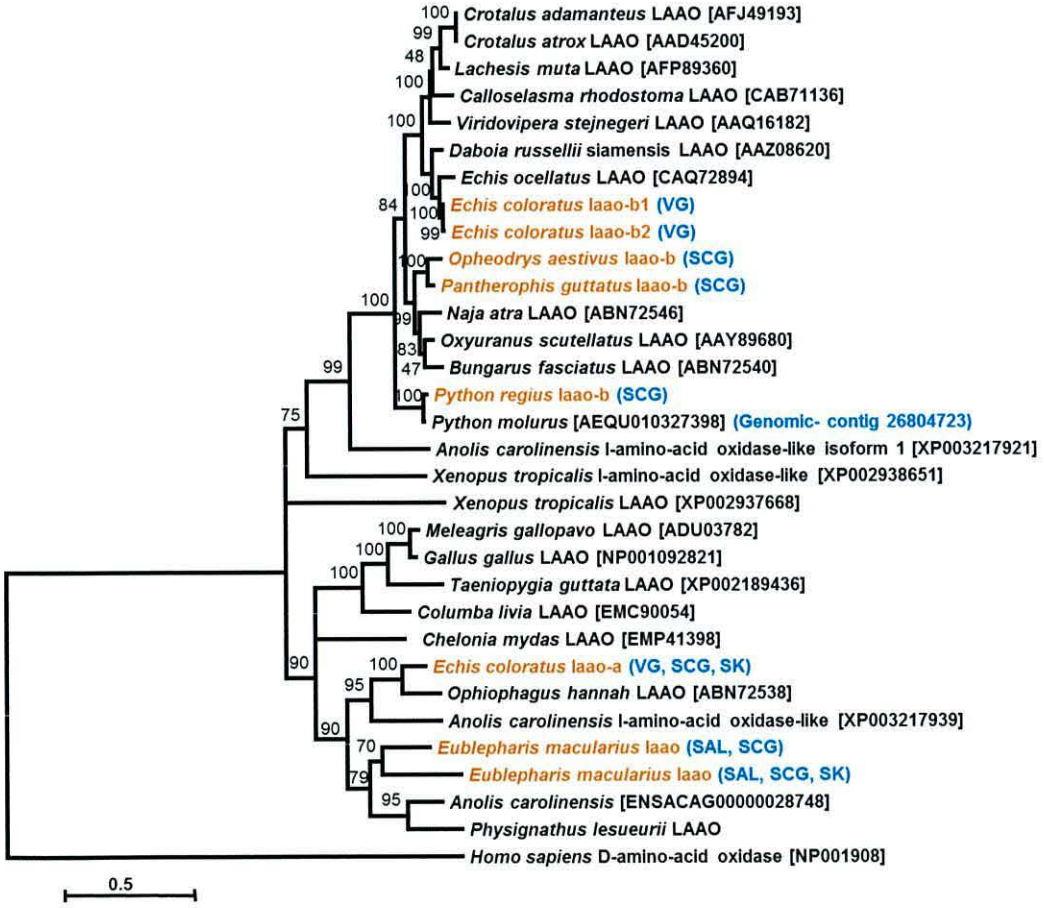


Figure 4.25. Maximum likelihood tree of *l-amino acid oxidase* (*laao*) sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets (the Australian Water Dragon (*Physignathus lesueurii*) sequence does not have a Genbank accession number and is derived from a sequence in (Fry et al. 2013)). Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

laao-a transcripts are found in all three *E. coloratus* and leopard gecko tissues. *laao-b* is venom gland-specific in *E. coloratus* (based on the available data) and transcripts of the orthologous gene are found in the scent glands of corn snake, rough green snake and royal python. The splice variant *laao-b2* may represent a venom toxin in *E. coloratus* based on its specific expression in the venom gland of this species and elevated expression level (628.84 FPKM).

Crotamine

A single *crotamine*-like transcript was found in the venom gland of *E. coloratus* (Figures 4.2 and 4.26). Related genes are found in a variety of tissues in other study species (including the scent gland of the rough green snake, the salivary gland and skin of the leopard gecko, and in all three corn snake tissues), although the short length of these sequences precludes a definitive statement of orthology. This gene may represent a toxic venom component in *E. coloratus* based on its tissue distribution, but due to its low transcript abundance (10.95 FPKM) it is likely to play a minor role, if any.

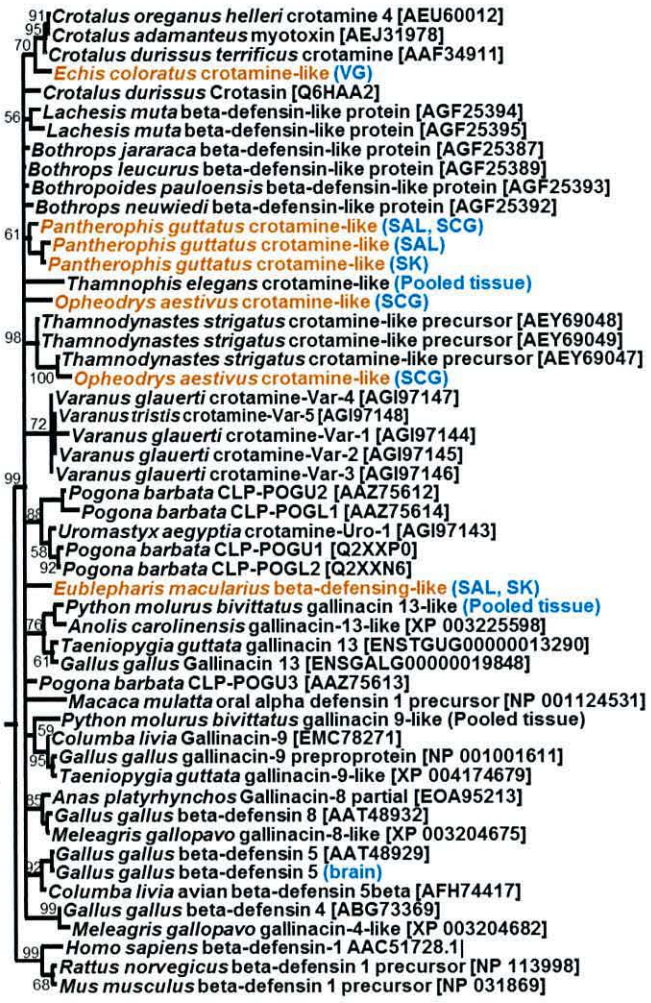


Figure 4.26. Maximum likelihood tree of *crotamine/β defensin* sequences. Genbank accession numbers are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

4.3.3 Proposed venom toxins in Echis coloratus

The following genes are found only in the venom gland of *E. coloratus* and clearly show an elevated expression level (Figure 4.32). Whilst these genes are likely to encode venom toxins in this species (Table 4.1) it should be noted that none of them support the monophyly of Toxicoferan venom toxins.

Cysteine-rich secretory proteins (CRISPs)

Transcripts encoding two distinct CRISPs are expressed in the *E. coloratus* venom gland, one of which is also found in skin and scent gland (Figure 4.2). Phylogenetic analysis of these genes (designated *crisp-a* and *crisp-b*) reveals that they appear to have been created as a result of a gene duplication event earlier in the evolution of advanced snakes (Figure 4.27). *crisp-a* transcripts are also found in all three corn snake tissues, as well as rough green snake skin and scent gland and royal python scent gland. *crisp-b* is also found in corn snake salivary gland (Figures 4.2 and 4.27) and the phylogenetic and tissue distribution of this gene suggest that it does indeed represent a venom toxin, produced via duplication of an ancestral *crisp* gene that was expressed in multiple tissues, including the salivary gland. The elevated transcript abundance of *crisp-b* (3,520.07 FPKM) in the venom gland of *E. coloratus* further supports its role as a venom toxin in this species. The phylogenetic and tissue distribution and low transcript abundance of *crisp-a* (0.61 FPKM in *E. coloratus* venom gland) shows that it is unlikely to be a venom toxin. There is no evidence of a monophyletic clade of reptile venom toxins and therefore, contrary to earlier reports (Fry et al. 2009b; Fry et al. 2010a), the CRISP genes of varanid and helodermatid lizards do not represent shared Toxicoferan venom toxins and, if they are indeed toxic venom components, they have been recruited independently from those of the advanced snakes. Regardless of their status as venom toxins, it appears likely that the diversity of CRISP genes in varanid lizards in particular (Fry et al. 2006) has been overestimated as a result of the use of negligible levels of sequence variation to classify transcripts as representing distinct gene products (see later section).

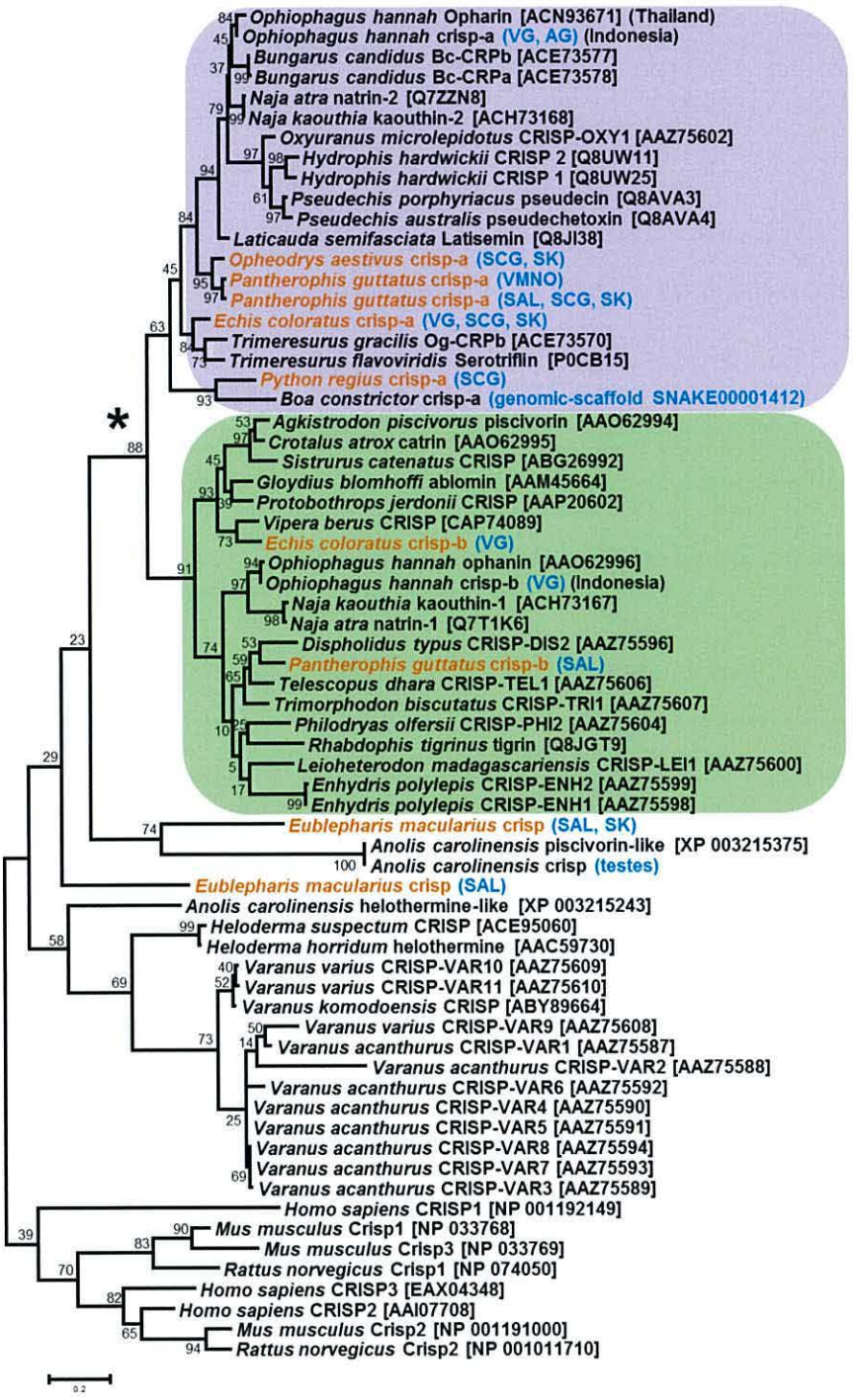


Figure 4.27. Maximum likelihood tree of *crisp* sequences. Genbank accession numbers are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland; AG, accessory gland; VMNO, vomeronasal organ). Numbers above branches correspond to Bootstrap values for 500 replicates. The * denotes the gene duplication event that produced *crisp-a* (shaded purple) and *crisp-b* (green).

C-type lectins

Transcripts encoding 11 distinct C-type lectin genes were found in the *E. coloratus* venom gland, one of which (*ctl-a*) is also expressed in the scent gland of this species. The remaining 10 genes (*ctl-b* to *k*) are found only in the venom gland and form a clade with other viper C-type lectin genes (Figures 4.2 and 4.28). Of these, 6 are highly expressed in the venom gland (*ctl-b* to *d*, *ctl-f* to *g* and *ctl-j*) with a transcript abundance range of 3,706.21-24,122.41 FPKM. The remainder of these genes (*ctl-e*, *ctl-h* to *i* and *ctl-k*) show lower transcript abundance (0.80-1,475.88 FPKM), with two (*ctl-i* and *k*) being more lowly expressed than *ctl-a* (230.06 FPKM). A number of different C-type lectin genes are found in the other study species, often expressed in multiple tissues (Figure 4.2). Therefore the 6 venom-gland specific C-type lectin genes that are highly expressed are likely to represent venom toxins in *E. coloratus* and it appears that these genes diversified via the duplication of an ancestral gene with a wide expression pattern, including in salivary/venom glands (see Chapter 5). Based on their selective expression in the venom gland (from available data) the remaining four C-type lectin genes cannot be ruled out as putative toxins, although their lower transcript abundance suggests that they are likely to be minor components in *E. coloratus* venom. It should also be noted that a recent analysis of king cobra (*Ophiohagus hannah*) venom gland transcriptome and proteome suggested that "...lectins do not contribute to king cobra envenoming" (Vonk et al. 2013).

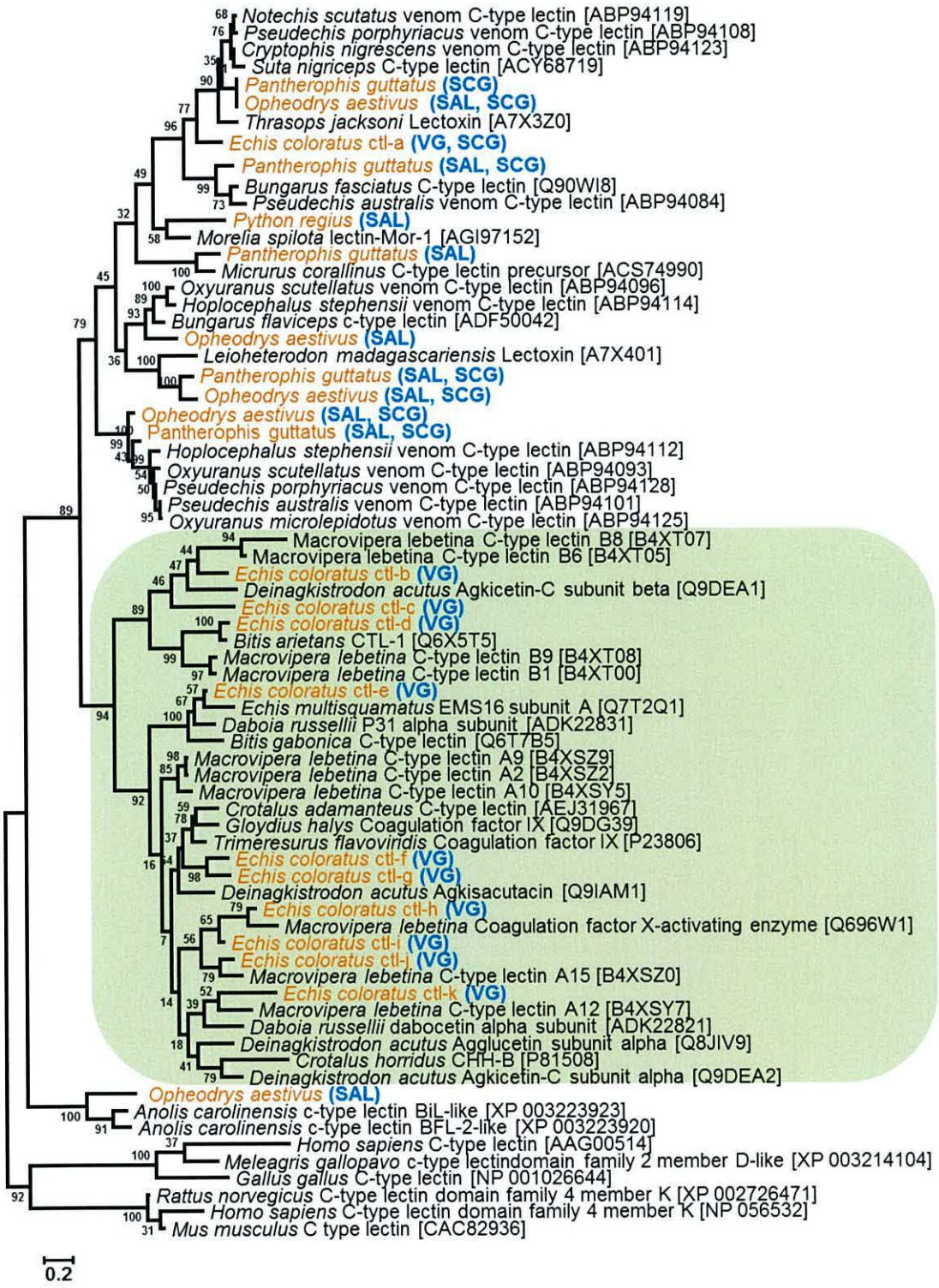


Figure 4.28. Maximum likelihood tree of *c-type lectin* (*ctl*) sequences. The putative viper venom-specific *ctl* clade is shaded green. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates

Phospholipase A_2 (PLA_2 Group IIA)

Five transcripts encoding Group IIA PLA_2 genes are expressed in *E. coloratus*, three of which are found only in the venom gland and two of which are found only in the scent gland (these latter two likely represent intra-individual variation in the same transcript) (Figures 4.2 and 4.29). The venom gland-specific transcript *PLA_2 IIA-c* is highly expressed (22,520.41 FPKM) and likely represents a venom toxin, and may also be a putative splice variant although further analysis is needed to confirm this. *PLA_2 IIA-d* and *IIA-e* show an elevated, but lower, expression level (1,677.15 FPKM and 434.67 FPKM respectively). Based on tissue and phylogenetic distribution it is proposed that these three genes may represent putative venom toxins (Table 4.1).

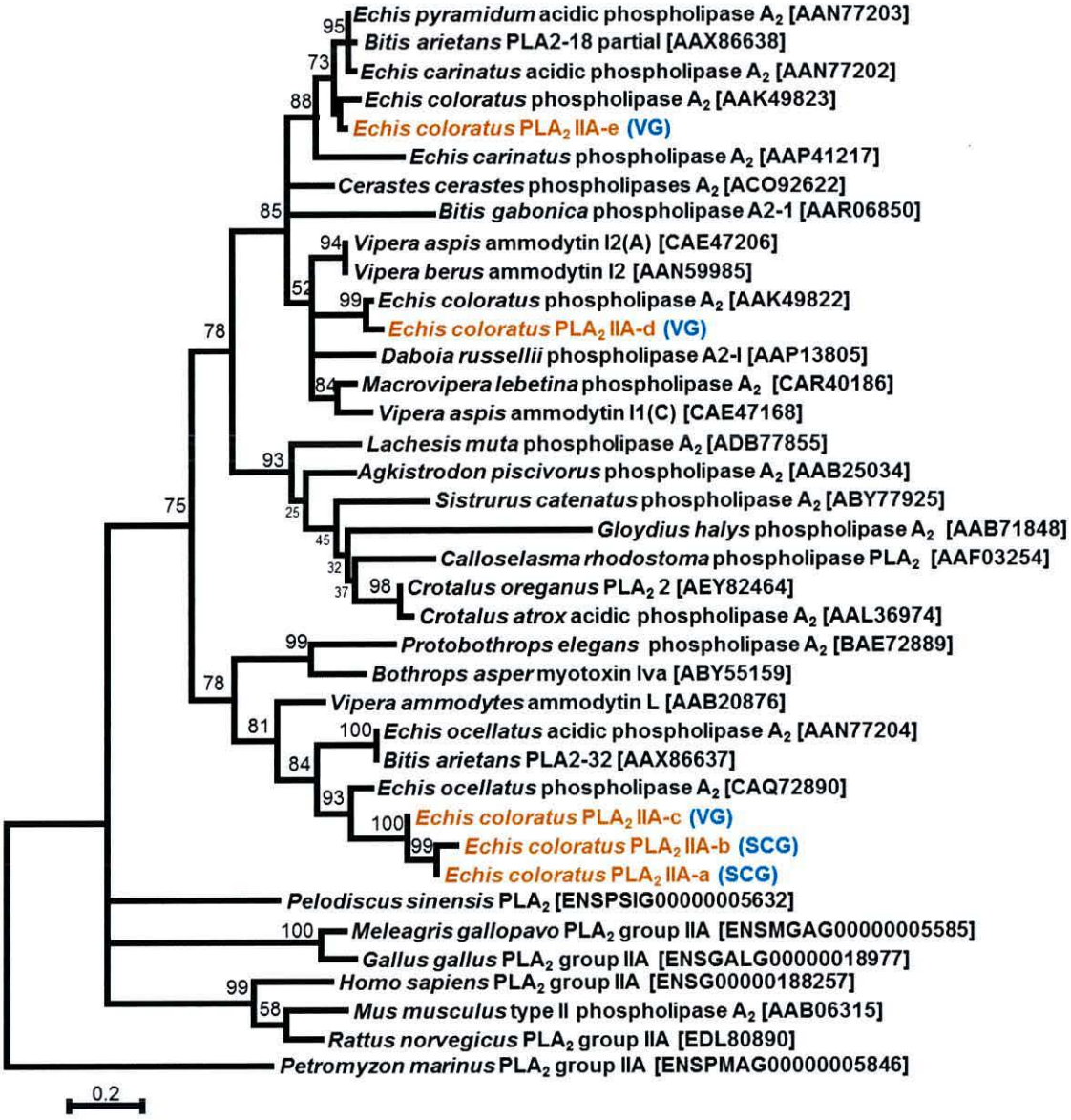


Figure 4.29. Maximum likelihood tree of *Group IIA Phospholipase A_2* (*PLA_2 Group IIA*) sequences. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates. The *Echis coloratus* sequences for *PLA_2 IIA-a* and *IIA-b* are likely to be alleles or due to individual variation between samples used in this study, and probably represent the same gene.

Serine proteases

Six transcripts encoding Serine proteases are expressed in *E. coloratus* (Figures 4.2 and 4.30) which (based on available data) are all venom gland specific. Four of these transcripts are highly expressed in the venom gland (*serine proteases a-c* and *e*; 3,076.01-7,687.03 FPKM) whilst two are expressed at a lower level (*serine proteases d* and *f*; 1,098.45 FPKM and 102.34 FPKM respectively). These results suggest that *serine proteases a, b, c* and *e* represent venom toxins whilst *serine proteases d* and *f* may represent putative venom toxins (Table 4.1).

Snake venom metalloproteinases

A total of 21 transcripts encoding snake venom metalloproteinases in *E. coloratus* were found and of these 14 are venom gland-specific, whilst another (*svmp-n*) is expressed in the venom gland and scent gland (albeit at a very low level, 0.03 FPKM, in the scent gland). Five remaining genes are expressed in the scent gland only whilst another is expressed in the skin (Figures 4.2 and 4.31). Of the 14 venom gland-specific SVMPs, 4 are highly expressed (5,552.84-15,118.41 FPKM). In the absence of additional data, the 13 venom gland-specific *svmp* genes are likely to repesent venom toxins in this species (Table 4.1).

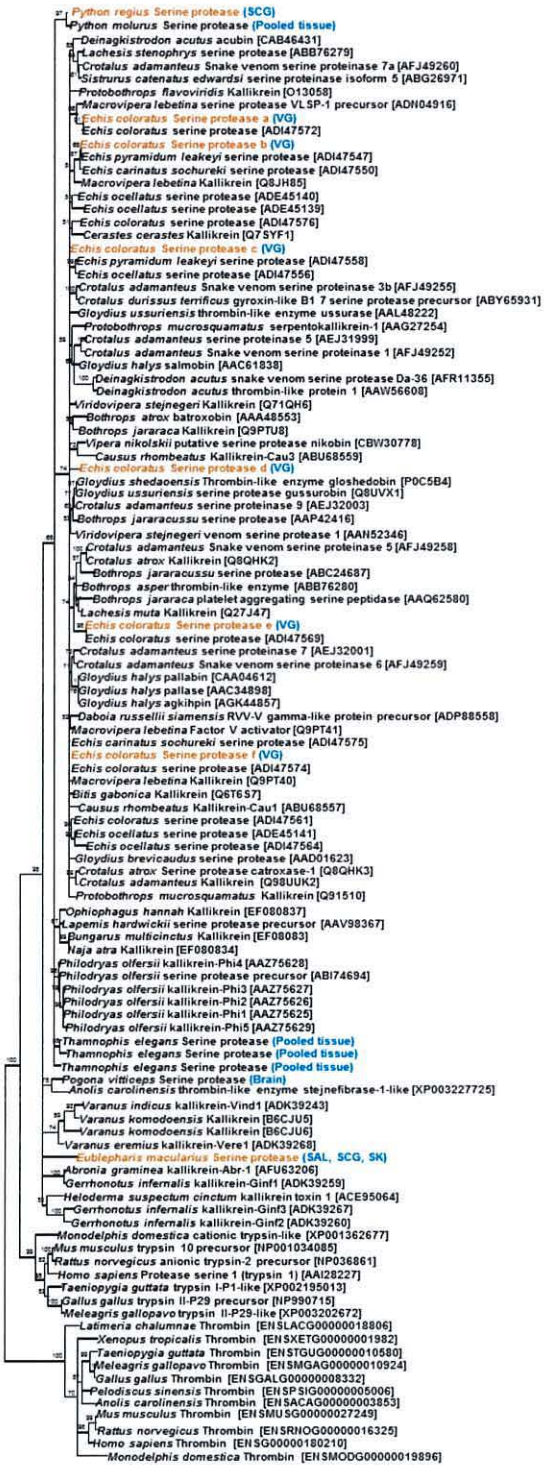


Figure 4.30. Maximum likelihood tree of *serine protease* (*sp*) genes. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

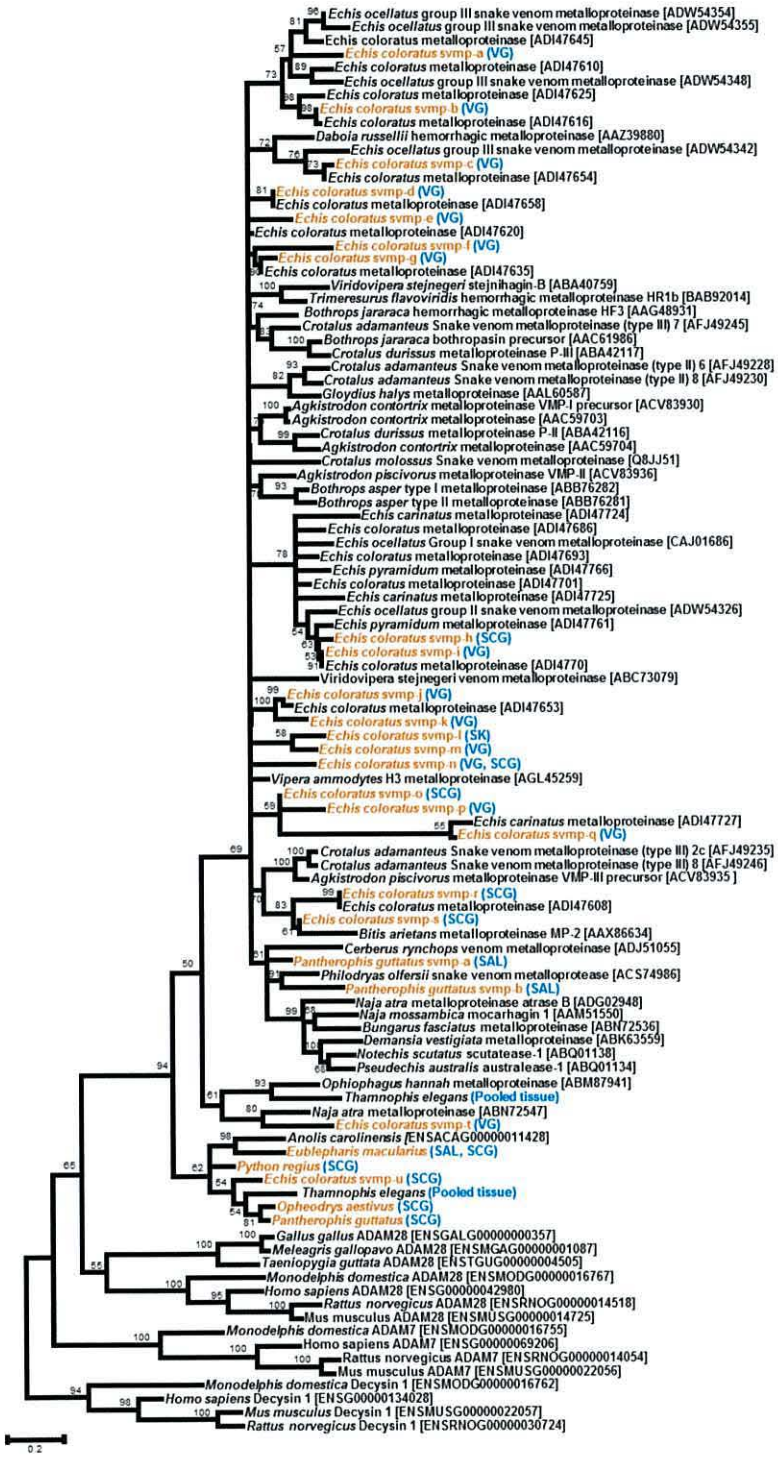


Figure 4.31. Maximum likelihood tree of *snake venom metalloproteinase* (*svmp*) genes. Genbank accession numbers or Ensembl gene IDs are given in square brackets. Sequences from this study are in orange text and the location of transcripts is provided in bracketed blue text (VG, venom gland; SAL, salivary gland; SK, skin; SCG, scent gland). Numbers above branches correspond to Bootstrap values for 500 replicates.

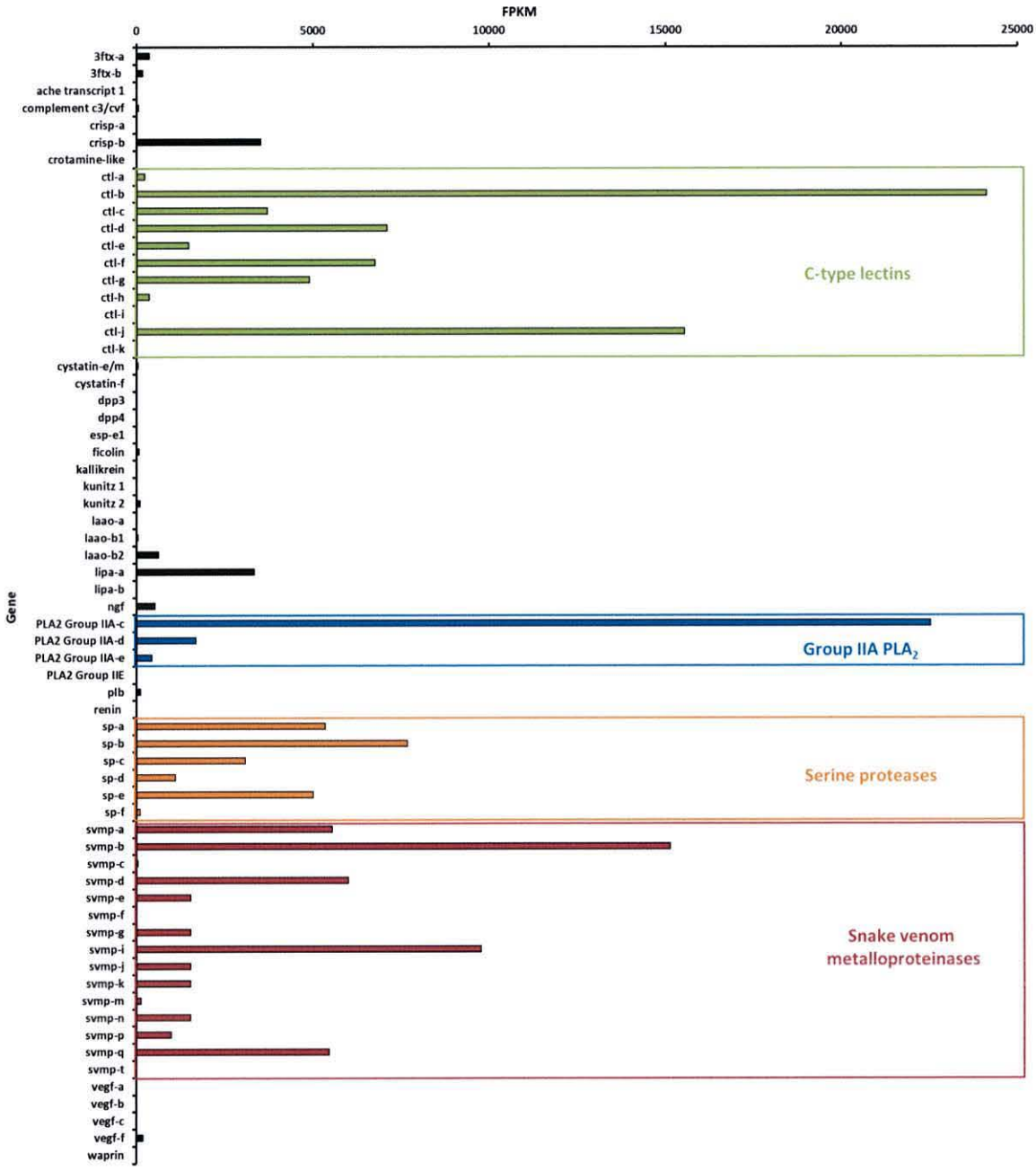


Figure 4.32. The majority of Toxicoferan transcripts are expressed at extremely low level, with the most highly expressed genes falling into only four gene families (C-type lectins, Group IIA phospholipase A_2, serine proteases and snake venom metalloproteinases). FPKM = Fragments Per Kilobase of exon per Million fragments mapped.

4.3.4 Evidence of misidentification and low sequence variation in Toxicoferan sequences

Whilst conducting phylogenetic analyses it became evident that some publicly available Toxicoferan toxin sequences had been annotated as different genes when their sequences were

actually 100% identical (Figure 4.33 and 4.34). In other cases, sequences had been annotated as different genes (not alleles of the same gene) despite being almost identical at the nucleotide level (Figures 4.35-4.38). As a result it appears that the number of reptile Toxicoferan genes has been incorrectly inflated.

```
                                   10        20        30        40        50        60        70
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       ATGGCATCTT TCATAGCTTT CCTTGTGTAC AGCTGGGTTC TCCTATCGCT TGCAGCAGGT CAAGTTACCT
EU195462.1 MMP2-Var2       .......... .......... .......... .......... .......... .......... ..........

                                   80        90       100       110       120       130       140
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       ATCCAGAATC TCCAGCCTGT AAGTTTCCCT TCATTTATGA GGGCAAAGCC TTCACTACAT GCACGGAATA
EU195462.1 MMP2-Var2       .......... .......... .......... .......... .......... .......... ..........

                                  150       160       170       180       190       200       210
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       TGGTTCTACT GACAAAACTC CATGGTGTGC CACAACCTCA AACTATGACA GGGATCGTAA ATGGAAGCCG
EU195462.1 MMP2-Var2       .......... .......... .......... .......... .......... .......... ..........

                                  220       230       240       250       260       270       280
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       TGTGCTGTCA AAGAGTATGG AGAACTGAGT GTCAACTACC CTGAAGATCC ATGTAGATTT CCTTTCACCT
EU195462.1 MMP2-Var2       .......... .......... .......... .......... .......... .......... ..........

                                  290       300       310       320       330       340       350
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       ACAAAGGCAA GACTTACTCT GGTTGTACAG GAGCTGGGAG AGAGGATGGG AAGCTTTGGT GCTCTATCAG
EU195462.1 MMP2-Var2       .......... .......... .......... .......... .......... .......... ..........

                                  360       370       380       390       400       410       420
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       CAAAAATCAT GATGACAACT CACAATTGGT ATTCTGTGAG CCTTCAGATC CAGCCCCCTG CTACTTTCCT
EU195462.1 MMP2-Var2       .......... .......... .......... .......... .......... .......... ..........

                                  430       440       450       460       470       480       490
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       TTCAAATACA AGCAAAAATC CTACTCTGAC TGCACCATGA ATGGGAGTTT TGATGGGCAT TGGTGGTGTG
EU195462.1 MMP2-Var2       .......... .......... .......... .......... .......... .......... ..........

                                  500       510       520       530       540       550       560
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       CCACAACCGC AGATTATGAC AAGGACAGCA AGTGGAAGGC ATGTATTACT GAAGAATATG GAGGAAACTC
EU195462.1 MMP2-Var2       .......... .......... .......... .......... .......... .......... ..........

                                  570       580       590       600       610       620       630
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       GAATGGTGGG CAGTGTGTCT TCCCCTTCAC CTACATGGAC AAGGAATATA CTACATGCAC CAATGAGGAT
EU195462.1 MMP2-Var2       .......... .......... .......... .......... .......... .....

                                  640       650       660       670       680       690       700
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       ACAGATGGGA GATTCTGGTG TGCCACTACG GCAAACTATG ACAAGGACAA GAAATGGAGC TTCTGTGCTG
EU195462.1 MMP2-Var2

                                  710       720       730       740       750       760       770
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       ACACAAGAAT GAGCAACGGG GAGATTGTCA AGCCAGGAAA TGTCGGCAGT TCTAAGGCGC CATGTTCCTT
EU195462.1 MMP2-Var2

                                  780       790       800       810       820       830       840
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       CCCCTTCATC TACAAAGGCA AGACTTACAC AGAGTGTACT TCCAAGGGGA CTTTAACATG GAAGCTCTGG
EU195462.1 MMP2-Var2

                                  850       860       870       880       890       900       910
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441521.1 ESP-Vkom1       TGCTCTCTCA CCAGCAACTA TGATAGAGAT AAGAGATGGA GATACTGTCA GCCTTCAGAT TTAGAAACAC
EU195462.1 MMP2-Var2

                                  920       930       940       950       960
                           ....|....| ....|....| ....|....| ....|....| ....|....| ....|.
GU441521.1 ESP-Vkom1       AAAATCAGTT GCAGTCTAAT GAAGAGAAGC AAACAATAAA TATCAGCAAG GAATAA
EU195462.1 MMP2-Var2
```

Figure 4.33. Alignments of *Varanus komodoensis* epididymal secretory protein (GU441521, "ESP-Vkom1") and matrix metalloproteinase (EU195462, MMP2-Var2) sequences (Fry et al. 2010a) showing 100% similarity at the nucleotide level. Sequence lengths reflect those on Genbank and have not been edited.

```
                                 10        20        30        40        50        60        70
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441522.1 ESP-Vkom2     ATGGCATCTT TCATAGCTTT CCTTGTGTAC AGCTGGGTTC TCCTATCGCT TGCAGCAGGT CAAGTTACCT
EU195454.1 MMP2 Var1     .......... .......... .......... .......... .......... .......... ..........

                                 80        90       100       110       120       130       140
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441522.1 ESP-Vkom2     ATCCAGAATC TCCAGCCTGT AAGTTTCCCT TCATTTATGA GGGCAAAGCC TTCACTACAT GCACGGAATA
EU195454.1 MMP2 Var1     .......... .......... .......... .......... .......... .......... ..........

                                150       160       170       180       190       200       210
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441522.1 ESP-Vkom2     TGGTTCTACT GACAAAACTC CATGGTGTGC CACAACCTCA AACTATGACG GGGATCGTAA ATGGAAGCCG
EU195454.1 MMP2 Var1     .......... .......... .......... .......... .......... .......... ..........

                                220       230       240       250       260       270       280
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441522.1 ESP-Vkom2     TGTGCTGTCA AAGAGTATGG AGGTAATTCC AATGGAGCGC CATGCACCTT TCCCTTCATT TACCTTGGTC
EU195454.1 MMP2 Var1     .......... .......... .......... .......... .......... .......... ..........

                                290       300       310       320       330       340       350
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441522.1 ESP-Vkom2     GTACTTACTA CACCTGCACC AACAAACTTG AGCATAAAGG ACGGTACTGG TGCGCCACAA CAGGGAGCTA
EU195454.1 MMP2 Var1     .......... .......... .......... .......... .......... .......... ..........

                                360       370       380       390       400       410       420
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441522.1 ESP-Vkom2     TGATAAGGAC CGGAAGTGGA GTTTCTGTGC AGACATCAAA CTGAGTGTCA ACTACCCTGA AGATCCATGT
EU195454.1 MMP2 Var1     .......... .......... .......... .......... .......... .......... ..........

                                430       440       450       460       470       480       490
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441522.1 ESP-Vkom2     AGATTTCCTT TCACCTACAA AGGCAAGACT TACTCTGGTT GTACAGGAGC TGGGAGAGAG GATGGGAAGC
EU195454.1 MMP2 Var1     .......... .......... .......... .......... .......... .......... ..........

                                500       510       520       530       540       550       560
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441522.1 ESP-Vkom2     TTTGGTGCTC TATCAGCAAA AATCATGATG ACAACTCACA ATTGGTATTC TGTGAGCCTT CAGATCCAGC
EU195454.1 MMP2 Var1     .......... .......... .......... .......... .......... .......... ..........

                                570       580       590       600       610       620       630
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441522.1 ESP-Vkom2     CCCCTGCTAC TTTCCTTTCA AATACAAGCA AAAATCCTAC TCTGACTGCA CCATGAATGG GAGTTTTGAT
EU195454.1 MMP2 Var1     .......... .......... .......... .......... .......... .......... ..........

                                640       650       660       670       680       690
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441522.1 ESP-Vkom2     GGGCATTTGT GGTGTGCCAC AACCGCAGAT TATGACAAGG ACAGCAAGTG GAAGGCATGT
EU195454.1 MMP2 Var1     .......... .......... .T........ .......... .....
```

Figure 4.34. Alignments of *Varanus komodoensis* epididymal secretory protein (GU441522, "ESP-Vkom2") and matrix metalloproteinase (EU195454, MMP2 Var1) nucleotide sequences (Fry et al. 2010a) showing near total sequence identity (1bp difference over 675bp of aligned sequence). Sequence lengths reflect those on Genbank and have not been edited.

```
                                  10        20        30        40        50        60        70
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441513.1| RNase-Ginf2   ATGGCTGCAA AGGGATCCTG TGTGTTCTGG TTCCTTTTGG CATCCGCCTG GATAGTGATG AAGAGTGATG
GU441515.1| RNase-Ginf4   .......... .......... .......... .......... .......... .......... ..........

                                  80        90       100       110       120       130       140
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441513.1| RNase-Ginf2   CAGCTGACAC CTTTGAATCC TTCAAAGAGC TGCACGTTGA TTATCCAAAG ACTGGAGCTC CAAATGATGA
GU441515.1| RNase-Ginf4   .......... .......... .......... .......... .......... .......... ..........

                                 150       160       170       180       190       200       210
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441513.1| RNase-Ginf2   TGAATACTGC AAAAAAGTTA TGGGGGGAAG AGGCCGGACC AAACTCAAGG CCAATACTTA CATTCATGCC
GU441515.1| RNase-Ginf4   .......... .......... .......... .......... .......... .......... ..........

                                 220       230       240       250       260       270       280
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441513.1| RNase-Ginf2   CCAGACAGTG AACTTTTAGC CGCCTGCAAT CGGAAGAAGT ATAAACTTAA TCATGAATAC GGCAGGACCT
GU441515.1| RNase-Ginf4   .......... .......... .......... .......... .......... .......... ..........

                                 290       300       310       320       330       340       350
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441513.1| RNase-Ginf2   CTCGCCTTCC AACCACTTTG TGTACGTATG GTGATAGAGT ATTCTTGGGA AGCTCCTTGC CAGGAACGAA
GU441515.1| RNase-Ginf4   .......... .......... .......... .......... .......... ...------- ----------

                                 360       370       380       390       400
                          ....|....| ....|....| ....|....| ....|....| ....|....|
GU441513.1| RNase-Ginf2   TAAGGTTCTG TGTGTGAATT GGAAGCCAGT TGCATTTAGA GGATTCAACG CCTAA
GU441515.1| RNase-Ginf4   ---------- ---------- ---------- ---------- ---------- -----
```

Figure 4.35. Alignments of *Gerrhonotus infernalis* ribonuclease sequences (Fry et al. 2010a) GU441513 ("Ginf2") and GU441515 ("Ginf4") showing 100% similarity at the nucleotide level. Sequence lengths reflect those on Genbank and have not been edited.

```
                                  10        20        30        40        50        60        70
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441517.1 RNase-Ginf6   ATGGCTGCAA AGGGATCCTG TGTGTTCTGG TTCCTTTTGG CATCCGCCTG GATAGTGATG AAGAGTGATG
GU441516.1 RNase-Ginf5   .......... .......... .......... .......... .......... .......... ..........

                                  80        90       100       110       120       130       140
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441517.1 RNase-Ginf6   CAGCTGACAC CTTTGAATCC TTCAAAGAGC TGCACGTTGA TTATCCAAAG ACTGAAGCTC CAAATGATAA
GU441516.1 RNase-Ginf5   .......... .......... .......... .......... .......... .......... ..........

                                 150       160       170       180       190       200       210
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441517.1 RNase-Ginf6   TGAATACTGC AAAAAAGTTA TGGGGGGAAG AGGCCAGACC AAACTCAAGG CCAATACTTA CATTCATGCC
GU441516.1 RNase-Ginf5   .......... .......... .......... .......... .......... .......... ..........

                                 220       230       240       250       260       270       280
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441517.1 RNase-Ginf6   CCAGACAGTG AACTTTTAGC CGCCTGCAAT CGGAAGAAGT ATAAACTTAA TCATGAATAC GGCAGGACCT
GU441516.1 RNase-Ginf5   .......... .......... .......... .......... .......... .......... ..........

                                 290       300       310       320       330       340       350
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
GU441517.1 RNase-Ginf6   CTCGCCTTCC AACCACTTTG TATGGCTGCA AAGGGATCCT GTGTGTTCTG GTTCCTTTTG GCATCCGCCT
GU441516.1 RNase-Ginf5   .......... .......... .......... .......... .......... .......... ..........

                                 360
                          ....|....| ....|....
GU441517.1 RNase-Ginf6   GCATAGTGAT GAAGAGTGA
GU441516.1 RNase-Ginf5   .......... .........
```

Figure 4.36. Alignments of *Gerrhonotus infernalis* ribonuclease sequences (Fry et al. 2010a) GU441516 ("Ginf5") and GU441517 ("Ginf6") showing 100% similarity at the nucleotide level. Sequence lengths reflect those on Genbank and have not been edited.

```
                                  10        20        30        40        50        60        70
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139885.1 CRISP-VAR5     ATGATCCTGC TCAAACTGTA TTTGACCCTA GCTGCAATCT TATGTCAATC CCGTGGCACG ACTTCTCTTG
DQ139884.1 CRISP-VAR4     .......... .......... .......... .......... .......... .......... ..........
DQ139886.1 CRISP-VAR6     .......... .......... .......... .......... .......... .......... ..........

                                  80        90       100       110       120       130       140
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139885.1 CRISP-VAR5     ATCTTGATGA TTTGATGACT ACCAACCCTG AGATACAAAA TGAGATCATT AACAAGCACA ATGACCTACG
DQ139884.1 CRISP-VAR4     .......... .......... .......... .......... .......... .......... ..........
DQ139886.1 CRISP-VAR6     .......... .......... .......... .......... .......... .......... ..........

                                 150       160       170       180       190       200       210
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139885.1 CRISP-VAR5     GAGAACAGTG GATCCCCCAG CTAAAAACAT GCTGAAGATG TCCTGGGACA ACATCATTGC AGAGAGTGCC
DQ139884.1 CRISP-VAR4     .......... .......... .......... .......... .......... .......... ..........
DQ139886.1 CRISP-VAR6     .......... .......... .......... .......... .......... .......... ..........

                                 220       230       240       250       260       270       280
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139885.1 CRISP-VAR5     AAACGTGCAG CACTGAGATG CAACCAAAAT GAGCACACAC CTGTCTCGGG AAGAACAATA GGTGGTGTGG
DQ139884.1 CRISP-VAR4     .......... .......... .......... .......... .......... .......... ..........
DQ139886.1 CRISP-VAR6     .......... .......... .......... .......... .......... .......... ..........

                                 290       300       310       320       330       340       350
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139885.1 CRISP-VAR5     TGTGCGGAGA AAATTACTTC ATGTCAAGTA ACCCTCGCAC ATGGTCTTTC GGCATTCAGA GTTGGTTTGA
DQ139884.1 CRISP-VAR4     .......... .......... .......... .......... .......... .......... ..........
DQ139886.1 CRISP-VAR6     .......... .......... .......... .......... .......... .......... ..........

                                 360       370       380       390       400       410       420
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139885.1 CRISP-VAR5     TGAAAGGAAC TACTTTAAGT TTGGTTTTGG ACCAACAAGA GCAGGTGTCA TGGTTGGCCA TTATACCCAG
DQ139884.1 CRISP-VAR4     .......... .......... .......... .......... .......... .......... ..........
DQ139886.1 CRISP-VAR6     .......... .......... .......... .......... .......... .......... ..........

                                 430       440       450       460       470       480       490
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139885.1 CRISP-VAR5     GTGGTCTGGT ATAAGTCTTA CAAAATGGGA TGTGCGATCA ACTTGTGCCC TAATGAGCCC CTGAAGTACT
DQ139884.1 CRISP-VAR4     .......... .......... .......... .......... .......... .......... ..........
DQ139886.1 CRISP-VAR6     AA..AAC.T. G..G.C.GG. AGT.

                                 500       510       520       530       540       550       560
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139885.1 CRISP-VAR5     TCCTGGTTTG CCAGTACTGC CCAGGAGGGA ACGTTGTAGG CCGGAAGTAT GAACCCTATG CAATCGGAGA
DQ139884.1 CRISP-VAR4     .......... ..........
DQ139886.1 CRISP-VAR6

                                 570       580       590       600       610       620       630
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139885.1 CRISP-VAR5     ACCATGTGCA GCTTGCCCCA ACAATTGTGA CAACGGACTG TGCACTAACC CCTGTGAGCA CAGCAATCAA
DQ139884.1 CRISP-VAR4
DQ139886.1 CRISP-VAR6

                                 640       650       660
                          ....|....| ....|....| ....|....|
DQ139885.1 CRISP-VAR5     TACATCAACT GCCCAGATTT AACAAAACAG
DQ139884.1 CRISP-VAR4
DQ139886.1 CRISP-VAR6
```

Figure 4.37. Nucleotide alignments of *Varanus acanthurus* CRISP sequences (Fry et al. 2006) DQ139885 ("CRISP-VAR5"), DQ139884 ("CRISP-VAR4") and DQ139886 ("CRISP-VAR6") showing almost total similarity at the nucleotide level. Sequence lengths reflect those on Genbank and have not been edited.

```
                                 10        20        30        40        50        60        70
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139888.1 CRISP-VAR8    ATGATCCTGC TCAAACTGTA TTTGACCCTA GCTGCAATCT TATGTCAATC CCGTGGCACG ACTTCTCTTG
DQ139887.1 CRISP-VAR7    .......... .......... .......... .......... .......... .......... ..........
DQ139883.1 CRISP-VAR3    .......... .......... .......... .......... .......... .......... ..........

                                 80        90       100       110       120       130       140
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139888.1 CRISP-VAR8    ATCTTGATGA TTTGATGACT ACCAACCCTG AGATACAAAA TGAGATCATT AACAAGCACA ATGACCTACG
DQ139887.1 CRISP-VAR7    .......... .......... .......... .......... .......... .......... ..........
DQ139883.1 CRISP-VAR3    .......... .......... .......... .......... .......... .......... ..........

                                150       160       170       180       190       200       210
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139888.1 CRISP-VAR8    GAGAACGGTG GATCCCCCAG CTAAAAACAT GCTGAAGATG TCCTGGGACA ACACCATTGC AGAGAGTGCC
DQ139887.1 CRISP-VAR7    .......... .......... .......... .......... .......... .......... ..........
DQ139883.1 CRISP-VAR3    .......... .......... .........C .......... .......... .......... ..........

                                220       230       240       250       260       270       280
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139888.1 CRISP-VAR8    AAACGTGCAG CACTGAGATG CAACCAAAAT GAGCACACAC CTGTCTCGGG AAGAACAATA GGTGGTGTGG
DQ139887.1 CRISP-VAR7    .......... .......... .......... .......... .......... .......... ..........
DQ139883.1 CRISP-VAR3    .......... .......... .......... .......... .......... .......... ..........

                                290       300       310       320       330       340       350
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139888.1 CRISP-VAR8    TGTGCGGAG- AAAATTACT TCATGTCAAG TAACCTTCGC ACATGGTCTT TCGGCATTCA GAGTTGGTTT
DQ139887.1 CRISP-VAR7    .......... .......... .......... .......... .......... .......... ..........
DQ139883.1 CRISP-VAR3    .......... .......... .......... .......... .......... .......... ..........

                                360       370       380       390       400       410       420
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139888.1 CRISP-VAR8    GATGAAAGGA ACTACTTTAA GTTTGGTTTT GGACCAACAA GAGCAGGTGT CATGGTTGGC CATTATACCC
DQ139887.1 CRISP-VAR7    .......... .......... .......... .......... .......... .......... ..........
DQ139883.1 CRISP-VAR3    .......... .......... .......... .......... .......... .......... ..........

                                430       440       450       460       470       480       490
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139888.1 CRISP-VAR8    AGGTGGTCTG GTATAAGTCT TACAAAATGG GATGTGCGAT CAACTTGTGC CCTAATGAGC CCCTGAAGTA
DQ139887.1 CRISP-VAR7    .......... .......... .......... .......... .......... ......---- ----------
DQ139883.1 CRISP-VAR3    .......... .......... .......... .......... .......... .......... ..........

                                500       510       520       530       540       550       560
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
DQ139888.1 CRISP-VAR8    CTTCCTGGTT TGCCAGTACT GCCCAGGAGG GAACGTTGTA GGCCGGAAGT ATGAACCCTA TGCAATCGGA
DQ139887.1 CRISP-VAR7    ---------- ---------- ---------- ---------- ---------- ---------- ----------
DQ139883.1 CRISP-VAR3    .......... .......... .......... .......... ......---- ---------- ----------

                                570       580       590       600
                         ....|....| ....|....| ....|....| ....|....| ..
DQ139888.1 CRISP-VAR8    GAACCATGTG CAGCTTGCCC CAACAACTGT GACAATGGAC TG
DQ139887.1 CRISP-VAR7    ---------- ---------- ---------- ---------- --
DQ139883.1 CRISP-VAR3    ---------- ---------- ---------- ---------- --
```

Figure 4.38. Nucleotide alignments of *Varanus acanthurus* CRISP sequences (Fry et al. 2006) DQ139888 ("CRISP-VAR8"), DQ139887 ("CRISP-VAR7") and DQ139883 ("CRISP-VAR3") showing almost total similarity at the nucleotide level. Sequence lengths reflect those on Genbank and have not been edited.

4.4 Discussion

All 16 of the basal venom toxin genes used to support the hypothesis of a single, early evolution of venom in reptiles (the Toxicofera hypothesis (Vidal and Hedges 2005; Fry et al. 2006; Fry et al. 2009a; Fry et al. 2009b; Fry et al. 2010; Fry et al. 2012b; Fry et al. 2013)), as well as a number of other genes that have been proposed to encode venom toxins in multiple species are in fact expressed in multiple tissues, with no evidence for consistently higher expression in venom or salivary glands. Additionally, only two genes in the entire dataset of 74 genes in five species were found to encode possible venom gland-specific splice variants (*l-amino acid oxidase b2* and *PLA_2 IIA-c*). Therefore many of the proposed basal Toxicoferan genes most likely represent housekeeping or maintenance genes and the identification of these genes as conserved venom toxins is due to incomplete tissue sampling. This lack of support for the Toxicofera hypothesis therefore prompts a return to the previously held view (Kardong et al. 2009) that venom in different lineages of reptiles has evolved independently, once at the base of the advanced snakes, once in the helodermatid (gila monster and beaded lizard) lineage and, possibly, one other time in monitor lizards, although evidence for a venom system in this latter group (Fry et al. 2009b; Fry et al. 2010; Vikrant and Verma 2013) may need to be reinvestigated in light of our findings. The process of reverse recruitment (Casewell et al. 2012), where a venom gene undergoes additional gene duplication events and is subsequently recruited from the venom gland back into a body tissue (which was proposed on the basis of the placement of garter snake and Burmese python "physiological" genes within clades of "venom" genes) must also be re-evaluated in light of our findings (see next chapter).

Since antivenoms are derived from the injection of crude venom into a host animal, they are not targeted to the most pathogenic venom components and therefore also include antibodies to weakly- or non-pathogenic proteins requiring the administration of large or multiple doses (Casewell et al. 2013), increasing the risks of adverse reactions. A comprehensive understanding of snake venom composition is therefore vital for the development of the next generation of antivenoms (Harrison 2004; Wagstaff et al. 2006; Casewell et al. 2013) as it is important that research effort is not spread too thinly through the inclusion of non-toxic venom gland transcripts.

4.4.1 Implications for snake venom complexity and evaluation of methodology used

Results suggest that erroneous assumptions about the single origination and functional conservation of venom toxins across the Toxicofera has led to the complexity of snake venom being overestimated by previous authors. The ruling out of the majority of proposed Toxicoferan toxins from the venom repertoire has a substantial effect on the number of genes and gene families which contribute to reptile venom composition. Based on the findings of this study the venom of the painted saw-scaled viper, *Echis coloratus,* is likely to consist of just 34 genes in 8 gene families (Table 4.1, based on venom gland-specific expression and a 'high' expression, as defined by presence in the top 25% of transcripts (Williford and Demuth 2012) in at least two of four venom gland samples), fewer than has been suggested for this and related species in previous EST or transcriptomic studies (Wagstaff and Harrison 2006; Casewell et al. 2009).

Table 4.1. Predicted venom composition based on the results of this study of the painted saw-scaled viper, *Echis coloratus*

Gene family	Number of genes
SVMP	13
C-type lectin	8
Serine protease	6
PLA_2	3
CRISP	1
L-amino acid oxidase	1
VEGF	1
Crotamine	1
Total 8	34

It is noteworthy that the results of these analyses accord well with proteomic analyses of venom composition in snakes, with an almost identical complement of 35 toxins in 8 gene families known from the related ocellated carpet viper, *Echis ocellatus* (Wagstaff et al. 2009), where SVMPs, CTLs and PLA_2s were found to be the most abundant proteins. However, on closer inspection this study considered DC-fragments (disintegrin/cysteine-rich fragment) and disintegrins to be separate gene families, whilst they are in fact (by the authors own admission) cleaved fragments derived from a snake venom metalloproteinase (Kini and Evans 1992; Calvete et al. 2003; Wagstaff et al. 2009b). The venom proteome of *E. ocellatus* can therefore be considered to be composed of 35 toxins belonging to 6 gene families (26 SVMPs, 4 PLA_2s,

2 C-type lectins, 1 CRISP, 1 L-amino acid oxidase and 1 Serine protease) (Wagstaff et al. 2009b).

Proteomic studies of a range of other venomous snake species have identified a typical complement of between 24-61 toxins in 6-14 families (Table 4.2). Far from being a "complex cocktail" (Izidoro et al. 2006; Calvete et al. 2007a; Wong and Belov 2012; Casewell et al. 2013), snake venom may in fact represent a relatively simple mixture of toxic proteins honed by natural selection for rapid prey immobilisation, with limited lineage-specific expansion in one or a few particular gene families.

It seems likely that the application of this approach to other species (together with proteomic studies of extracted venom) will lead to a commensurate reduction in claimed venom diversity, with clear implications for the development of next generation antivenoms: since most true venom genes are members of a relatively small number of gene families, it is likely that a similarly small number of antibodies may be able to bind to and neutralise the toxic venom components, especially with the application of "string of beads" techniques (Whitton et al. 1993) utilising fusions of short oligopeptide epitopes designed to maximise the cross-reactivity of the resulting antibodies (Wagstaff et al. 2006).

Table 4.2. Predicted numbers of venom toxins and venom toxin families from proteomic studies of snake venom accord well with transcriptomic results.

Species	Reference	Number of toxins	Number of toxin families
Bitis caudalis	(Calvete et al. 2007b)	30	8
Bitis gabonica gabonica	(Calvete et al. 2006)	35	12
Bitis gabonica rhinoceros	(Calvete et al. 2007b)	33	11
Bitis nasicornis	(Calvete et al. 2007b)	28	9
Bothriechis schlegelii	(Lomonte et al. 2008)	?	7
Cerastes cerastes	(Fahmi et al. 2012)	25-30	6
Crotalus atrox	(Calvete et al. 2009)	~24	~9
Echis ocellatus	(Wagstaff et al. 2009a)	35	8
Lachesis muta	(Sanz et al. 2008)	24-26	8
Naja kaouthia	(Kulkeaw et al. 2007)	61	12
Ophiophagus hannah	(Vonk et al. 2013)	?	14
Vipera ammodytes	(Georgieva et al. 2008)	38	9

4.4.2 Suggestions for future studies

In order to avoid the continued overestimation of the number of venom genes in reptiles, many of which have been used in the support and propagation of the Toxicofera hypothesis, several suggestions are made and discussed below.

In order to avoid continued over-inflation of venom complexity, future transcriptome-based analyses of venom composition must include quantitative comparisons of multiple body tissues from multiple individuals and robust phylogenetic analysis that includes known paralogous members of gene families. In this way the true expression pattern of a gene or transcript may be inferred, any occurrences of pleiotropy may be detected, and correct orthology can be assigned to genes based on phylogenetic analysis rather than BLAST-based methods. Transcriptomic analysis of solely venom gland (as is a frequent methodology) is perfectly acceptable for descriptive studies which seek to characterise the transcriptome of this tissue. However, in order to assign a potential toxic role to a gene (and especially to infer its true evolutionary history, or the evolution of the venom repertoire in an entire lineage), sequencing the venom gland alone is insufficient as no "housekeeping" tissue is present as a reference.

The use of clearly explained, justifiable criteria for classifying highly similar sequences as new paralogs rather than alleles or the result of PCR or sequencing errors is also essential, as it seems likely that some available sequences from previous studies have been presented as distinct genes on the basis of extremely minor (or even non-existent) sequence variation (see figures 4.33 and 4.34 for examples of the same sequence being annotated as two different genes and figures 4.35-4.38 for examples of identical or nearly identical ribonuclease and CRISP sequences). As a result, the diversity of "venom" composition in these species may have been inadvertently inflated.

Finally, inconsistent nomenclature of snake venom genes has contributed to a prolonged incidence of reptile salivary proteins being claimed as homologs of characterised venom toxins. Whilst a standardised nomenclatural system is established for human (Shows et al. 1979), mouse (Eppig 2006) and zebrafish (Mullins 1995) genes, and such a system has been suggested for spiders (King et al. 2008), scorpions (Tytgat et al. 1999), sea anemones (Oliveira et al. 2012) and centipedes (Undheim et al., 2014), there is to date no nomenclatural system in place for reptile toxins. As a result the orthology, paralogy and evolutionary history of snake venom genes is not always easily apparent. For example, *ophanin* (accession AY181984, (Yamazaki et al. 2003a)) and *opharin* (accession AY299475, Direct submission) are both cysteine-rich secretory proteins (CRISPs), whilst *ohanin* (accession DQ103590, (Pung et al. 2005)) is a vespryn-like gene. As all three are similar in pronunciation and spelling, and are all derived

from the king cobra (*Ophiophagus hannah*), it is obvious that their gene names could lead to confusion as they do not provide useful information about the proteins they encode nor their evolutionary origin. Therefore, the adoption of a standard nomenclature for reptile genes is encouraged (Hargreaves and Mulley 2014), as the overly-complicated and confusing nomenclature used currently (Table 4.3) may also contribute to the perceived complexity of snake venom. Rather than develop an entirely novel nomenclatural system for reptile toxins, it is more logical that the adopted nomenclature system should be based on the comprehensive standards developed for anole lizards (Kusumi et al. 2011), for example:

- "Gene symbols for all…species should be written in lower case only and in italics, e.g., *gene2*."
- "Whenever criteria for orthology have been met… the gene symbol should be comparable to the human gene symbol, e.g., if the human gene symbol is *GENE2*, then the gene symbol would be *gene2*."
- "Duplication of the ortholog of a mammalian gene will be indicated by an "a" or "b" suffix, e.g., *gene2a* and *gene2b*. If the mammalian gene symbol already contains a suffix letter, then there would be a second letter added, e.g., *gene4aa* and *gene4ab*."

In addition, where toxin sequences are derived from proteomic analyses without a corresponding gene sequence they should be named based on similarity to existing toxin sequences in public databases (e.g. Swiss-Prot (Bairoch and Boeckmann 1991)) with the suffix "-like" to acknowledge sequence similarity but also to identify that the protein is currently uncharacterised. However, the increase in genomic and transcriptomic data available for reptile species should facilitate the identification of toxins derived from proteomic studies.

Table 4.3. Venom gene nomenclature. Lack of a formal set of nomenclatural rules for venom toxins has led to an explosion of different gene names and may have contributed to the overestimation of reptile venom diversity.

Gene/gene family	Alternative name and accession number
CRISP	Piscivorin [AAO62994] Ablomin [AAM45664] Tigrin [Q8JGT9] Kaouthin [ACH73167, ACH73168] Natrin-1 [Q7T1K6] Pseudechetoxin [Q8AVA4] Pseudechin [Q8AVA3] Serotriflin [P0CB15] Latisemin [Q8JI38] Ophanin [AAO62996] Opharin [ACN93671]
Serine proteases	Acubin [CAB46431] Gyroxin [B0FXM3] Ussurase [AAL48222] Serpentokallikrein [AAG27254] Salmobin [AAC61838] Batroxobin [AAA48553] Gloshedobin [POC5B4] Gussurobin [Q8UVX1] Pallabin [CAA04612] Pallase [AAC34898]
Snake venom metalloproteinase (SVMP)	Stejnihagin-B [ABA40759] Bothropasin [AAC61986] Atrase B [ADG02948] Mocarhagin 1 [AAM51550] Scutatease-1 [ABQ01138] Austrelease-1 [ABQ01134]
Vascular endothelial growth factor (VEGF)	Barietin [ACN22038] Cratrin [ACN22040] Apiscin [ACN22039] Vammin [ACN22045]
Vespryn	Ohanin [AAR07992] Thaicobrin [P82885]
Waprin	Nawaprin [P60589] Porwaprin [B5L5N2] Stewaprin [B5G6H3] Veswaprin [B5L5P5] Notewaprin [B5G6H5] Carwaprin [B5L5P0]

4.4.3 Difficulty in assigning justifiable expression level cut-offs

One of the specific problems encountered when analysing the transcript abundance estimation data was determining an appropriate and justifiable criteria for assigning an expression level to transcripts (e.g. highly or lowly expressed). Normally transcripts with an FPKM value of less than 1.0 are classed as being not expressed, which would apply to a considerable number of transcripts in this study (Figure 4.39). The datasets (particularly for venom gland) were also highly skewed, with a small number of highly expressed transcripts (i.e. large FPKM values), but a large number of lowly expressed transcripts. As a result any cut-off values assigned were purely arbitrary with no statistical justification.

Values below 1.0 FPKM were not discarded as this would have eliminated a considerable number of proposed Toxicoferan gene transcripts from the analysis. As all transcriptomes were trimmed to sequences 300bp in length or above, transcripts must have been assembled from several reads, and would not be representative of unplaced read pairs. A cut-off value of above 1,000 FPKM was initially chosen to represent highly expressed transcripts, with the next Toxicoferan gene below this value being ~600 FPKM. As can be seen in Figure 4.39 this resulted in only 16 transcripts belonging to 5 gene families being classed as encoding venom toxins. Whilst this may be the correct number of venom genes in *E. coloratus*, there is no statistically justifiable reason for this cut-off, and the results are also lower than the number of toxins found in the venom proteome of *E. ocellatus*.

An alternative strategy was to use the expression of housekeeping genes as a base level of transcription and use this to determine justifiable cut-off values. However, as the expression of housekeeping genes was highly variable (see Chapter 3) it was not possible to establish a base level. The sequencing depth between samples (particularly between venom gland samples) was also highly variable, which must be considered when conducting future experiments. A high amount of variation in the number of reads per sample may lead to biases in mapping and transcript abundance estimation, even despite FPKM values being normalised for transcript length and sequencing depth. A future strategy could be to use the transcriptome sequences (along with genomic sequences to infer exon/intron boundaries) to design primers for qPCR experiments, where the expression level of a specific gene can be assessed across a larger number of samples and compared to several reference genes.

	Ema			Pre			Eco			Pgu			Oae		
	SAL	SCG	SK	SAL	SCG	SK	VG	SCG	SK	SAL	SCG	SK	SAL	SCG	SK
Venom															
crisp-b	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-
c-type lectins b, c, d, f, g, j	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
Group IIA PLA_2 c	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
serine proteases a, b, e	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
svmps a, b, d, l, q	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
Not venom															
3 finger toxin a	-	-	-	+	-	-	+	+	+	+	+	+	+	-	+
3 finger toxin b	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-
ache - transcript 1	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-
ache - transcript 2	-	-	-	-	+	-	-	+	+	-	-	-	-	+	-
complement c3/cvf	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+
crisp-a	-	-	-	-	+	-	+	+	+	+	+	+	-	+	+
crotamine-like	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
c-type lectin a	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
cystatin-e/m	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
cystatin-f	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+
dipeptidyl peptidase 3	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
dipeptidyl peptidase 4	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
esp-e1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ficolin	+	+	+	+	-	-	+	+	+	+	+	+	+	+	-
kallikrein	+	+	-	+	+	+	+	+	+	-	+	-	-	+	-
kunitz 1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
kunitz 2	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
l-amino acid oxidase a	+	+	+	-	-	-	+	+	+	-	-	-	-	-	-
l-amino acid oxidase b1	-	-	-	-	+	-	+	-	-	-	+	-	-	+	-
l-amino acid oxidase b2	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
lysosomal acid lipase a	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
lysosomal acid lipase b	-	-	-	-	+	-	+	+	-	-	-	-	-	-	-
nerve growth factor	+	+	+	-	-	+	+	+	+	+	+	-	-	+	+
Group IIE PLA_2	+	-	-	+	-	-	+	-	-	+	-	-	+	-	-
phospholipase b	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
renin	-	+	-	-	-	-	+	+	-	-	-	-	-	-	-
vegf-a	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
vegf-b	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
vegf-c	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
vegf-f	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
vespryn	-	+	+	-	+	+	-	-	-	+	+	-	-	+	+
waprin	-	-	-	-	-	-	+	-	+	+	-	-	+	+	+

Figure 4.39. Gene expression levels are represented by colour shading with "High" (>1000 FPKM) shaded red, medium (100-1000 FPKM) shaded yellow, low (1-100 FPKM) shaded green, and transcripts which would usually be considered to be not expressed (<1 FPKM) shaded white. The absence of expression of a transcript is indicated with a "-" and is shaded grey.

4.4.4 Conclusions

The identification of the apparently conserved Toxicofera venom toxins in previous studies is most likely a side effect of incomplete tissue sampling, compounded by incorrect interpretation of phylogenetic trees and the use of BLAST-based gene identification methods. It should perhaps not be too surprising that homologous tissues in related species would show similar gene complements and the restriction of most previous studies to only the "venom" glands means that monophyletic clades of reptile sequences in phylogenetic trees have been taken to represent monophyletic clades of venom toxin genes. Whilst it is true that some of these genes do encode toxic proteins in some species (indeed, this was often the basis for their initial discovery) the discovery of orthologous genes in other species does not necessarily demonstrate shared toxicity. The complexity of reptile venoms appears to have been greatly exaggerated, which has implications both for our understanding of how venom gene families evolve, and for the development of more refined and targeted antivenom treatments for snakebite. Whilst this study does not constitute grounds to fully refute the Toxicofera hypothesis, it is certainly enough to call the widespread acceptance of it into question, and will hopefully represent a point of reference for future research into elucidating the true evolutionary history of venom in reptiles.

Chapter 5

Gene duplication and venom gene recruitment

Snake venom has been hypothesised to have originated via a process that involves the duplication of genes encoding non-toxic body proteins with subsequent "recruitment" of the copy to the venom gland, which then becomes weaponized through the gradual accumulation of mutations in protein coding regions. 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 requiring the evolution of novel *cis*-regulatory architecture. Nevertheless, this hypothesis has proved to be pervasive and is often accepted as established fact, despite being supported by only a single publication. A comparative transcriptomic analysis of multiple body tissues from a range of reptile species was used to critically evaluate this hypothesis, and revealed that snake venom does not evolve via the duplication and recruitment of genes encoding non-toxic body proteins. Instead, many proposed venom toxins are expressed in a wide variety of body tissues, including the salivary gland of non-venomous reptiles. Snake venom therefore evolves via the duplication and subfunctionalisation of genes encoding existing salivary proteins prior to them becoming toxic, and not the recruitment of genes from multiple tissues to the venom gland.

5.1 Gene duplication and the evolution of phenotypic novelty

Gene duplication has been suggested to be the major source of novel genetic material (Ohno 1970) and an essential (if not the most important) process in evolutionary adaptation and diversification (Ohno 1967). Arguably the topic of gene duplication has been the most influential and informative area of modern genetics in the last century, and discussions of gene and genome duplication and their role in the evolutionary diversification of species began in 1911 (Kuwada 1911; Taylor and Raes 2004), with subsequent publications in the 1930s (Bridges 1936; Serebrovsky 1938; Taylor and Raes 2004), 1950's (Stephens 1951; Taylor and Raes 2004) and late 1960's (Britten and Davidson 1969; Taylor and Raes 2004). However, its popularisation amongst biologists is often attributed to Susumu Ohno and his book "*Evolution by gene duplication*" (Ohno 1970). At the time these were bold statements to make, especially as there was little in the way of experimental evidence to support them. However, with the dawning of DNA and whole genome sequencing, it is apparent that gene duplications have played a major role in evolution (Taylor and Raes 2005).

Indeed, there is limited scope for evolutionary innovation simply by modifying an existing gene through mutation. Conversely, the *de novo* formation of a gene occurs extremely rarely (if at all), and therefore duplicating an already established and functional gene provides a template with which to experiment.

Following duplication a duplicate gene must become fixed within a population, and then be preserved in that population or it will eventually be deleted from the genome. The rate of gene duplication in vertebrates is proposed to vary from 1 gene per 100 to 1 gene per 1000 per million years (Cotton and Page 2005; Lynch and Conery 2000; Lynch and Conery 2003), with the fixation rate of those genes being much lower. Gene duplication can therefore be considered to be a rare process, especially in vertebrate genomes. Nevertheless, some duplicate genes are occasionally fixed and can evolve new functions, which may lead to speciation or phenotypic novelty allowing adaptation to an ecological niche.

5.2 Mechanisms of gene duplication

5.2.1 Unequal crossing over (ectopic recombination)

Unequal crossing over occurs when homologous chromosomes become misaligned during meiosis, resulting in the duplication of a chromosomal region from one chromosome which may be coupled with a deletion of a region from the other (Hurles 2004) (Figure 5.1). As a result, duplicated genes are linked on the same chromosome (i.e. they are tandemly arranged) and there is potential to

duplicate multiple genes in a single duplication event. This type of duplication is facilitated by the presence of repetitive sequence elements which cause false homology between chromosomal regions and can lead to slipped-strand mispairing and misalignment during recombination (Levinson and Gutman 1987).

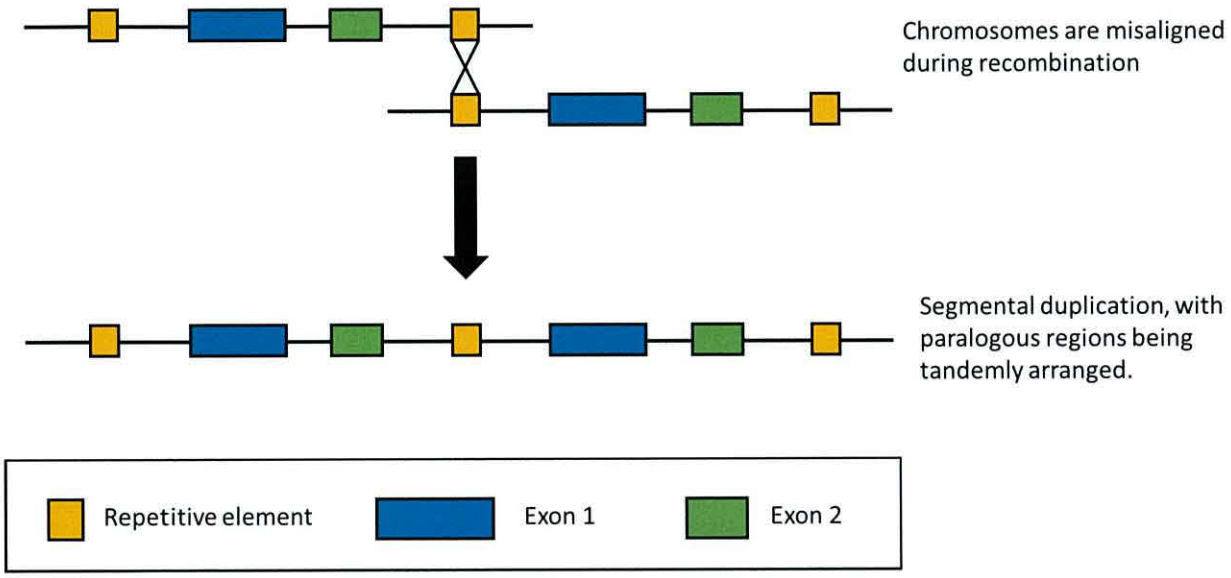


Figure 5.1. Recombination occurs between two homologous chromosomes which are misaligned. As a result, a gene is duplicated, with both paralogs being immediately adjacent to each other (tandemly arranged). This figure is adapted from Figure 1 in (Hurles 2004).

5.2.2 Chromosomal or whole genome duplication (polyploidy)

Whole genome duplication (polyploidy) causes the multiplication of a chromosome set within an organism, and can be a result of the formation of diploid gametes during meiosis (Van de Peer and Meyer 2005). If two of these diploid gametes fuse the resulting progeny will have four sets of homologous chromosomes although they will be unable to inbreed with the parents due to having a differing karyotype and consequently polyploidy can be considered to be a cause of sympatric speciation (Bolnick and Fitzpatrick 2007).

It is distinct from aneuploidy in that a full set of chromosomes is duplicated, and as such there is no over- or under-representation of gene dosage (Van de Peer and Meyer 2005). Aneuploidy results in either the loss or gain of a chromosome, caused by the failed separation of homologous

chromosomes during meiosis (Taylor and Raes 2005). As this can result in the gain or loss of a considerable number of genes, and therefore alter gene dosage, there can be significant phenotypic effects (For example Down syndrome in humans is the result of an additional copy of chromosome 21 (Taylor and Raes 2005)).

Polyploidy has been found to be common in plants (Adams and Wendel 2005) but much rarer in animals (although examples do exist such as some African clawed frog species (*Xenopus spp.*) (Hughes and Hughes 1993)). Interestingly, all polyploid reptiles are triploid (i.e. have three sets of homologous chromosomes) and all reproduce by parthenogenesis (Song et al. 2012) which is likely to be necessary due to one set of chromosomes being unable to form homologous pairs during meiosis.

It is believed that there has been two rounds of whole genome duplication early in vertebrate evolution (the 2R hypothesis) (Kasahara 2007), with a proposed additional round of genome duplication in teleost fish (Meyer and Van de Peer 2005). It is easy to imagine how this process could have been a major driver in vertebrate adaptation and diversity due to the creation of many additional copies of genes, forming a large reservoir of genetic material with which to experiment. Whilst many gene copies would have been functionally redundant following duplication and eventually lost from genome, some were retained such as Hox gene clusters (Kuraku and Meyer 2009).

5.2.3 Retrotransposition

Retrotransposition occurs when an mRNA molecule is reverse transcribed to cDNA and subsequently inserted into the genome, giving rise to a retrogene (Zhang 2003). The resulting sequence can be integrated at a random position in the genome, and as such is not linked to the ancestral gene on the same chromosome as in tandem gene duplication resulting from unequal crossing-over (Hurles 2004). As the duplicated sequence originates from an mRNA molecule there are several characteristic features associated with this mode of duplication, namely a lack of introns and regulatory regions and the presence of poly-A tracts and flanking short repeat sequences (Long 2001). As a consequence, the default fate of a retrogene is pseudogenisation as it lacks the regulatory sequences required for expression (Zhang 2003). However, if the retrotransposed duplicate is inserted downstream of a functional promoter it may still be expressed, potentially also inheriting a novel expression pattern depending on the regulatory sequence (Long 2001; Zhang 2003).

5.3 The potential fate of duplicate genes

Following fixation within a population, there are three possible fates for a gene duplicate: nonfunctionalisation, neofunctionalisation and subfunctionalisation (Hurles 2004).

5.3.1 Nonfunctionalisation (pseudogenisation)

The most common fate for a duplicate gene is the loss of its function, thus becoming a pseudogene (Mighell et al. 2000; Lynch and Conery 2000; Presgraves 2005). That is to say a gene which is highly similar to a functional gene but due to mutation (e.g. introduction of a premature stop codon) is unable to be expressed, or encodes a defective product (Hurles 2004). As potentially only a single point mutation is required to disrupt a necessary transcription factor binding site, disrupt protein function, or result in a premature stop codon, nonfunctionalization is considered to be the most frequent fate of duplicate genes (Lynch and Conery 2000). This process can occur regardless of whether the duplicated paralog confers a selective advantage or not, and duplication of a partial sequence of a gene can give rise to a pseudogene by default (Mighell et al. 2000). Pseudogenes are common in eukaryotic genomes (Wilde 1986) and it has been suggested that the window between duplication and nonfunctionalization is relatively small (Lynch and Conery 2000). Dollo's law states that the evolutionary trajectory of an organism is not reversible (Dollo 1893), and as such pseudogenes should be considered to be rendered permanently non-functional. However, it has been shown that silenced genes can be reactivated over relatively short time-scales (Marshall et al. 1994) meaning that pseudogenes could provide a starting point for the evolution of new genes.

5.3.2 Neofunctionalisation

In some cases a duplicate gene is retained and undergoes neofunctionalisation (where one of the duplicates assumes a new role, independent of the ancestral function (Force et al. 1999)). This can either involve the modification of the protein coding region of a gene, leading to a novel function, or the modification of the gene regulatory regions of a gene, leading to a novel expression domain (Figure 5.2). The process of evolving an entirely new function is known to be incredibly rare and there are few conclusive examples of it in the literature (Deng et al. 2010; Escriva et al. 2006; Van Damme et al. 2007).

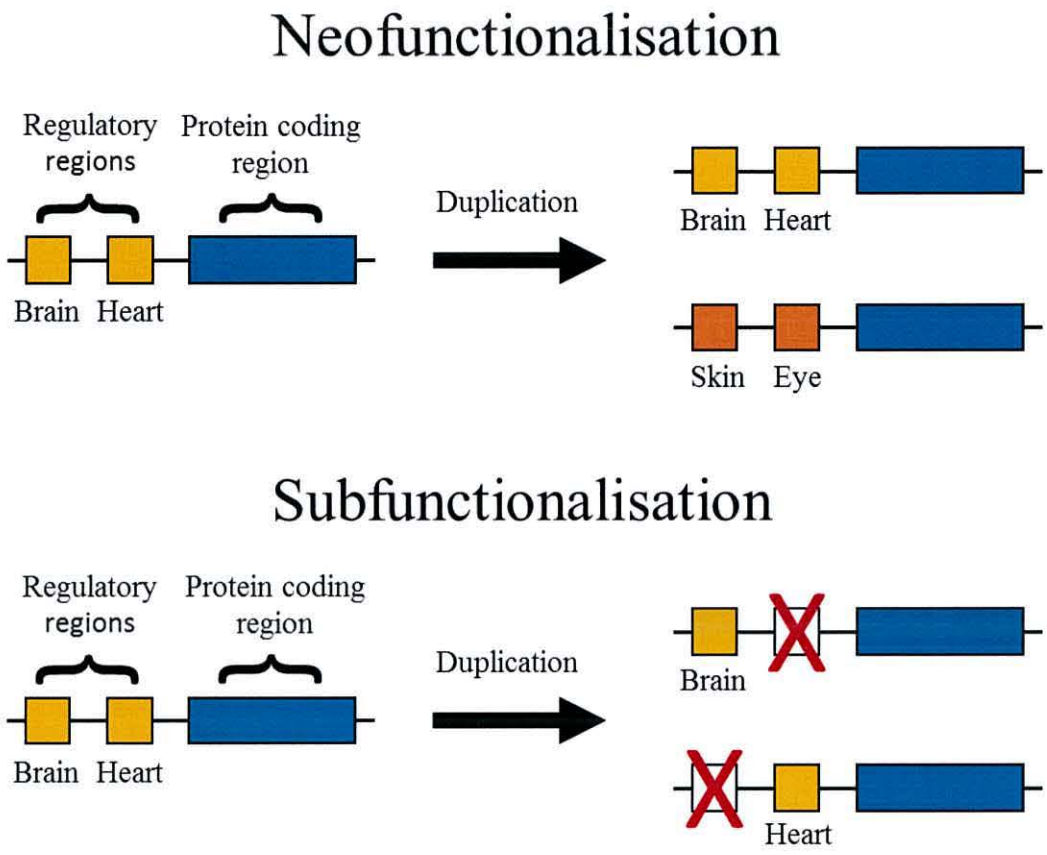


Figure 5.2. Neofunctionalisation and subfunctionalisation. In neofunctionalisation, one of the gene duplicates develops new *cis*-regulatory regions through mutation, and as a consequence evolves an entirely novel tissue expression pattern. In subfunctionalisation, the ancestral tissue expression pattern is divided between the two gene duplicates, with tissue-specific *cis*-regulatory regions being lost/rendered non-functional through mutation. Consequently the role of the original gene is divided between the two paralogs.

Ohno proposed that following gene duplication one duplicate is subject to "relaxed" selective pressure, as the other duplicate is carrying out the ancestral role (Ohno 1970), and as a consequence it is free to undergo mutation and eventually evolve a novel function. However, Bergthorsson et al. (Bergthorsson et al. 2007) postulated that this process is only possible if both gene duplicates are maintained in the population for a sufficient period of time for mutation to occur. They suggested that this would be achieved by selection for the retention of both gene copies, meaning that there would be no relaxed selective pressure allowing for one gene copy to develop a new function

(which they refer to as "Ohno's dilemma"). Instead they proposed that an ancestral gene develops a side-activity along with its usual function prior to duplication, which is amplified and maintained within the population if the side-activity confers a selective advantage. As a consequence the gene copies are maintained and any improvement to their function is favoured by selection, and so they are free to diverge (Bergthorsson et al. 2007).

5.3.3 Subfunctionalisation

Subfunctionalisation occurs when the role of an ancestral gene is partitioned between the two paralogs resulting from a gene duplication event (Figure 5.2) (Hurles 2004). As such, neither paralog can evolve a novel function and this process can be achieved through neutral mutation, making it more parsimonious than neofunctionalisation. In the Duplication, Degeneration, Complementation (DDC) model proposed by Force et al. (Force et al. 1999), both gene paralogs undergo deleterious mutation following gene duplication leading to an inability of either copy to carry out their ancestral role. As a result both gene copies are essential to carry out the function of the ancestral gene, and the loss of functionality in one paralog is complemented by the retention of that function in the other (Force et al. 1999). Subfunctionalisation can however lead to the specialisation of paralogous genes. For example, if the ancestral gene is pleiotropic (i.e. a single gene fulfils multiple roles) it is constrained as an alteration to one role may negatively affect the other(s). However, if this gene is then duplicated and the ancestral role is partitioned between paralogs, they are then able to specialise which could potentially lead to an improvement in functionality. This is especially true for paralogs which were ancestrally expressed in multiple tissues but subsequently become restricted to a specific tissue (Li et al. 2005).

5.4 Venom gene duplication and recruitment into the venom gland

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 (Casewell et al. 2012; Casewell et al. 2013; Fry et al. 2009b; Fry et al. 2012a; Kwong et al. 2009; Lynch 2007; 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 the term neofunctionalisation is used in reference to the acquisition of novel sites of expression at the level of individual tissues (Figure 5.2) and not the acquisition of novel functions at a molecular level as this latter interpretation is separate from the claims of the duplication/recruitment hypothesis. Even so, the evolution of novel functions through the modification of protein coding regions has only been shown to have occurred for a small number of venom toxins (Kini 2002; Kini 2003; Kini and Doley 2010; Lynch 2007) as the majority of duplicated toxins retain their ancestral bioactivity (Fry 2005; Warrell 2010)). Therefore, venom gene isozymes (paralogs which catalyse the same biochemical reaction) may have been erroneously claimed to have been neofunctionalised (Lynch 2007; Brust et al. 2013).

However, the mechanisms underlying repeated gene duplications and, more importantly, the gene regulatory changes that occur to facilitate "recruitment" into the venom gland are currently unknown. Given that whole genome duplication is a rare event in vertebrates in general and reptiles in particular (Mable 2004; Otto and Whitton 2000), 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 (Piskurek et al. 2006; Di-Poi et al. 2009; Shedlock et al. 2007; Kordiš and Gubenšek 1997) provides regions of pseudo-homology that facilitate unequal crossing-over during homologous recombination, producing tandemly-arranged duplicates. This process requires neither germline 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 snake venom metalloproteinases (SVMPs), cysteine-rich secretory proteins (CRISPs) and lectin genes in the king cobra genome (Vonk et al. 2013) and for phospholipase A_2 (PLA_2) genes in the Okinawan habu (*Protobothrops 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 this scenario explains how existing venom genes may undergo repeated rounds of gene duplication whilst retaining their required *cis*-regulatory regions, it does not explain how a gene may be originally "recruited" into the venom gland. The paralogous genes resulting from a gene duplication will be 100% identical and will initially be functionally redundant (i.e. their spatial and temporal expression patterns will be identical) (Force et al. 1999; Lynch and Force 2000).

Therefore, in order to recruit a gene from a body tissue into the venom gland, a novel combination of transcriptional regulatory sequences must arise in order to alter its expression pattern to be venom gland-specific. 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 transcription factor binding sites of venom genes are discussed in more detail in chapter 6). Therefore, the unlikeliness of evolving new combinations of transcription factor binding sites to give rise to novel tissue expression, coupled with the rarity of gene duplication in vertebrate genomes, should make the process of gene duplication and recruitment to the venom gland an extremely rare event.

5.5 Reverse recruitment

It has further been proposed that the process of recruitment may in fact be reversible, where a duplicated venom gene is recruited back into a body tissue, loses toxic function, and undergoes further neofunctionalisation to fulfil a non-toxic physiological role (Casewell et al. 2012). This "reverse recruitment" hypothesis was based largely upon non-toxin sequences from several body tissues of non-venomous snakes grouping together with toxin sequences in phylogenetic trees. The authors suggest either reverse recruitment or the co-expression of toxin genes in multiple tissues as an explanation for the non-monophyly of Toxicoferan toxin genes which their phylogenetic trees display. Whilst this dynamic toing and froing of both a genes function and its expression may add another layer of intrigue and complexity to the snake venom origin story, this hypothesis again assumes that both gene duplication and neofunctionalisation are extremely common processes within the reptile lineage.

5.6 Venom gene restriction-an alternative hypothesis

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 or salivary gland. Therefore following duplication these genes 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 5.3) 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.

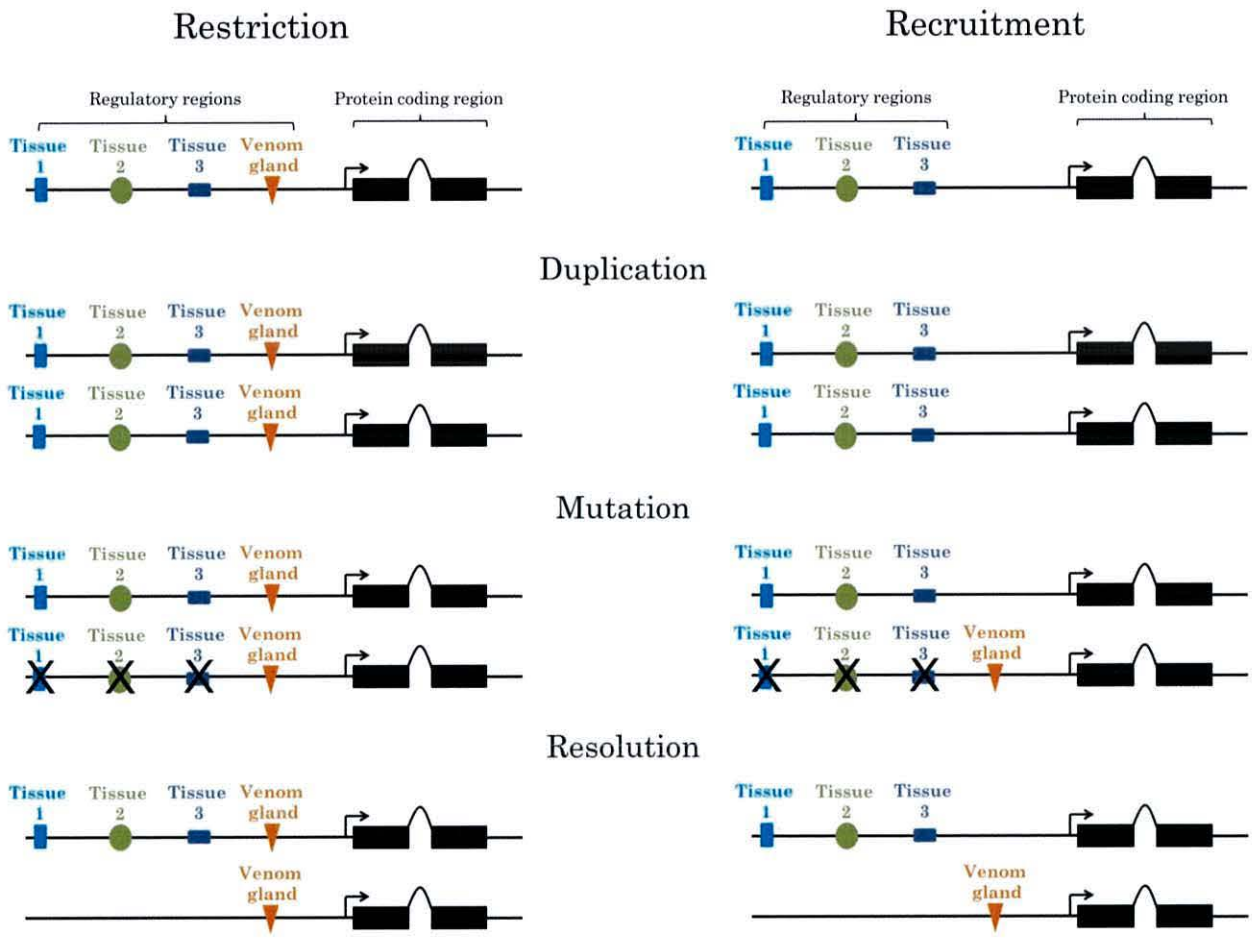


Figure 5.3. Restriction and recruitment. Duplicated genes may be either restricted or recruited to the venom gland, with the recruitment dependent on the evolution of new combinations of transcription factor binding sites in upstream regulatory regions. Mutation/loss of regulatory regions is indicated with an X.

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 the existing evidence for the duplication and recruitment of genes into the venom gland is reviewed. Furthermore, a comparative transcriptomic survey of gene expression in the venom glands and body tissues of a number of reptile species was carried out, including the painted saw-scaled viper (*Echis coloratus*), a venomous, medically important viperid; the corn snake (*Pantherophis guttatus*) a non-venomous colubrid that kills its prey via constriction ; the rough green snake (*Opheodrys aestivus*) a non-venomous colubrid that grasps prey and simply swallows it; the royal python (*Python regius*), a non-venomous pythonid and member of the "primitive" superfamily, Henophidia, and the leopard gecko (*Eublepharis macularius*, Gekkonidae), a lizard that belongs to one of the most basal lineages of squamate reptiles. The phylogenetic position of *E. macularius* is particularly important, as it lies outside of the proposed Toxicofera clade (Chapter 4) (Fry et al. 2006; Fry et al. 2009a; Fry et al. 2012b; Fry et al. 2013). Therefore genes found in the salivary gland of this species can be taken to represent the ancestral squamate expression pattern. Available transcriptomic resources for body tissues in a number of other reptile species were also incorporated into analyses, 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).

5.7 Methods

RNA sequencing and transcriptome assembly methods are discussed in detail in chapter 3. Briefly, 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*) (the general term 'salivary gland' is used for simplicity, to encompass the oral glands of the leopard gecko, rictal glands of the royal python and Duvernoy's gland of the corn snake and rough green snake and no homology to mammalian salivary glands is implied). 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 Appendix 23-24). 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 project accession #SRP042007. Assembled and annotated sequences used in phylogenetic trees have been deposited in the GenBank Transcriptome Shotgun Assembly (TSA) database under the project accession #PRJNA255316.

5.8 Results

5.8.1 Venom genes are ancestrally expressed in multiple tissues

Analysis of newly generated and publicly available transcriptomic data revealed that many of the gene families which have subsequently evolved and diversified to encode venom toxins are expressed in a multitude of tissues, including the venom gland and the salivary gland of non-venomous reptiles (Figure 5.4). Gene families which are unlikely to represent toxins in reptiles such as cystatins and waprin (as discussed in chapter 4) were found to have a wide expression pattern (Figure 5.4), further supporting the hypothesis that these simply represent housekeeping or maintenance genes.

	Tissue/species																						
	Sal/VG						Scent gland					Skin					Bra		Liv		K	O	P
	Ema	Pre	Oae	Pgu	Eco	Oha	Ema	Pre	Oae	Pgu	Eco	Ema	Pre	Oae	Pgu	Eco	Eco	Pgu	Eco	Tel	Eco	Eco	Oha
3ftx		■	■	■	■	■			■	■	■			■	■	■	■		■		■	■	■
ADAM	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	
Acetlycholinesterase	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■		■			
Complement c3	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■		■	■	■		■
Crisp	■			■	■	■	■	■	■	■	■	■		■	■	■		■			■	■	
Crotamine/β-defensin	■			■	■		■	■	■	■	■	■			■	■			■	■	■		■
Cystatin	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Factor V	■						■		■		■	■				■	■		■	■	■	■	
Factor X	■		■	■	■	■	■	■	■	■	■	■		■	■	■	■		■	■		■	■
Kallikrein	■	■			■		■	■	■	■	■		■			■							■
Kunitz	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
L-amino acid oxidase	■				■	■	■	■	■	■	■	■				■	■					■	
Lectin	■	■	■	■	■		■	■	■	■	■	■	■		■	■	■	■		■	■	■	■
Natriuretic peptide													■								■		
Nerve growth factor	■			■	■	■	■		■	■	■	■	■	■		■				■		■	
Phospholipase A2	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■		■	■	■	■	
Vegf	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Vespryn				■		■	■	■	■	■		■	■	■									
Waprin	■		■	■	■	■	■		■			■		■		■	■		■	■	■		■

Figure 5.4. Tissue distribution of putative toxin gene families. 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*).

5.8.2 Evaluation of previously proposed venom gene recruitment events in snakes

The study cited most frequently in support of the duplication and recruitment hypothesis is that of Fry (Fry 2005) in his paper *"From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins"* (see for example (Casewell et al. 2012; Casewell et al. 2013; Jiang et al. 2011; Warrell 2010)) and therefore this hypothesis will be herein referred to 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 (1) snake venom toxin sequences derived primarily from EST-based studies of only the venom gland could be considered to be venom gland-specific and (2) if they were related to a gene known to be expressed in a "body" 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 multiple tissues of the snake and the omission of data from non-venom gland tissues makes it impossible to elucidate the true extent of a genes expression pattern. 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 could possibly show any evidence in support of gene duplication and recruitment into the venom gland: crotamine; natriuretic peptide; complement C3 and Group IB phospholipase A_2 (Fry 2005). As nerve growth factor has been suggested to have undergone a duplication somewhere within the Elapid lineage (Sunagar et al. 2013; Hargreaves et al. 2014a; Hargreaves et al. 2014b) (see also chapter 4), and coagulation factors V and X have been suggested to have been duplicated and recruited from genes expressed in the liver (Reza et al. 2006; Reza et al. 2007; Kwong et al. 2009), they have also been included in this analysis and are discussed below.

Crotamine

The South American rattlesnake (*Crotalus durissus terrificus*) *crotamine*-like sequence labelled as 'Pancreas' (accession number Q6HAA2) used in Fry's study 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). This newly generated 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 5.4) and that the toxic duplicate has been restricted, not recruited, to the venom gland.

Natriuretic peptides

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 (Hayashi et al. 2003; Hayashi and Camargo 2005; Murayama et al. 1997). Few natriuretic peptides are found in this dataset (Figure 5.4), and the low number of these sequences previously characterised would suggest that they play little role in the venom of snakes

other than *Bothrops spp.*, where they appear to have undergone duplication and subfunctionalisation.

Complement C3 ("Cobra venom factor")

For *complement C3*, Fry's analysis (Fry 2005) 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 new transcriptomic 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 5.4 and 5.5) and that a paralogous copy of this gene has therefore been restricted to the venom gland following duplication within cobras. An additional gene duplication also appears to have taken place in the Australian lowland copperhead, *Austrelaps superbus*, although the genes expressed in the venom gland appear to be highly similar to those expressed in the liver, and it is not stated whether the liver and venom gland samples were derived from the same individual (Rehana and Kini 2007; Rehana and Kini 2008). Therefore, it is possible that the venom gland and liver sequences represent the same gene, with both duplicates of *complement c3* being expressed in the venom gland and liver (and possibly other tissues) simultaneously.

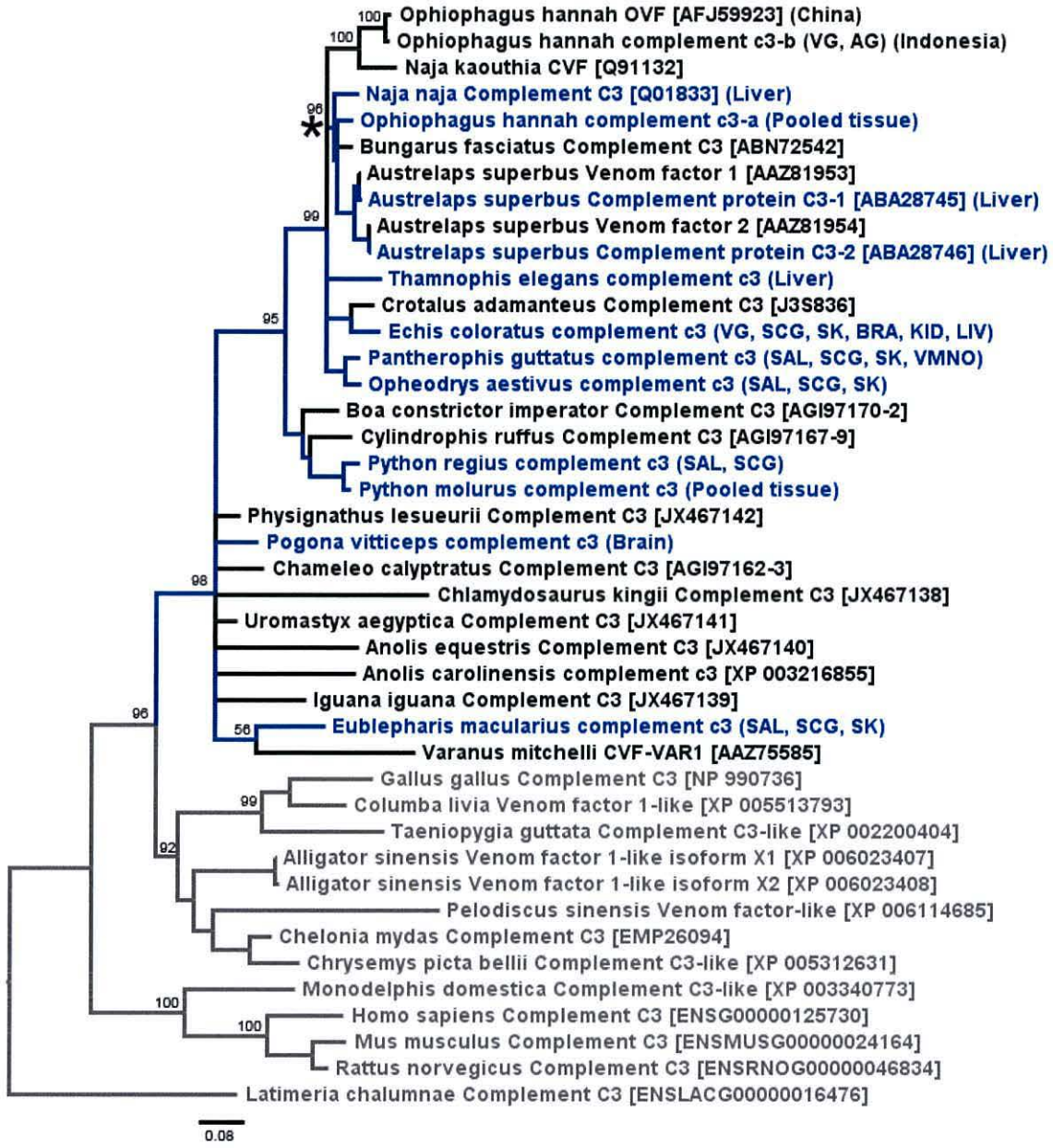


Figure 5.5. *complement C3* genes are expressed in a diversity of tissues, including venom and salivary glands. Following a gene duplication event (marked with *, shaded dark grey) one paralog has been restricted to the venom gland in the king cobra (*Ophiophagus hannah*) and the monocled cobra (*Naja 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.

Group IB Phospholipase A_2

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 the recruitment of this gene family. However, *PLA_2 IB* genes were found to be expressed in several body tissues, including the leopard gecko salivary gland (Figures 5.4 and 5.6), suggesting a wider ancestral expression pattern than previously claimed.

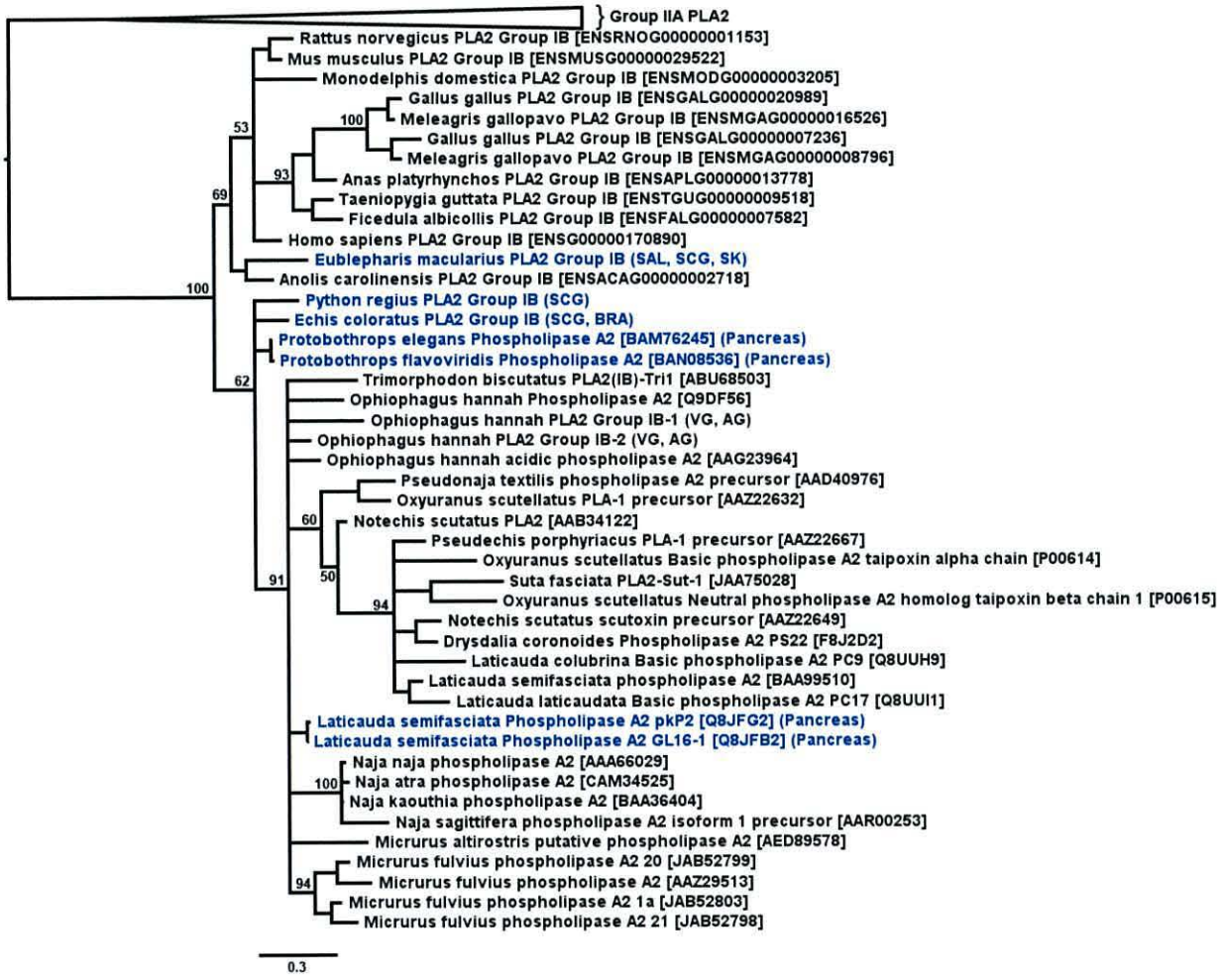


Figure 5.6. *Phospholipase A_2 group IB* genes are expressed in a diversity of tissues, including leopard gecko (*Eublepharis macularius*) salivary glands. Following a gene duplication event somewhere in the advanced snake lineage one paralog has been restricted to the venom gland. Lineages for which body (non-venom gland) sequences are available are coloured blue and bootstrap values for 500 replicates are shown above branches.

Nerve growth factor

It has recently been suggested that there has been a duplication of *nerve growth factor* (*ngf*) genes in some snake species (Sunagar et al. 2013), although the presence of additional copies of *ngf* in certain species of cobra has been known for some time (Koh et al. 2004; Lipps 2000). The non-toxic form of *ngf* (more specifically *ngfa*) is in fact expressed in a diversity of tissues, including the salivary glands of non-venomous reptiles (Figures 5.4 and 5.7). The expression of the putatively toxic version (*ngfb*) has therefore also been restricted to the venom gland following duplication.

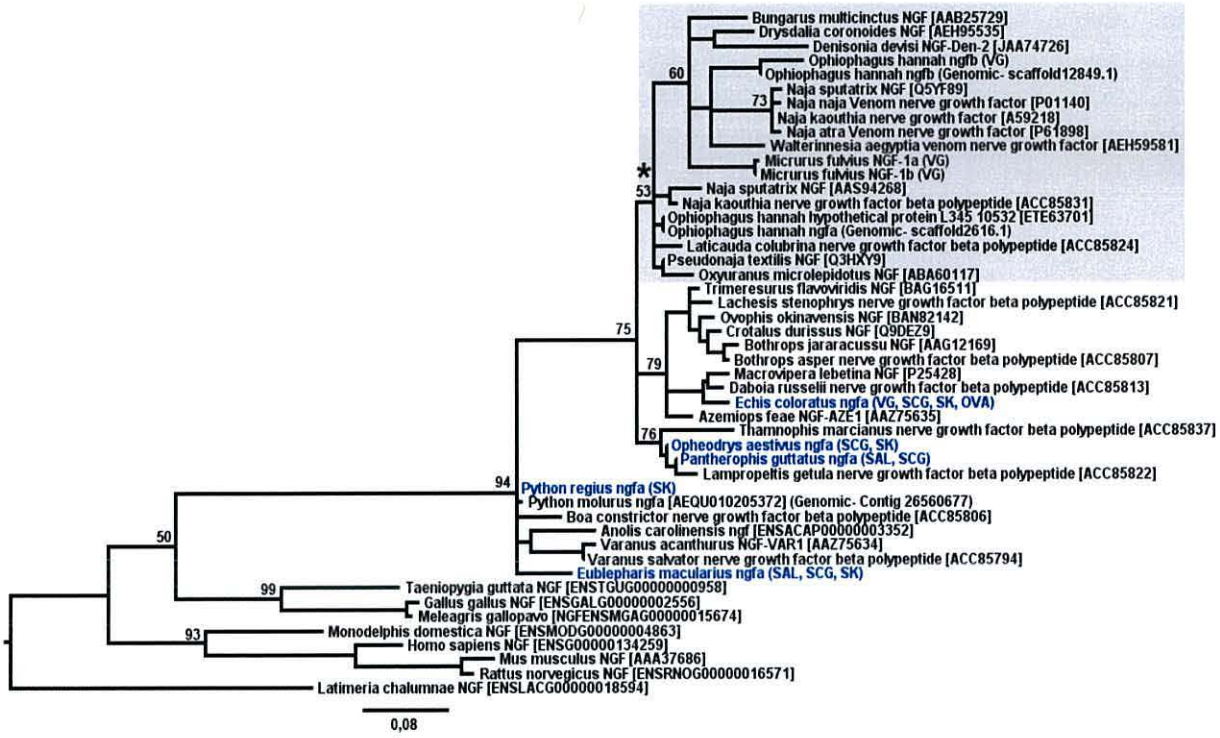


Figure 5.7. *nerve growth factor* (*ngf*) genes are expressed in a diversity of tissues, including venom and salivary glands. Following a gene duplication event (marked with * and shaded) one paralog has been restricted to the venom gland. Lineages for which body (non-venom gland) sequences are available are coloured blue and bootstrap values for 500 replicates are shown above branches.

Coagulation factors V and X

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 and Kini; Kwong et al. 2009). 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. Results from this analysis in fact show *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 (Figures 5.4 and 5.8).

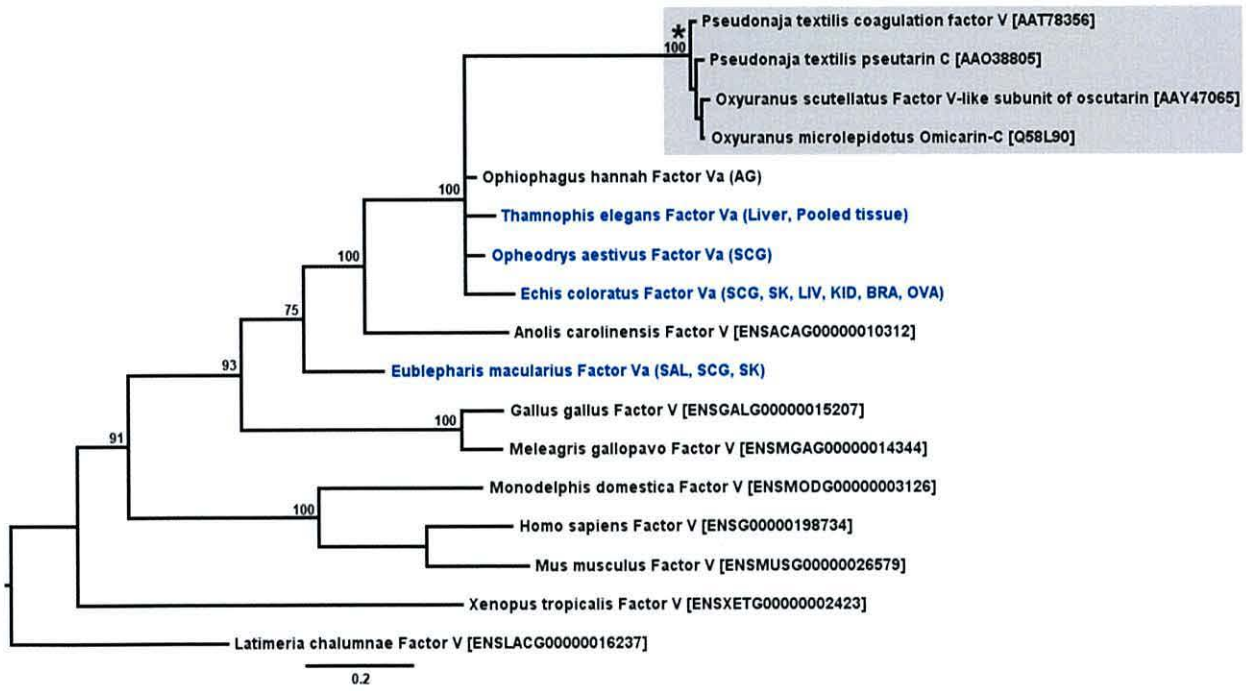


Figure 5.8. *factor V* genes are expressed in a diversity of tissues, including leopard gecko (*Eublepharis macularius*) salivary glands. Following a gene duplication event (marked with * and shaded) one paralog has been restricted to the venom gland. Lineages for which body (non-venom gland) sequences are available are coloured blue and bootstrap values for 500 replicates are shown above branches.

Factor X is also expressed in multiple tissues (Figures 5.4 and 5.9), 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 restriction to the venom gland following duplication.

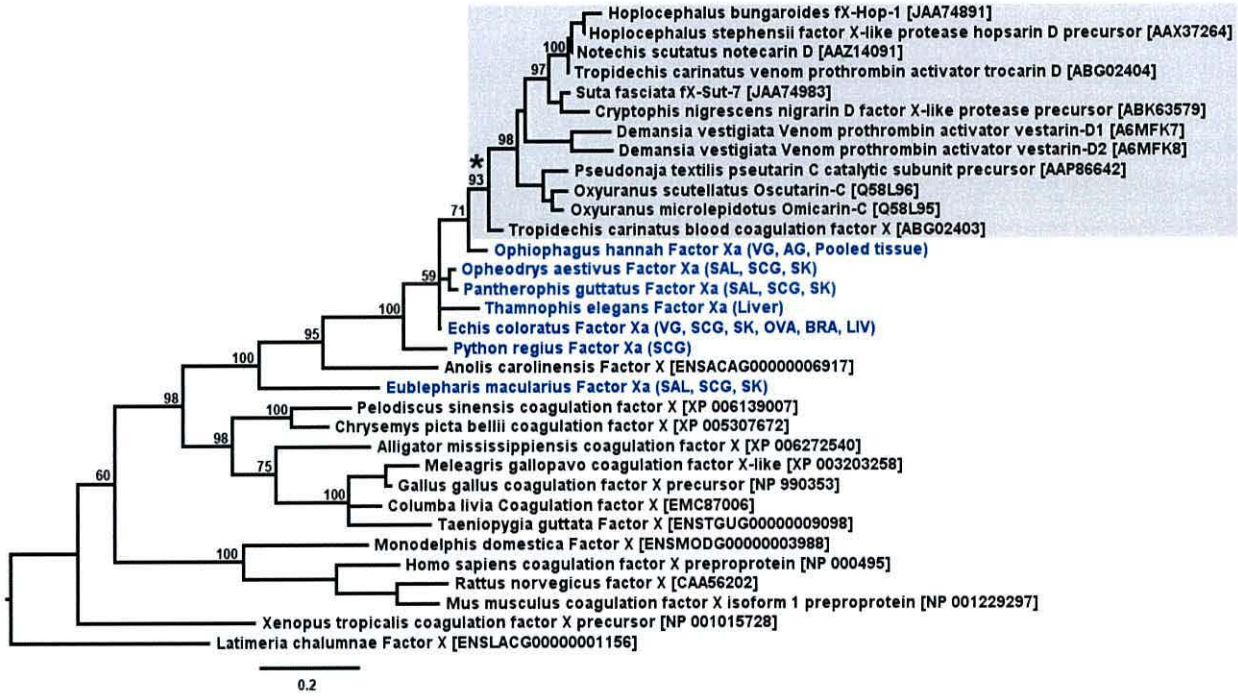


Figure 5.9. *factor X* genes are expressed in a diversity of tissues, including venom and salivary glands. Following a gene duplication event (marked with * and shaded) one paralog has been restricted to the venom gland. Lineages for which body (non-venom gland) sequences are available are coloured blue and bootstrap values for 500 replicates are shown above branches.

5.9 Discussion

The hypothesis that snake venom evolves via the duplication of physiological or body genes and subsequent recruitment into the venom gland is 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 there is evidence of a diverse tissue expression pattern, including the salivary gland of non-venomous reptiles (Figure 5.4), demonstrating that, if they do encode toxic venom components (Hargreaves et al. 2014a), 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)), this analysis shows that the majority of snake venom toxins are likely derived from pre-existing salivary proteins. This would mean that the evolution of a novel gene regulatory network to achieve venom gland-specific expression is unnecessary, as the gene is ancestrally already expressed in the venom gland prior to becoming toxic.

The "reverse recruitment" study (Casewell et al. 2012) was the first publication to include a large amount of non-venom gland tissue transcriptomic data in its analyses. However, it should be noted that these sequences were derived from transcriptomic studies of the heart and liver of the Burmese python (*Python molurus bivittatus*) (Castoe et al. 2011) and mixed body tissues of the garter snake (*Thamnophis elegans*) (Schwartz et al. 2010), i.e. from non-venomous species. Therefore, due to the lack of non-venom gland tissues from a venomous species, the proposed duplication history of venom genes in this study are not supported. Furthermore, the authors made the assumption that "body" sequences nesting within a "venom clade" represented former toxins. The more parsimonious explanation is that these sequences actually represent housekeeping genes which have never been toxic, requiring no gene duplications at all. As such, the default conclusion should more rightly have been reptile sequences grouping together in a reptile clade. Several of these phylogenetic trees also have sequences from other species such as human (*Homo sapiens*) and rainbow trout (*Oncorhynchus mykiss*) nested between reptile clades, implying that these trees in fact contain sequences for multiple genes, and so the formation of multiple clades is inevitable. Taking this into account, whilst Casewell *et al.* propose reverse recruitment as an explanation for

the non-monophyly of Toxicoferan toxin genes in their phylogenetic trees, this new data coupled with analyses in chapter 4 suggest that these genes are in fact housekeeping or maintenance genes, and not former toxins.

The proposal that venom is an incredibly complex cocktail of proteins (Casewell et al. 2013; Fox and Serrano 2008; Kini 2002; Wagstaff et al. 2006) recruited from multiple body tissues (Casewell et al. 2013; Fry 2005; Fry et al. 2009a; Warrell 2010), requiring extensive gene duplication and neofunctionalisation appears to be largely unsupported based on available data. Instead, snake venom should be considered to be 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 (Fahmi et al. 2012; Kulkeaw et al. 2007; Vonk et al. 2013; Wagstaff et al. 2009). It is possible that the elevated expression level caused by two duplicates of a pre-existing salivary protein could confer an increased efficacy to prey capture, and as such the expression of this paralog in the venom or salivary gland is maintained. The known increased expression of a *factor X* paralog following an insertion in the promoter region (Han et al. 2013; Kwong and Kini; Kwong et al. 2009; Reza et al. 2007) and the increased expression of *crotamine* in the venom gland following duplication (Rádis-Baptista et al. 2003; Rádis-Baptista et al. 2004) would suggest that this may be the case. The additional duplication of *complement c3* in *Austrelaps superbus* (Figure 5.5) may indicate the beginning phase of this process. It is possible that these paralogs are both expressed in the liver (Rehana and Kini 2008), venom gland (Rehana and Kini 2007) and potentially other tissues. If the expression of one paralog in the venom gland confers an advantage to the function of the venom, this paralog may subsequently be restricted to the venom gland and may undergo mutation to develop toxicity.

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*, and 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 (Casewell et al. 2013; Durban et al. 2011; Fry et al. 2008)). 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 newly available data, but that it was never supported originally. 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 (Castoe et al. 2013; Wall et al. 2011). 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.

In addition to the implications these findings have for snake venom evolution, they also highlight the problem of 'just-so stories' (Kipling 1902) in evolutionary biology, here used as a pejorative term to describe when a hypothesis becomes accepted as established fact due to its appealing nature, whilst its fundamental lack of scientific evidence is overlooked. The 'genome to venome' hypothesis has been widely and unquestioningly cited and treated neither as a hypothesis to be tested and refuted (Popper 2005), nor as a scientific research programme to provide predictions to be investigated (Lakatos 1980). As a consequence an incorrect hypothesis has been accepted for almost a decade, and further investigation into the true origin of venom genes in reptiles has been hindered. Whilst the role of gene duplication should rightly be considered as part of the core of the snake venom evolution research programme, further testing and further scrutiny are required in order to elucidate how the venom gene repertoire in snakes first evolved.

Chapter 6

The regulation of snake venom production

Snake venom is an essential prey capture tool that must be rapidly replenished following expenditure. Whilst previous chapters have shed light on the processes in which venom genes can arise, the regulatory mechanisms underlying the expression of these genes are mostly unknown. Using a comparative transcriptomic approach the first large-scale survey of the transcription factors and signaling pathways involved in venom production has been carried out. Venomous and non-venomous species produce similar numbers of secreted products in their venom or salivary glands and only one transcription factor (*Tbx3*) is expressed in venom glands but not salivary glands. As this and other proposed transcription factors involved in venom gene regulation are repressors, it is possible that the rapid initiation of venom production is controlled by negative regulation where the removal of transcription factors and activation of transcriptional complexes already bound to DNA is the key first step. Using the draft genome sequence for the painted saw-scaled viper, *Echis coloratus*, along with other available genomic sequences, conserved transcription factor binding sites in the upstream regions of venom genes were identified. Binding sites appear to be conserved across members of the same gene family, but not between families, indicating that multiple gene regulatory networks (GRNs) are involved in venom production. There is also evidence to suggest that the expression of venom genes shows temporal variation, which may add further support to the hypothesis of multiple GRNs. Finally, venom gene promoters are located close to the start of the gene, facilitating the duplication of venom genes along with their associated regulatory architecture. Taken together, these results suggest that the venom gland is a relatively simple tissue, and a small number of modifications to pre-existing gene expression networks may have led to the evolutionary adaptation of venom in reptiles.

6.1 Venom gene regulation

It is apparent from chapters 4 and 5 that the origin and evolution of venom in reptiles is not as complex as has previously been suggested, with a relatively small number of genes having their expression pattern restricted to the venom gland where they have subsequently diversified via gene duplication and mutations in protein coding regions. However, the gene regulatory mechanisms underpinning the expression of these toxin genes in the venom gland are currently poorly understood, as the majority of research efforts focus on the genes themselves.

Mechanisms affecting the transcription, translation and post-translational modification of venom toxins have been suggested to be responsible for the inter- and intra-specific variation in venom composition (Casewell et al. 2014), which have implications for the efficacy of antivenom treatment (Fry et al. 2003; Gutiérrez et al. 2009; Sunagar et al. 2014). Such variation can be a product of both genome-level and gene-level events, reflecting either variation in gene presence/absence between individuals and populations (as has been demonstrated for the presence or absence of Mojave toxin in the Mojave rattlesnake *Crotalus scutulatus scutulatus* (Wooldridge et al. 2001)) or changes in gene regulation that alter temporal, spatial or quantitative expression of the target genes (for example ontogenetic changes in venom composition in the South American pit vipers *Bothrops asper* (Alape-Girón et al. 2008), *Bothrops insularis* (Zelanis et al. 2007; Zelanis et al. 2008) and *Bothrops atrox* (Guércio et al. 2006; López-Lozano et al. 2002)).

Understanding the genetic basis of the regulation of snake venom production can also inform our understanding of general evolutionary processes, such as the origin and fate of duplicate genes and the molecular-level mechanisms underpinning evolutionary innovation. In Chapter 5 it was shown that snake venom in fact evolves via the duplication and restriction of genes, rather than recruitment (Hargreaves et al. 2014b), and this process of subfunctionalisation rather than neofunctionalisation does not require the *de novo* creation of novel transcription factor binding sites. As such there is no requirement for the evolution of a novel gene regulatory network for venom gene expression. Even so, the transcription of toxin genes must be initiated or upregulated following venom expenditure (either through envenomation of prey or induced manually by "milking") and, given the energetic costs associated with venom production (McCue 2006) it must be assumed that this will be an efficient and rapid process (although these costs may not be as high as those related to other physiological processes such as digestion and shedding (Pintor et al. 2010)). Historically, a window 3-4 days post-milking has been taken to represent a peak period of venom synthesis (Boldrini-França et al. 2009; Kochva 1978; Paine et al. 1992; Wagstaff and Harrison 2006),

although there has long been evidence to suggest that some venom components are transcribed almost immediately following (and certainly within a few hours of) venom expenditure (Currier et al. 2012; Lachumanan et al. 1999). Many previous studies of snake venom composition have understandably focussed on the period of peak synthesis (see for example, (Boldrini-França et al. 2009; Casewell et al. 2009; Margres et al. 2013; Neiva et al. 2009; Wagstaff and Harrison 2006)) and are therefore unable to provide insights into the signalling pathways and transcription factors that may regulate the production of snake venom. Whilst Ca^{2+} mobilisation, α- and β-adrenoceptors, the MAP kinase (ERK1/2) signalling pathway and nuclear factor κB (NFκB) and activator protein 1 (AP-1) transcription factors have been implicated in the initiation of venom production following milking in the South American pit viper *Bothrops jararaca* (Kerchove et al. 2004; Kerchove et al. 2008; Luna et al. 2009; Yamanouye et al. 2000), it is not known how widespread these putative gene regulatory network components might be, nor what other transcription factors and signalling pathways might be involved in the control of snake venom production. In addition, little is currently known about promoter and enhancer regions associated with venom genes, although binding sites for NFκB, *Sp1* and *Nf1* transcription factors have been identified upstream of the crotamine gene in the South American rattlesnake, *Crotalus durissus terrificus* (Rádis-Baptista et al. 2003). Binding sites for some of the proposed transcription factors have also been found in the promoter regions of three-finger toxins present in *Boiga dendrophilia* (*Nf1*) (Pawlak and Kini 2008), *Naja sputatrix* (NFκB, AP-1, *Nf1*, *Sp1*) (Lachumanan et al. 1998; Ma et al. 2001) and *Naja atra* (*Nf1*, *Sp1*) (Chang et al. 2004); and in Phospholipase A_2 group IB genes of *Bungarus multicinctus* (*Sp1*) (Chu and Chang 2002) and *N. sputatrix* (AP-1, *Sp1*) (Jeyaseelan et al. 2000). It is worth noting that no studies of the promoter regions of venom genes in true vipers (Viperinae) are currently available in the literature. Therefore, the aims of this study were to determine the transcription factors and signalling pathways involved in the regulation of venom production through the use of comparative genomic and transcriptomic approaches in a range of reptile species.

6.2 Methods

6.2.1 Genome and transcriptome sequencing

Methods for genome and transcriptome sequencing and assembly are described in detail in chapters 2 and 3. All transcriptomic datasets used in this analysis were also used in chapters 4 and 5 except the *Echis pyramidum* venom gland transcriptome. Transcriptome reads were deposited in the European Nucleotide Archive (ENA) database under accession #ERP001222 and in the NCBI Sequence Read Archive (SRA) under the study accessions #SRP042007 and #SRP043460. Sequencing reads for the *Echis coloratus* genome were deposited in the SRA under study accession #SRP043211.

6.2.2 Downstream transcriptomic analyses

Putative open reading frames (ORFs) were extracted using longorf.pl (available online at http://bip.weizmann.ac.il/adp/bioperl/bioSeq/longorf) using the strict option (where ORFs must start with an ATG).

```
perl longorf.pl -i Transcriptome_assembly.fasta -v > output.txt
```

The output file of predicted ORFs can then be imported into a sequence alignment editor (such as BioEdit (Hall 1999)) and saved as a .fasta file. Signal peptides were predicted by analyzing the ORF sequences using SignalP (Petersen et al. 2011) (version v4.1) which is run using the following command:

```
./signalp -t euk -f short ORFs.fasta > inputfile_OUT
```

The file `inputfile_OUT` was then imported into Microsoft Excel and the contig names of sequences with a predicted signal peptide were extracted and saved as a list in text (.txt) file format. Using this list, ORF sequences with a predicted signal peptide were then extracted using a BioPerl

script obtained online (https://www.biostars.org/p/2822/) (additional material CD). This was used in preference to other, similar scripts as it does not require complete and exact contig name identifiers in order to run. The script was run using the following command:

```
perl findSeqs.pl ORFs.fasta SignalP_contiglist.txt > output.fasta
```

Resulting in a .fasta file of ORF sequences with a predicted signal peptide, i.e. sequences from each transcriptome assembly which are likely to be secreted.

Transcript annotation and assignment of gene ontology (GO) terms was performed using BLAST2GO (Conesa et al. 2005), the KEGG Automatic Annotation Server (KAAS) (Moriya et al. 2007) and by local BLAST using BLAST+ (Camacho et al. 2009) (version v2.2.27). Enrichment analyses using Fisher exact tests were implemented using BLAST2GO. Tissue distribution of transcripts was determined through the creation of a global all tissue assembly and read mapping/transcript abundance estimation using RSEM (Li and Dewey 2011) as a downstream application of Trinity (version r2013-02-25), with an FPKM (Fragments Per Kilobase of exon per Million mapped reads) value of ≥1 taken as confirmation of expression.

To obtain the numbers of shared and unique transcripts to each tissue, the FPKM results files were first merged into one file using the command:

```
ls FPKM_1.txt FPKM_2.txt FPKM_3.txt > combined_FPKMs.txt
```

These results were then placed into tabular format with values below a specified cut-off (in this case 1.0 FPKM) being replaced with a "-" using the python script trinMergeMappings.py obtained from Dr. Martin Swain at the University of Aberystwyth. Usage:

```
python trinMergeMappings.py combined_FPKMs.txt 1.0 >
FPKM_output.txt
```

In this format it is now possible to search for patterns within the file to determine the number of shared or unique transcripts.

For the number of shared transcripts between all samples:

```
grep -v "-" FPKM_output.txt | wc -l
```

The Linux command `grep` can be used to search a file for a specified pattern. In the above example, `-v` is used to specify a reverse of the pattern given, i.e. search the file `FPKM_output.txt` for lines which do not contain the character "-". As FPKM values less than 1.0 are considered to be not expressed and have been replaced with the – symbol, this command will identify lines containing only FPKM values ≥ 1.0 i.e. the transcript is expressed in all samples. The result of this is then "piped" using the `|` symbol to the command `wc`, or word count and the parameter `-l` is used to specify that the output should only be the number of lines counted.

To determine the number of unique transcripts in a sample:

The file `FPKM_output.txt` was opened using Microsoft Excel, the FPKM values belonging to the sample of interest were moved so that they were the first column of results in the file, a column containing only the word "END" all the way down was inserted to be the last column in the file, and then the file was re-saved as a comma-separated values (.csv) file.
A new text file was then created, which was specific for the number of samples in question. For example, when determining the number of unique transcripts in Eco8 venom gland there was 4 venom gland samples in total (4 columns of FPKM values with those of Eco8 venom gland being in the first column) and a column containing the word END in the final 5th column. Therefore, to determine unique transcripts in Eco8 venom gland it is necessary to count the number of lines where there is a numerical value in the first column, followed by the symbol – in the following 3 columns, followed by the word END. So for this example the text file would be:

```
-,-,-,END
```

(values are separated by commas as the file is in .csv format). The text file was saved as `pattern.in`. Again using the `grep` command it was now possible to determine the number of transcripts unique to a particular sample:

```
grep -f pattern.in FPKMfile.csv | sort | uniq -c | wc -l
```

6.2.3 Transcription factor binding sites analysis

Putative transcription factor binding sites were annotated using MultiTF, implemented through Mulan (Ovcharenko et al. 2005) with a 100bp conserved region length and 70% conservation limit. Upstream region sequences were first aligned using ClustalW (Larkin et al. 2007) and then trimmed to leave only contiguous sequence leading up to the start (ATG) codon. Sequences were then surveyed for all available evolutionarily conserved transcription factor binding sites, specifically for transcription factors previously implicated to have a role in the regulation of snake venom production, including NFκB and AP-1, as well as those predicted by the current study, such as Tbx3. The genome assembly for the king cobra was downloaded from GenBank under the accession AZIM00000000.1. The assembly for the Burmese python genome (v2.0) was downloaded from the snake genomics website (http://www.snakegenomics.org/SnakeGenomics/Available_Genomes.html) but this assembly is also available on GenBank under the accession AEQU00000000.2. All genome assemblies of the Boa constrictor generated for Assemblathon 2 (Bradnam et al. 2013) are available for download from GigaDB (http://gigadb.org/dataset/view/id/100060). The assembly "snake_1C_scaffolds" which was assembled by the BCM-HGSC team was used in this analysis (for more details on the assembly method used see Additional file 3 in (Bradnam et al. 2013)).

6.3 Results

This study represents the first large-scale comparative analysis of the regulation of snake venom production, along with the first analysis of venom gene upstream regulatory regions in a member of the Viperinae (true vipers), and the first insight into changes in venom gene expression during the venom replenishment cycle. Results from analyses of potential regulatory mechanisms underpinning venom production are discussed in detail below.

6.3.1 Transcriptomics

Assembled transcriptomes contained between 18,169 and 134,236 contigs of ≥300bp and contig N50 values ranged from 950bp to 3,552bp (full assembly statistics are available in Appendix 23-28). Most contigs encoded an ORF of at least 20aa and of these between 3.72% and 10.64% were predicted to encode a signal peptide and so are likely to be secreted (Appendix 26-28 and next section). Gene ontology was assigned to contigs using BLAST2GO (Conesa et al. 2005) and in order to provide a broad overview of the assigned gene ontologies generic GOSlim annotations were generated. Venom and salivary gland compositions appear to be broadly similar and comparisons of the venom gland and other body tissues in *E. coloratus* do not highlight any major differences (Figures 6.1-6.6).

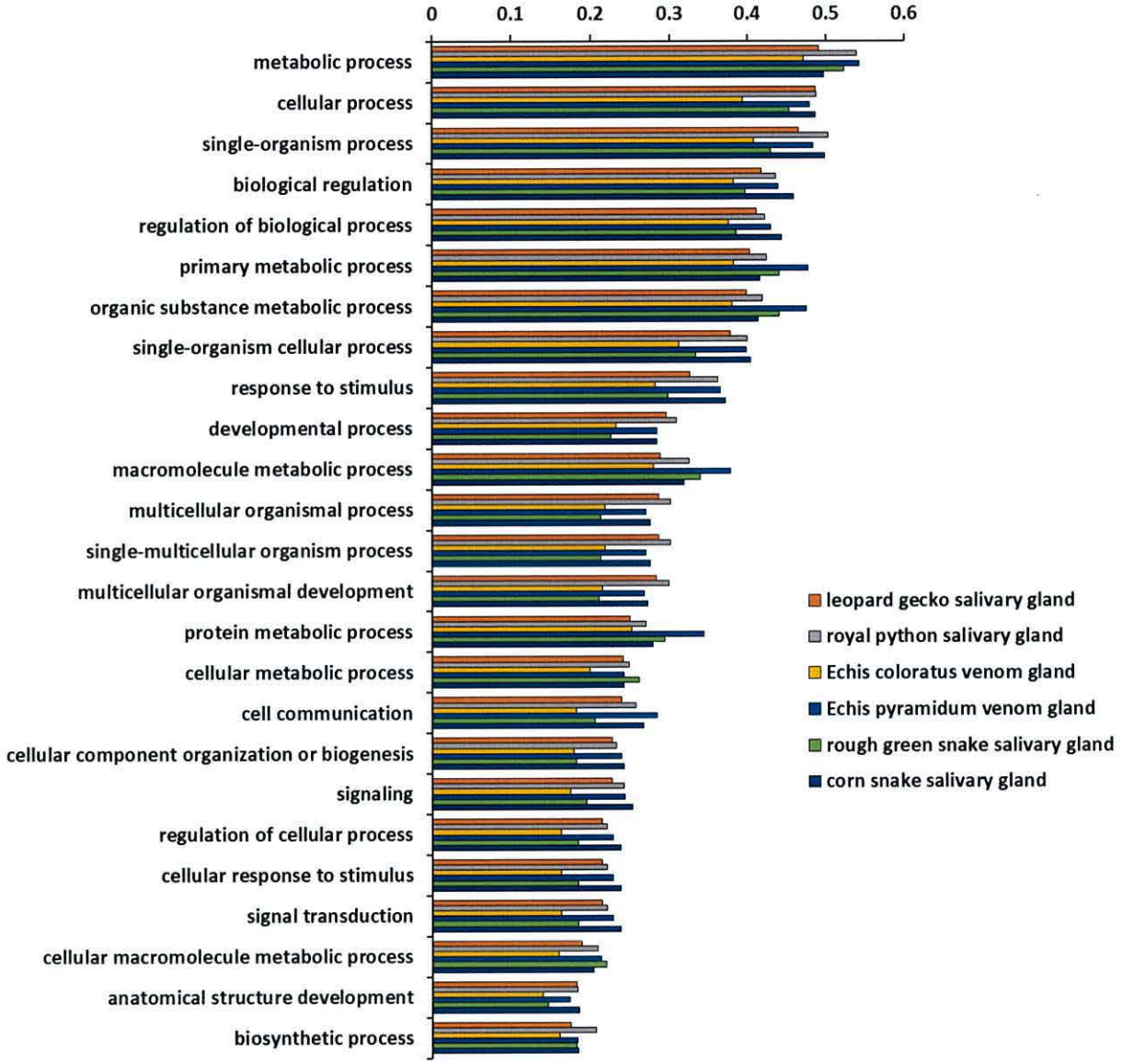


Figure 6.1. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Biological Process' terms for painted and Egyptian saw-scaled viper (*Echis coloratus* and *Echis pyramidum*) venom glands and corn snake (*Pantherophis guttatus*), rough green snake (*Opheodrys aestivus*), royal python (*Python regius*) and leopard gecko (*Eublepharis macularius*) salivary glands.

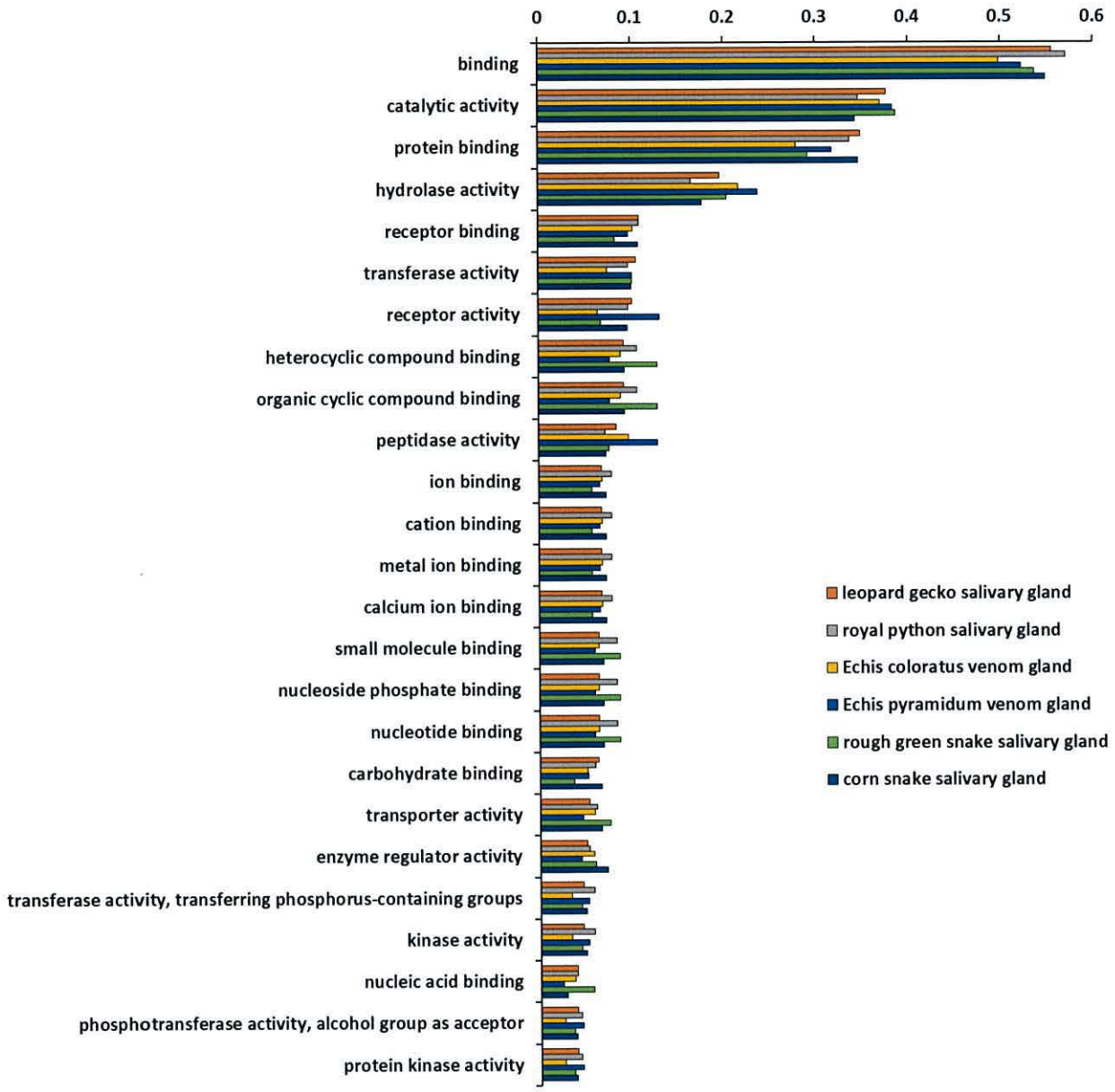


Figure 6.2. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Molecular function' terms for painted and Egyptian saw-scaled viper (*Echis coloratus* and *Echis pyramidum*) venom glands and corn snake (*Pantherophis guttatus*), rough green snake (*Opheodrys aestivus*), royal python (*Python regius*) and leopard gecko (*Eublepharis macularius*) salivary glands.

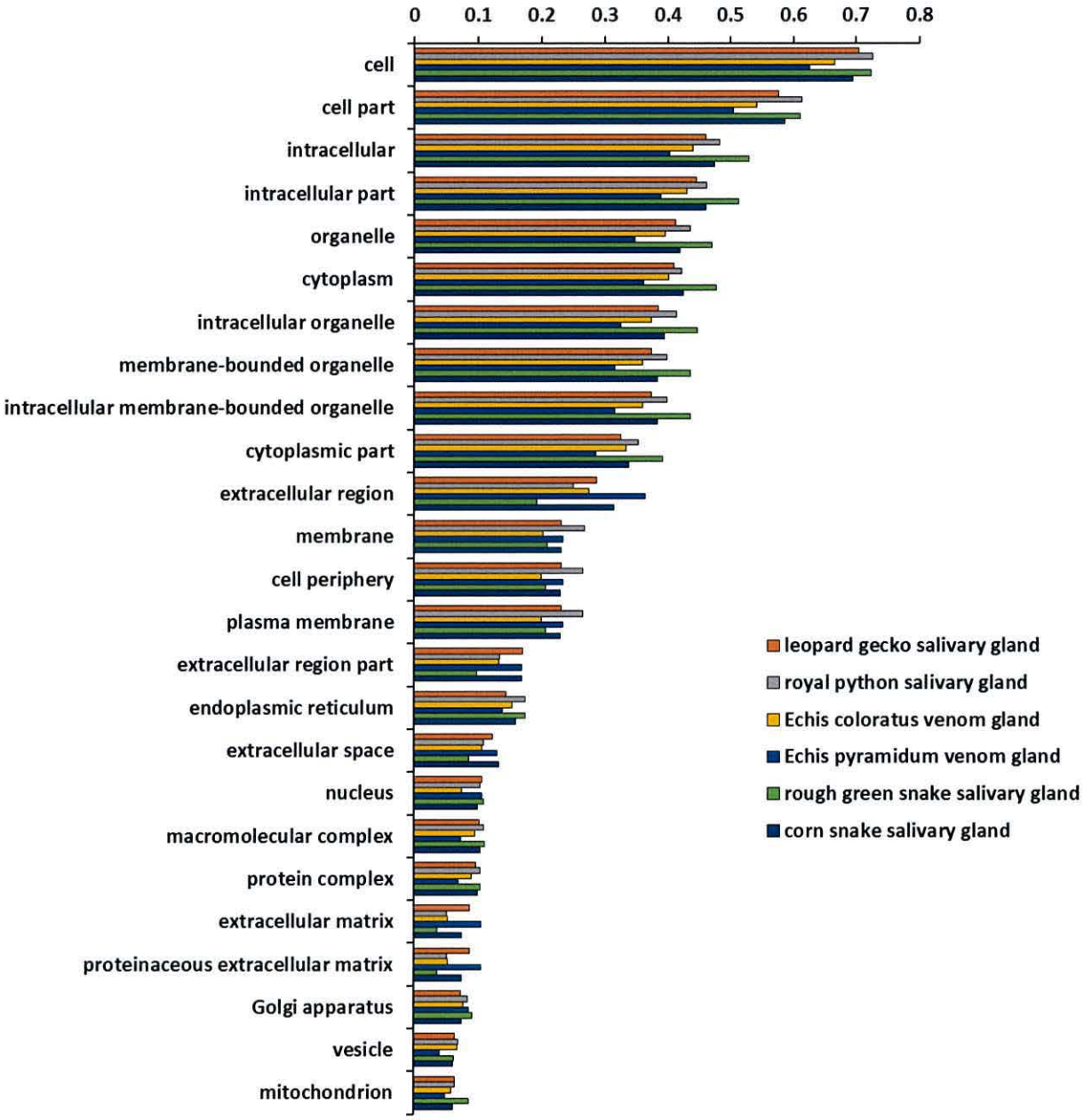


Figure 6.3. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Cellular component' terms for painted and Egyptian saw-scaled viper (*Echis coloratus* and *Echis pyramidum*) venom glands and corn snake (*Pantherophis guttatus*), rough green snake (*Opheodrys aestivus*), royal python (*Python regius*) and leopard gecko (*Eublepharis macularius*) salivary glands.

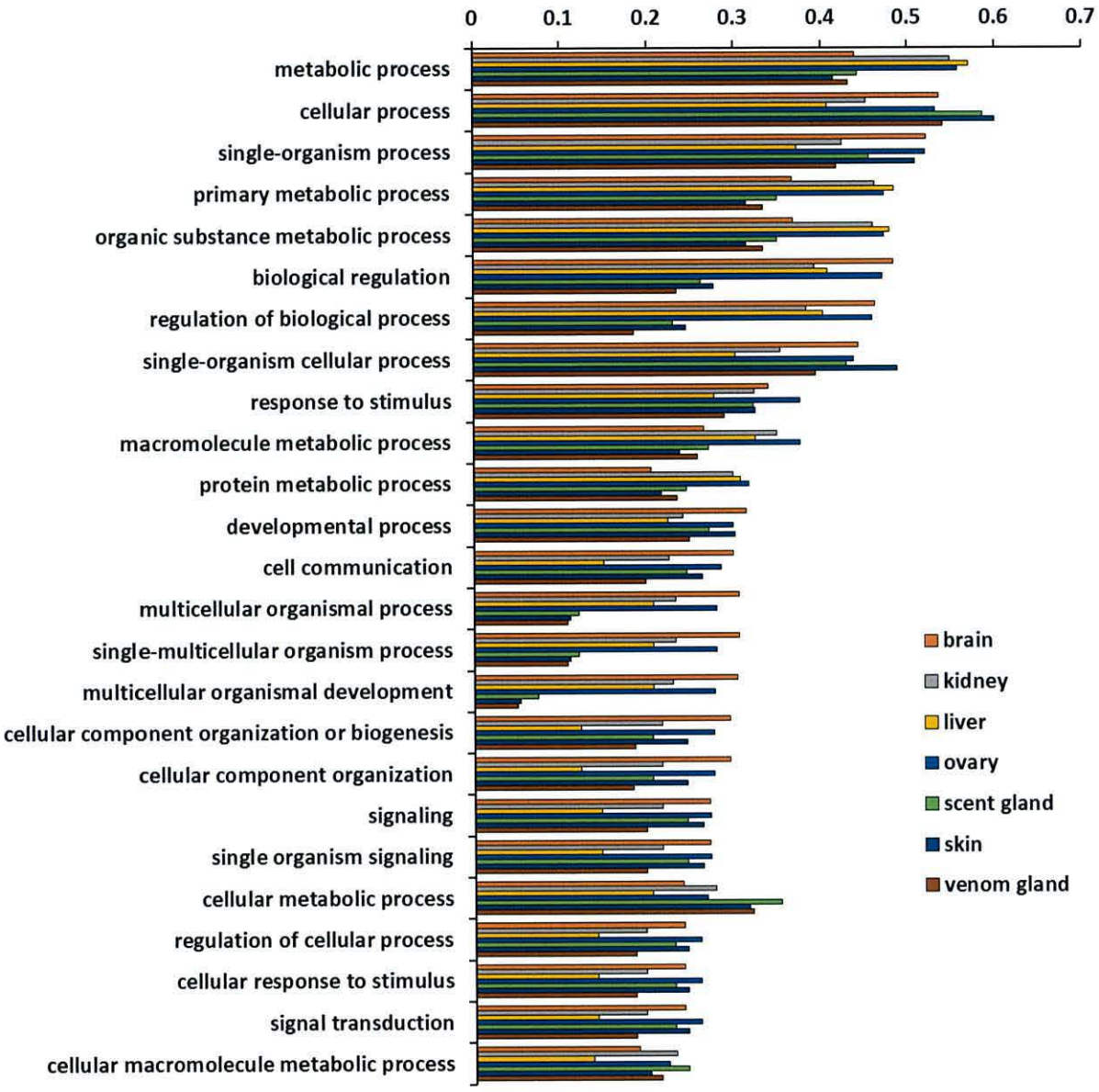


Figure 6.4. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Biological process' terms for painted saw-scaled viper (*Echis coloratus*) tissues.

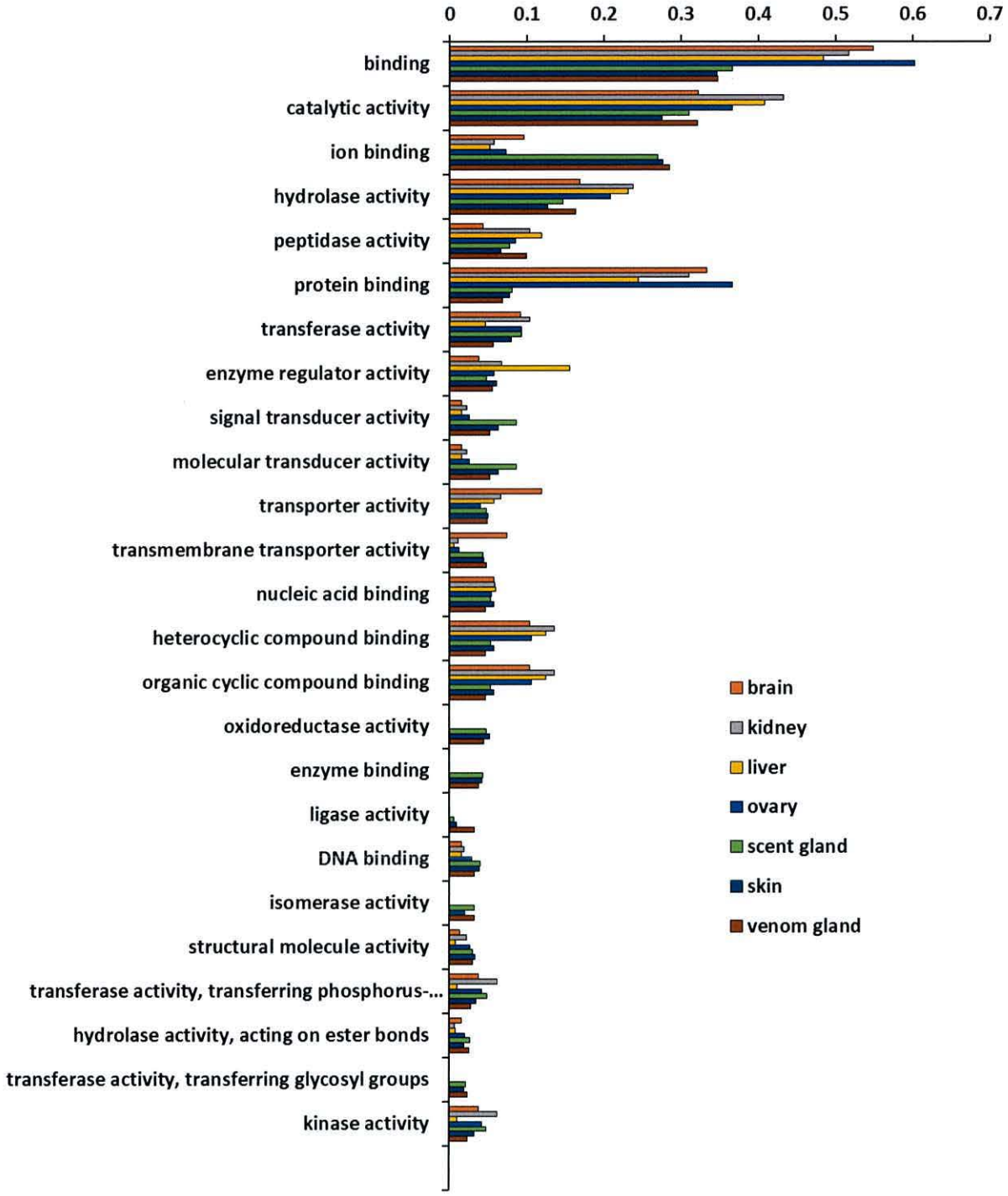


Figure 6.5. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Molecular function' terms for painted saw-scaled viper (*Echis coloratus*) tissues.

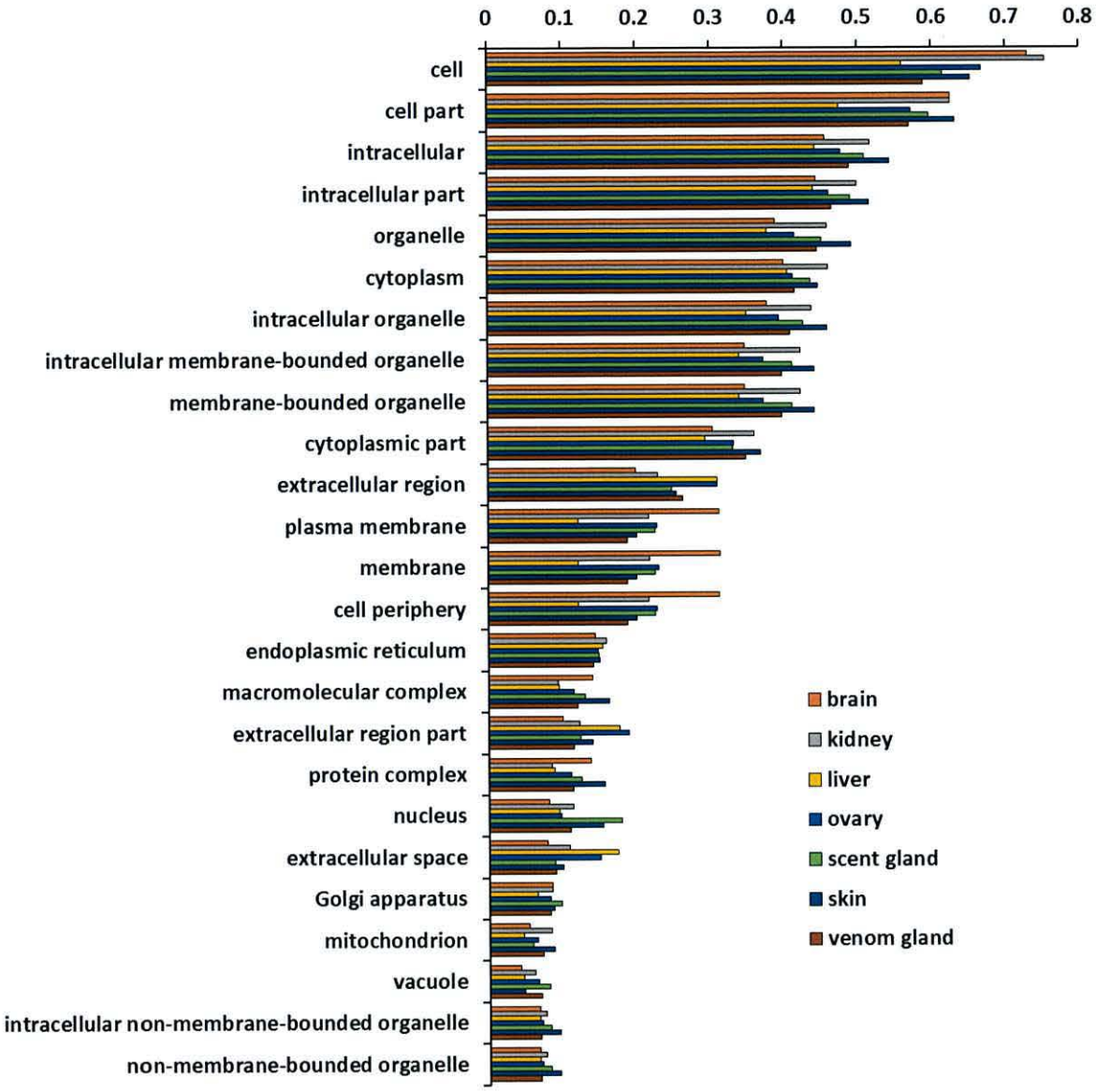


Figure 6.6. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Cellular component' terms for painted saw-scaled viper (*Echis coloratus*) tissues.

A global *E. coloratus* tissue assembly containing 147,966 contigs of ≥300bp was created and the tissue distribution of these transcripts was determined by mapping sequencing reads from each tissue to this assembly. 11,570 transcripts were found to be expressed in all 7 tissues (Figure 6.7) which suggests that these transcripts most likely represent ubiquitous maintenance or housekeeping genes common to all cells. Just 2,965 transcripts were found to be uniquely expressed in the venom gland, accounting for 8.27% of the total transcripts expressed in this tissue (far fewer than any other body tissues (Figure 6.7)), although it may be possible that highly expressed venom genes are drowning out some lowly-expressed transcripts.

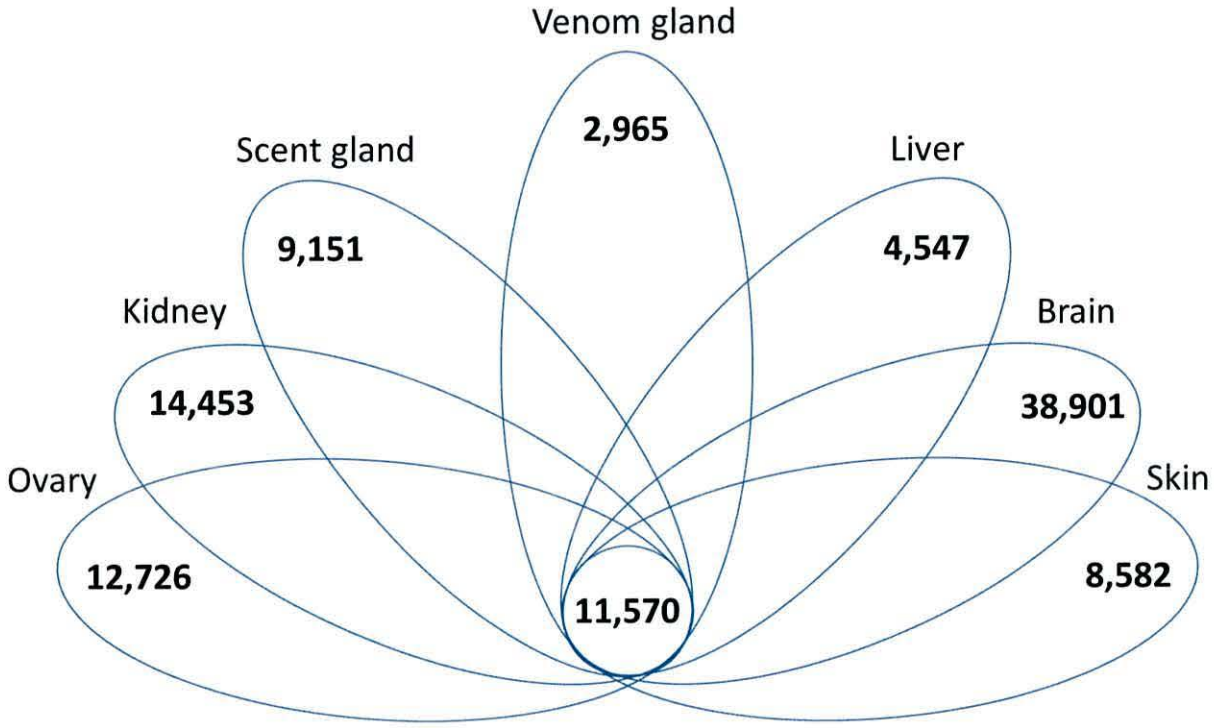


Figure 6.7. Tissue distribution of painted saw-scaled viper transcripts, determined by mapping sequencing reads derived from each tissue to a combined, all-tissue assembly with contig values of ≥1 FPKM (Fragments Per Kilobase of exon per Million fragments mapped) taken as evidence for expression. Figures represent the number of unique transcripts expressed in each tissue, with the number of transcripts expressed in all 7 tissues indicated in the centre.

Within these venom gland-specific transcripts were a number that encode members of known venom toxin gene families, including three contigs encoding group III SVMPs, eight contigs for both metalloproteinases and serine protease and four contigs encoding C-type lectins (see Figures 6.8-6.10 for full GO annotation graphs of these transcripts). Interestingly, the venom gland shares 26,181 transcripts in common with the scent gland, more than with any other body tissue (Table 6.1) and this may reflect their similar functions as secretory tissues.

Table 6.1. Number of shared expressed transcripts between the venom gland and other body tissues of the painted saw-scaled viper, *Echis coloratus*.

Tissues	Number of shared transcripts
Venom gland + Scent gland	26,181
Venom gland + Brain	25,172
Venom gland + Skin	24,427
Venom gland + Ovary	23,255
Venom gland + Kidney	22,854
Venom gland + Liver	15,433

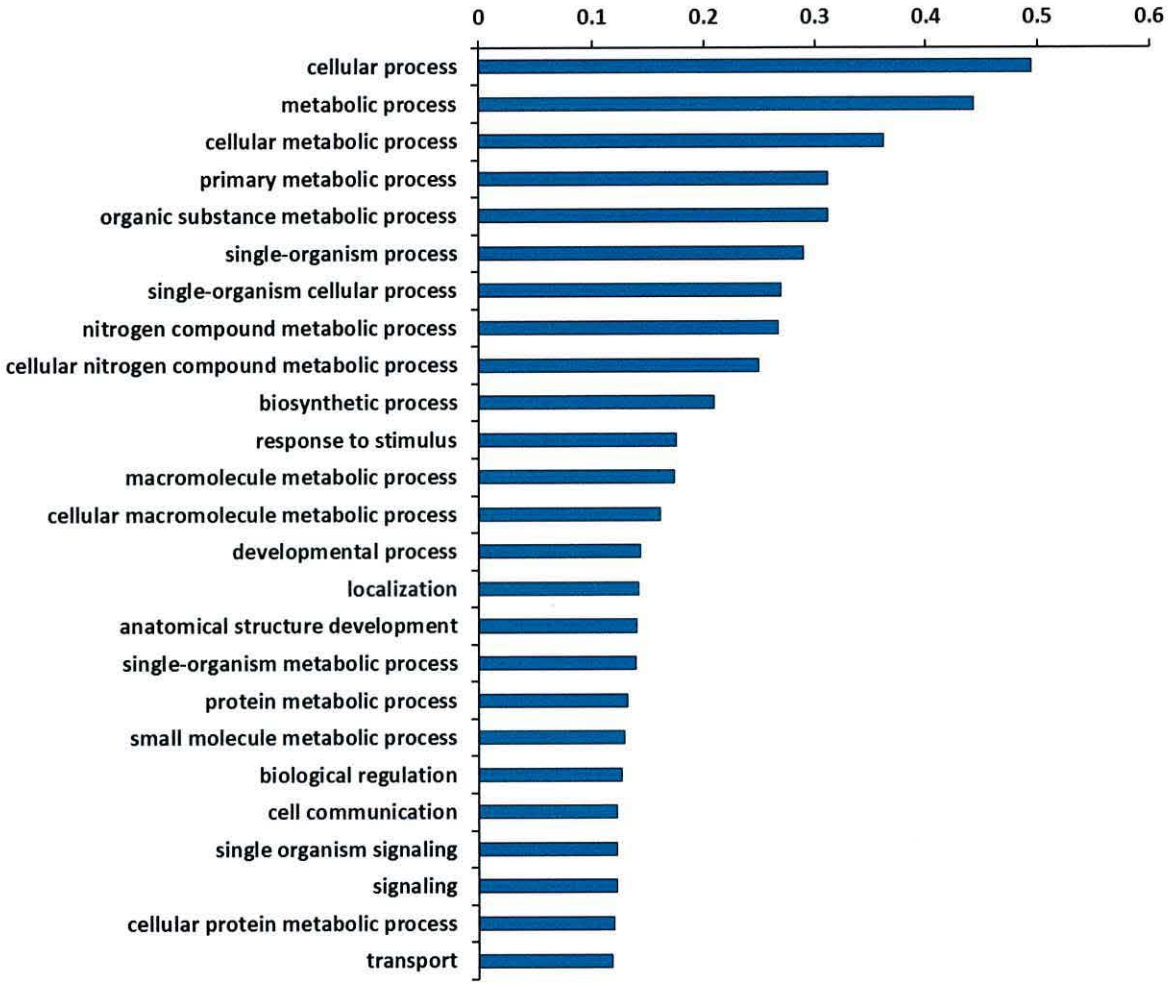


Figure 6.8. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Biological process' terms for expressed transcripts which are unique to the painted saw-scaled viper (*Echis coloratus*) venom gland compared to the remaining 6 body tissues.

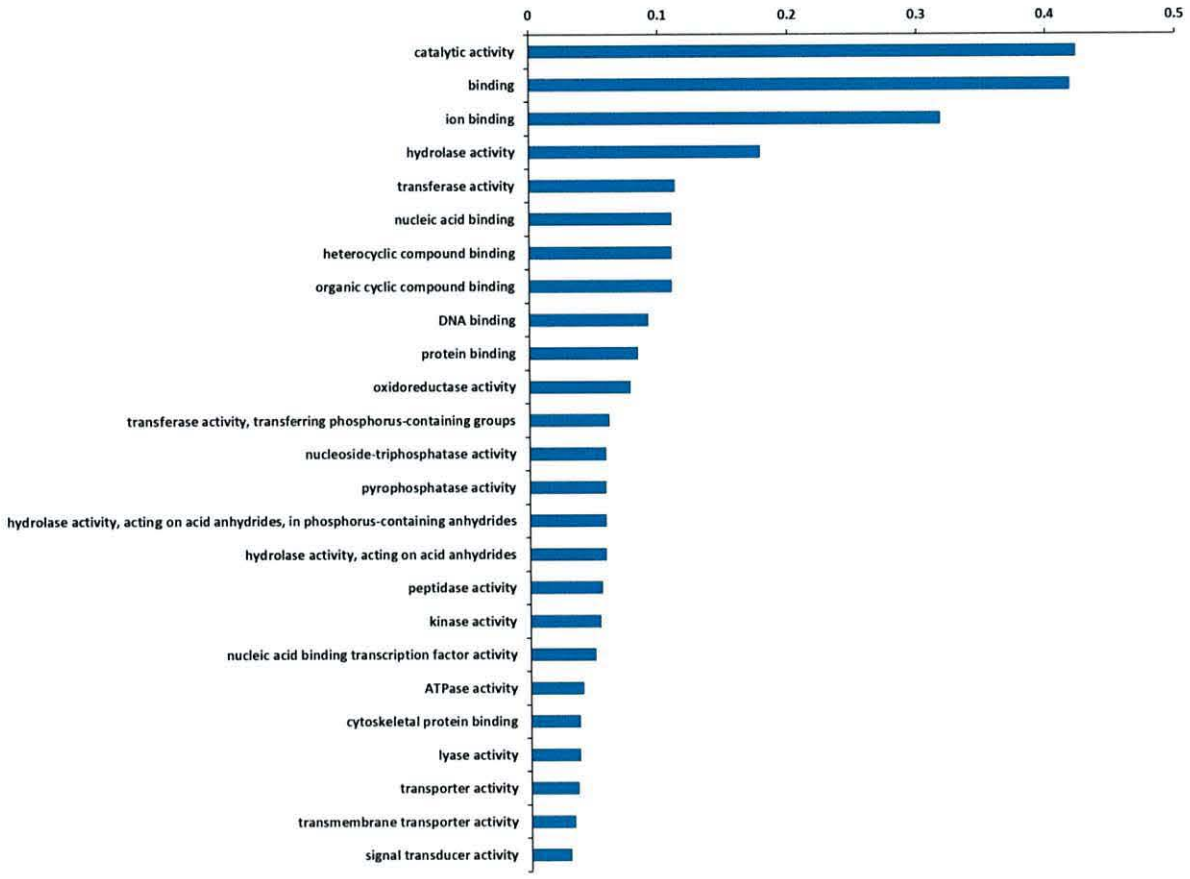


Figure 6.9. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Molecular function' terms for expressed transcripts which are unique to the painted saw-scaled viper (*Echis coloratus*) venom gland compared to the remaining 6 body tissues.

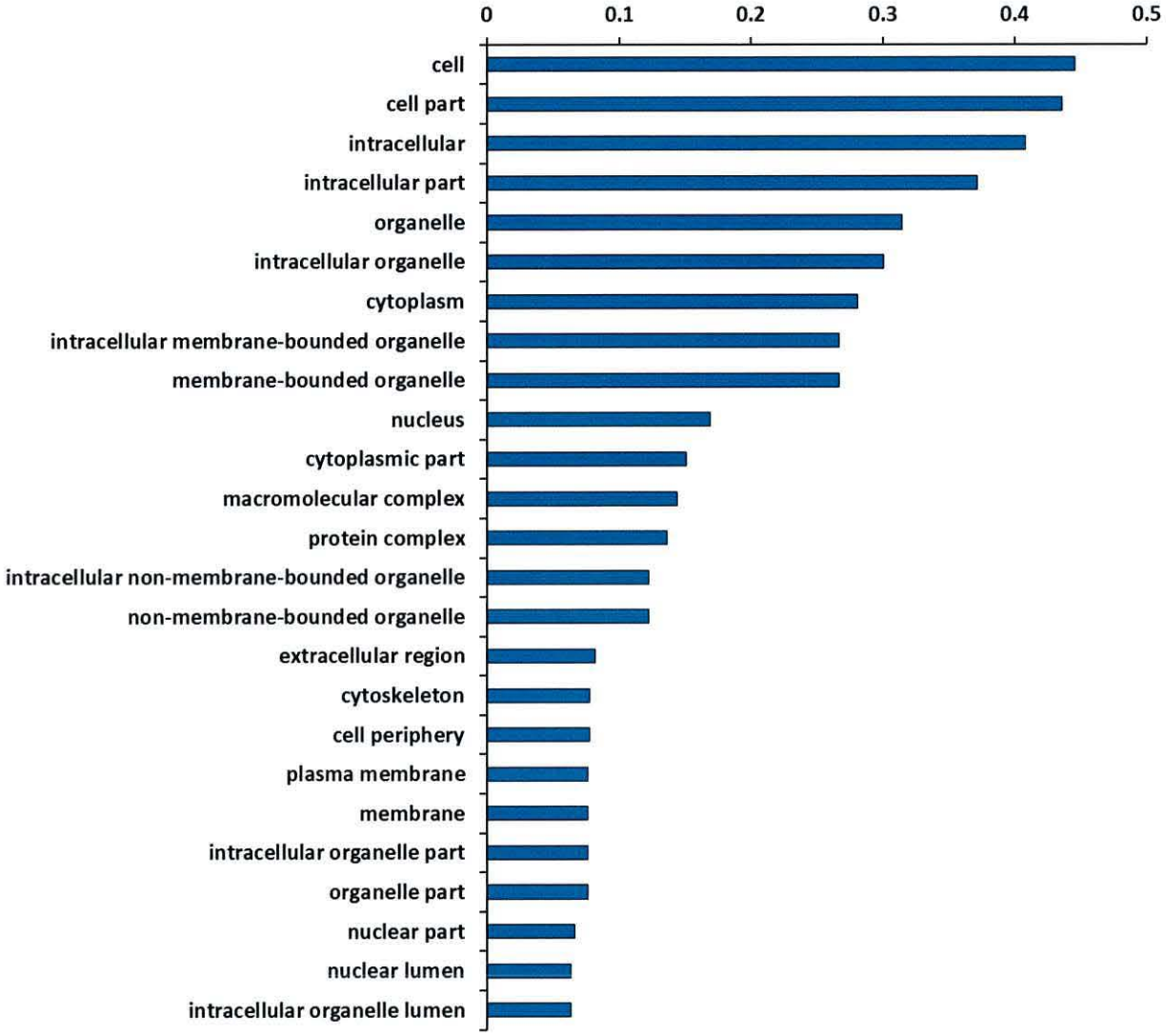


Figure 6.10. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Cellular component' terms for expressed transcripts which are unique to the painted saw-scaled viper (*Echis coloratus*) venom gland compared to the remaining 6 body tissues.

Venom gland samples for *Echis coloratus* were taken at different time points following manual venom extraction ("milking"); one sample approximately 16 hours post-milking, two samples at 24 hours and a final sample at 48 hours. 14,829 transcripts were common to all 4 venom gland samples, suggesting that these constitute the venom gland maintenance gene repertoire. However, there were also genes unique to each sample, with the 16 hour time point having 5,082 unique transcripts, the two 24 hour time points having 1,707 and 7,325 unique transcripts respectively, and the 48 hour time point having the highest number of unique transcripts with 12,535 (Appendix 29). Fishers exact tests revealed that the unique sequences expressed at 16 hours post-milking were mainly enriched for GO terms associated with transcription, such as "positive regulation of transcription from RNA polymerase II promoter", "transcription initiation from RNA polymerase II promoter" and "RNA polymerase II transcription factor binding transcription factor activity" as well as "histone H3-K4 methylation" which is a known histone modification at the promoter of genes which are being actively transcribed (Liang et al. 2004; Santos-Rosa et al. 2002; Schneider et al. 2003; Schubeler et al. 2004). However, several categories relating to post-translational modification were also enriched, such as "protein SUMOylation" and "peptidyl-serine phosphorylation". The GO terms "biosynthetic process" and "signaling" are also elevated compared to the other samples (Figure 6.11) as are "nucleic acid binding" and "protein binding" (Figure 6.12). At 24 hours, unique sequences are enriched for GO terms associated with translation, such as "translational initiation", "translational elongation", "structural constituent of ribosome" and "SRP dependent cotranslational protein targeting to membrane". The mRNA surveillance pathway, "nuclear-transcribed mRNA catabolic process, nonsense-mediated decay" is also enriched. The 48 hour time point unique sequences had no significantly enriched GO terms when compared to either of the 24 hour samples, but were enriched for "catalytic activity" and "hydrolase activity" compared to the 16 hour sequences. A local BLAST survey of the venom gland transcriptomes revealed all previously characterized venom genes in *Echis coloratus* (Hargreaves et al. 2014a, chapter 4) were present by 16 hours post-milking.

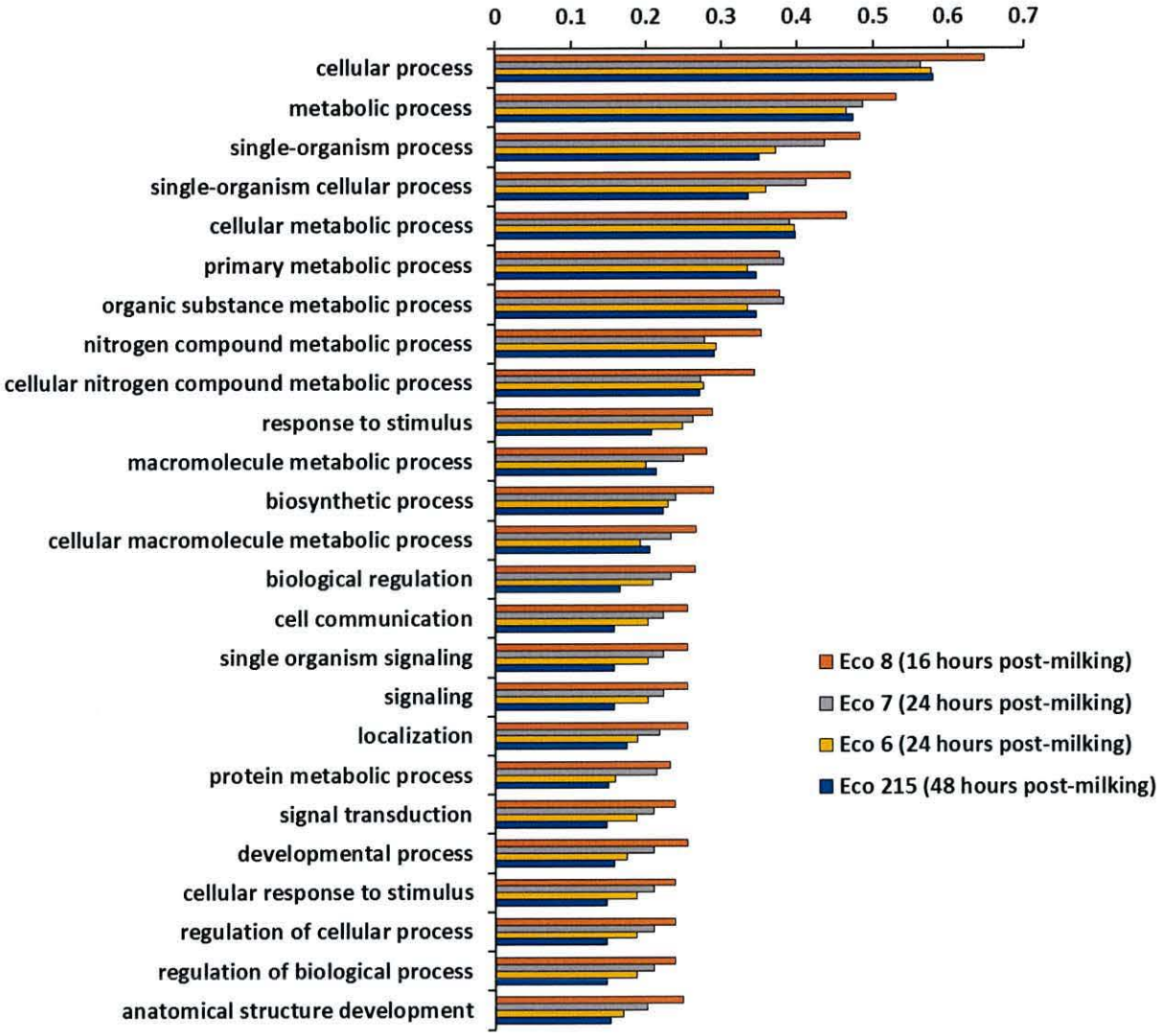


Figure 6.11. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Biological process' terms for transcripts which are unique to each individual timepoint following milking in the venom gland secretome of the painted saw-scaled viper (*Echis coloratus*).

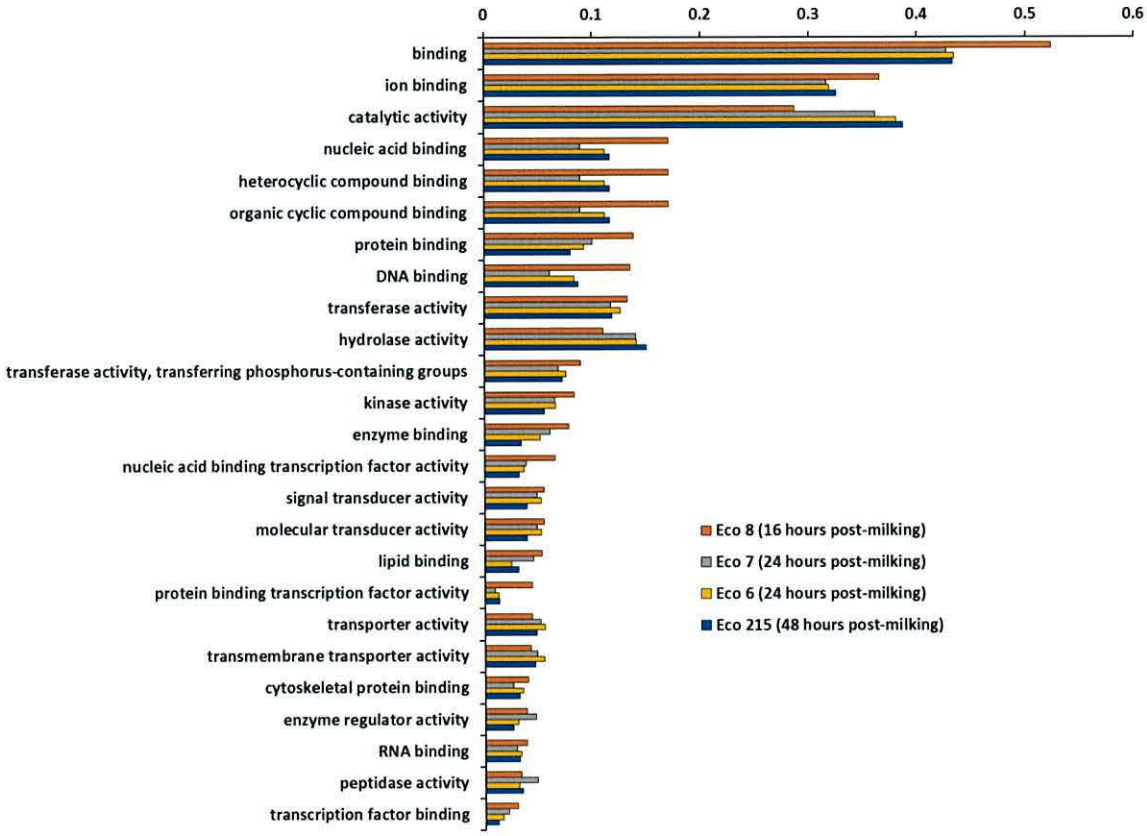


Figure 6.12. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Molecular function' terms for transcripts which are unique to each individual timepoint following milking in the venom gland secretome of the painted saw-scaled viper (*Echis coloratus*).

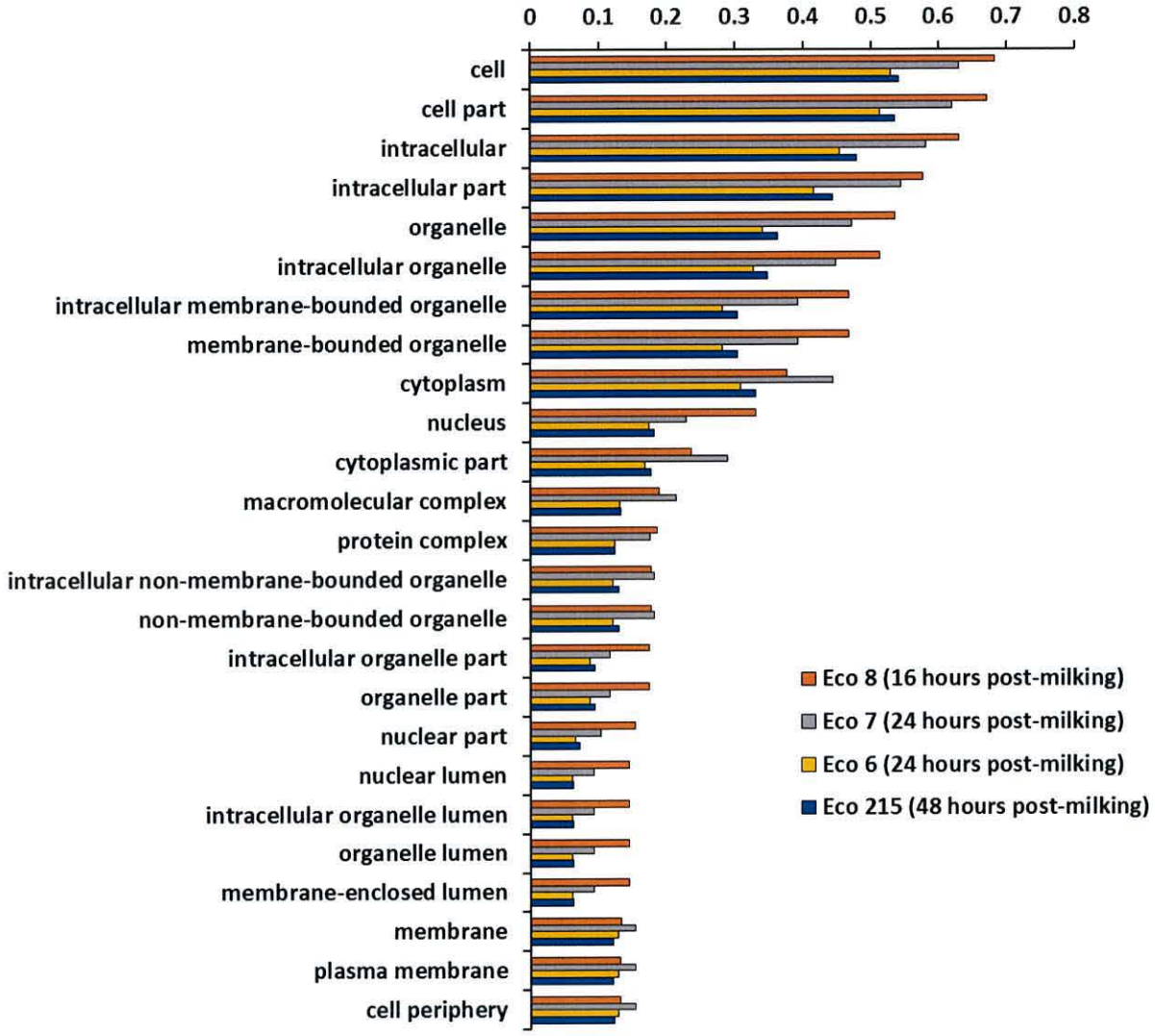


Figure 6.13. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Cellular component' terms for transcripts which are unique to each individual timepoint following milking in the venom gland secretome of the painted saw-scaled viper (*Echis coloratus*).

6.3.2 "Secretomics"

Between 3.72% and 10.64% of predicted open reading frames in the venom or salivary glands of the study species encoded a signal peptide and so are likely to be secreted (Appendix 26). Fishers exact tests show that the venom gland secretome of *E. coloratus* (based on pooled venom gland samples) is enriched for the GO terms "serine-type peptidase activity", "peptidase activity acting on L-amino acid peptides", "serine hydrolase activity" and "proteolysis" compared to the salivary gland secretomes of all non-venomous study species (Figures 6.1-6.3). As viper venom contains primarily proteases including serine proteases, and is also known to contain L-amino acid oxidase, these results support an increased amount of these components being expressed in the venom gland compared to the salivary gland of non-venomous species. Interestingly the *E. coloratus* venom gland was also enriched for "serine-type peptidase activity" and "serine hydrolase activity" in comparison to the venom gland of the eastern diamondback rattlesnake (*Crotalus adamanteus*), which may be indicative of interspecific variation in serine protease content in the venoms of these two species. When compared to the venom gland secretome of *E. pyramidum*, significant results were found only in the 16 and 48 hours post-milking samples, perhaps due to differences in the stages of the venom replenishment cycle between these samples and that of *E. pyramidum* venom gland (which was taken 24 hours post-milking), rather than any interspecific differences in gene expression. At 16 hours the venom gland secretome of *E. coloratus* was enriched for the GO terms "DNA binding", "chromatin remodeling", "nucleosome disassembly" and "transcription factor binding" compared to *E. pyramidum*. The GO categories "spliceosomal complex", "protein polyubiquitination", "protein transport", "L-amino acid oxidase activity", "serine-type endopeptidase inhibitor activity" and multiple categories for histone deacetylation were enriched in the 48 hours post-milking sample.

Sampling venom gland transcriptomes at different timepoints following milking allowed a comparison between venom gland secretomes at different stages of the venom replenishment cycle (Figures 6.14-6.16). At 16 hours post-milking, ion binding and transferase activity appear to be elevated compared to other timepoints (Figure 6.15), and the GO terms "protein phosphorylation", "transmembrane signaling receptor activity" and signal transducer activity" are significantly enriched compared to the remaining samples. The venom gland secretome 24 hours post-milking is significantly enriched for "protein ubiquitination", "protein modification by small protein conjugation or removal", "serine-type peptidase activity", "zinc ion binding", "peptidase activity" and "peptidase activity acting on L-amino acid peptides". Peptidase activity appears to be

considerably elevated in one 24 hour sample (Eco 7) but not in the other (Eco 6) (Figure 6.15) which may be suggestive of variation between individuals, but it may be more likely that this is a reflection of the difference in sequencing depth between these two samples. After 48 hours several GO categories related to cellular components are elevated (Figure 6.16). This timepoint was enriched for the highest number of GO categories, the majority of which were involved in histone deacetylation and chromatin modification. These include "chromatin modification", "Histone deacetylase activity (H3-K9 specific)", "protein deacetylase activity" and also "RNA splicing". The prevalence of histone deacetylation at this timepoint suggests a reduction in the rate of gene transcription, as this process is known to cause transcriptional repression and gene silencing through the modification of higher-order chromatin structure (Hui Ng and Bird 2000; Kouzarides 2007).

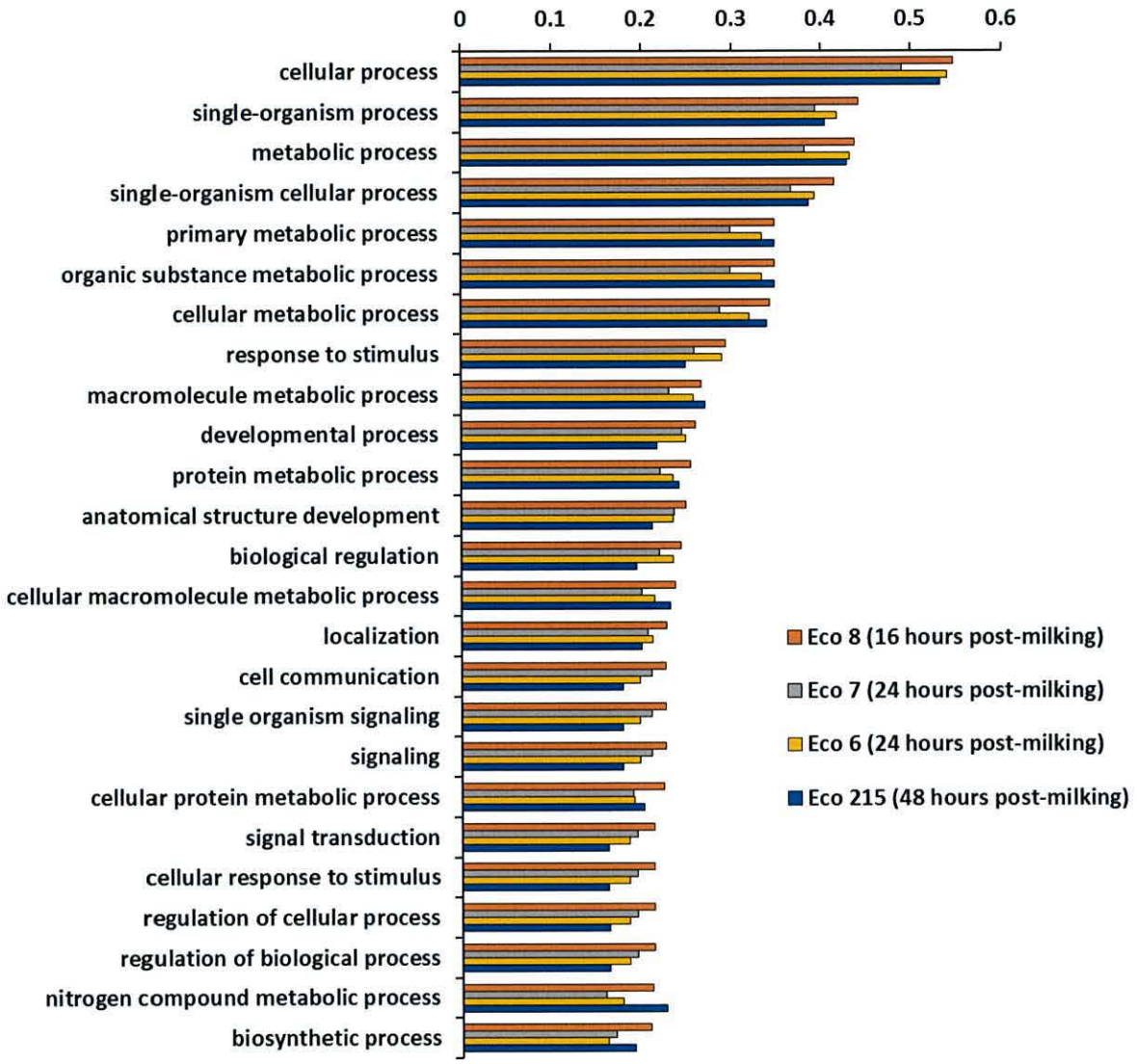


Figure 6.14. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Biological process' terms for painted saw-scaled viper (*Echis coloratus*) venom gland secretomes taken at different timepoints post-milking

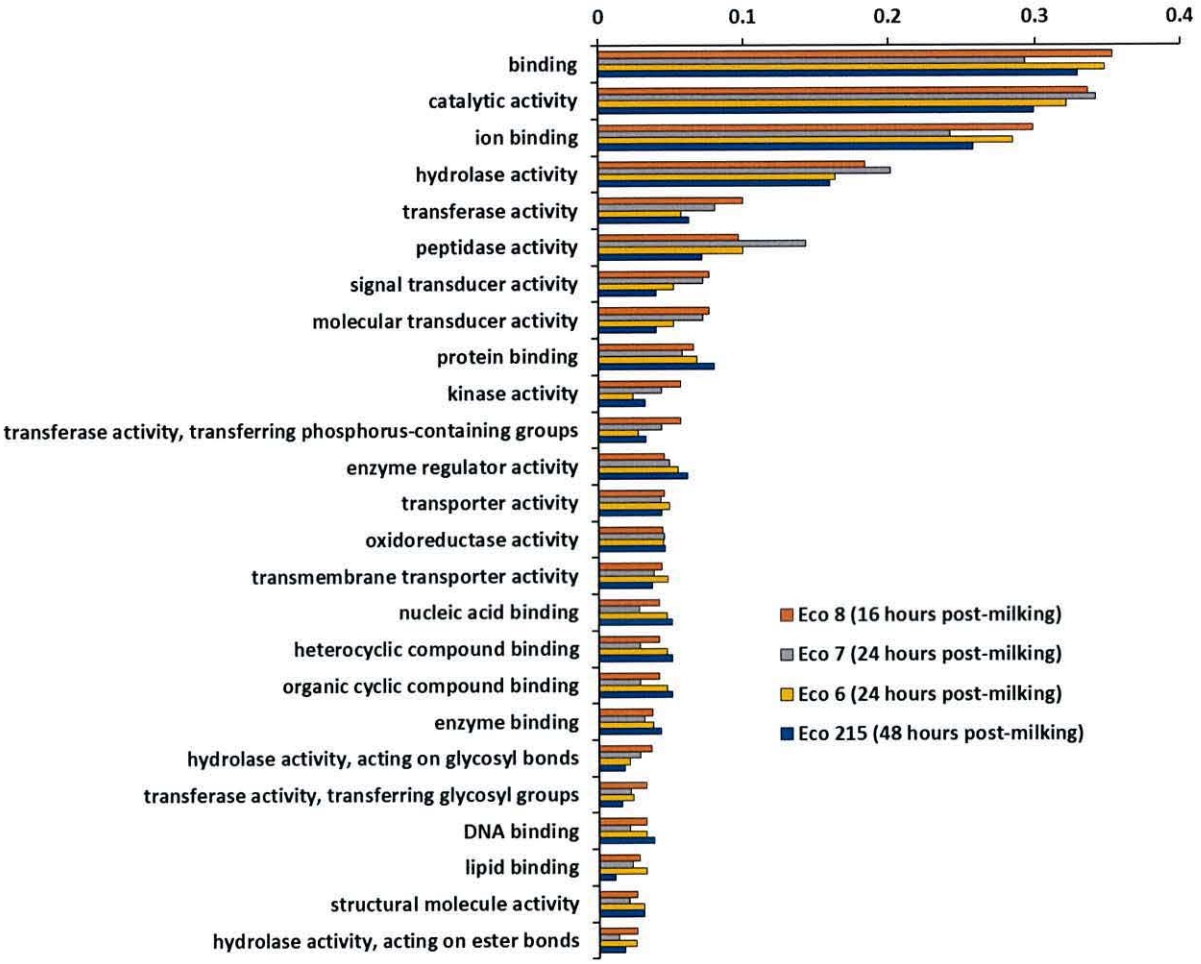


Figure 6.15. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Molecular function' terms for painted saw-scaled viper (*Echis coloratus*) venom gland secretomes taken at different timepoints post-milking

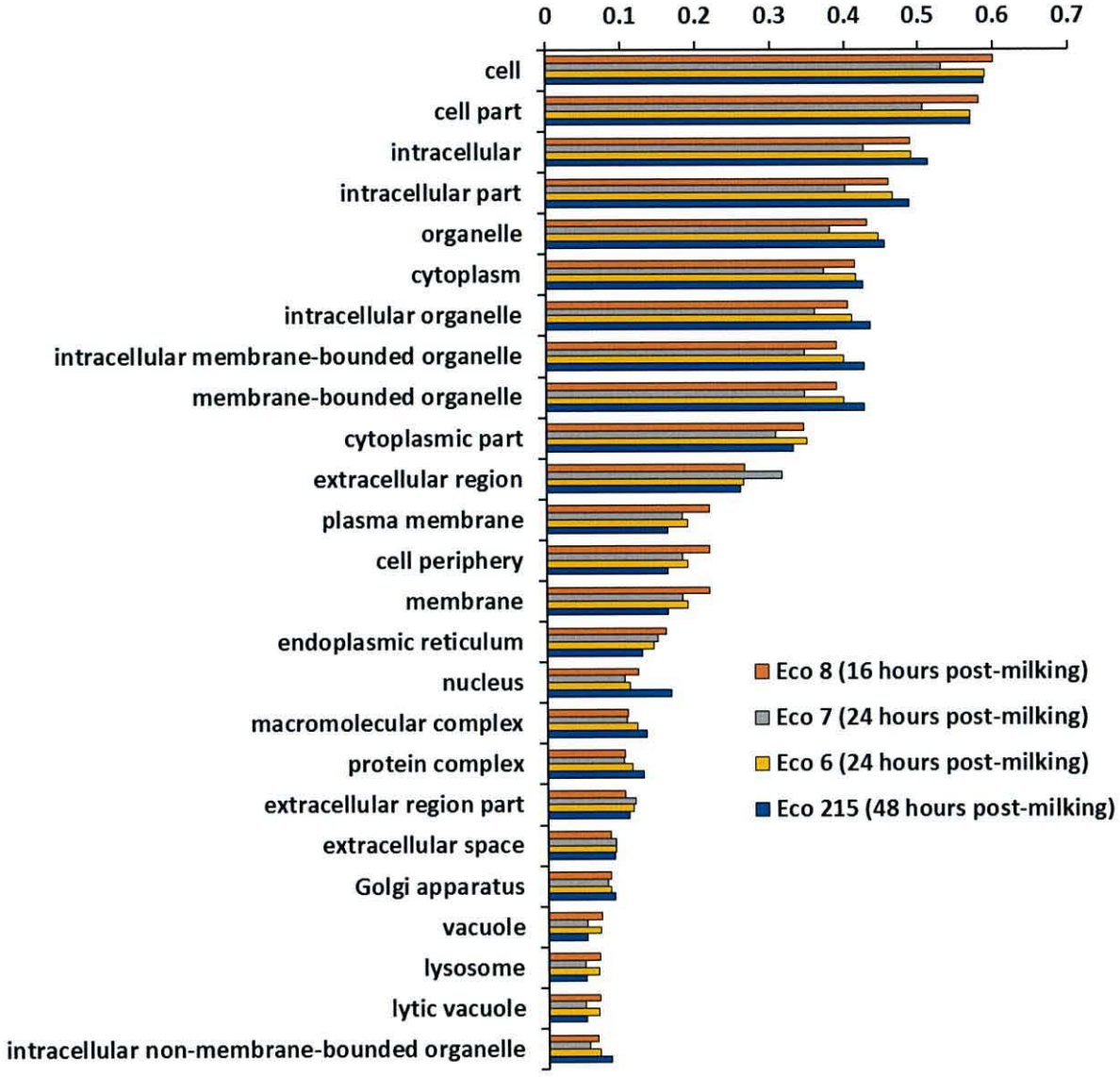


Figure 6.16. Proportion of transcripts assigned to each of the top 25 gene ontology (GO) slim 'Cellular component' terms for painted saw-scaled viper (*Echis coloratus*) venom gland secretomes taken at different timepoints post-milking

Finally, there was also some variation in the number of secreted transcripts belonging to toxin gene families over time, suggesting that their expression may show temporal differences during the venom replenishment cycle (Figure 6.17). In general the number of secreted transcripts appears to reduce towards the 48 hour post-milking timepoint (as can be seen for metalloproteinase, C-type lectin, CRISP, VEGF-F and serine protease in Figure 6.17). This may be a result of the reduced rate of transcription indicated by the increase in histone deacetylation as mentioned previously. The notable exception to this is L-amino acid oxidase, with a single transcript being identified in the 16 hours and both 24 hours post-milking samples, but ten transcripts being identified at 48 hours post-milking. On further inspection these transcripts in fact represent alternative splice variants which were previously characterized as *laao-b1* and *laao-b2* (Hargreaves et al. 2014a, chapter 4). The individual transcripts expressed at 16 and 24 hours post-milking encode *laao-b1*, which was found to be expressed in the venom gland of *E. coloratus* but also in the scent gland of royal python (*Python regius*), corn snake (*Pantherophis guttatus*) and rough green snake (*Opheodrys aestivus*) (i.e. all other snake species studied). The ten transcripts expressed 48 hours post-milking encode the variant *laao-b2* which was previously found to be venom gland-specific, and suggested to be a putative venom component in *E. coloratus* due to its tissue specificity and elevated expression level (Hargreaves et al. 2014a, chapter 4). Whilst more sampling (including earlier and possibly later timepoints) is required to confirm temporal expression differences, and it is likely that not all potential venom gene sequences have been included unless they are present as a full length ORF in these datasets, it is certainly suggestive that venom gene expression may show temporal variation following venom expenditure.

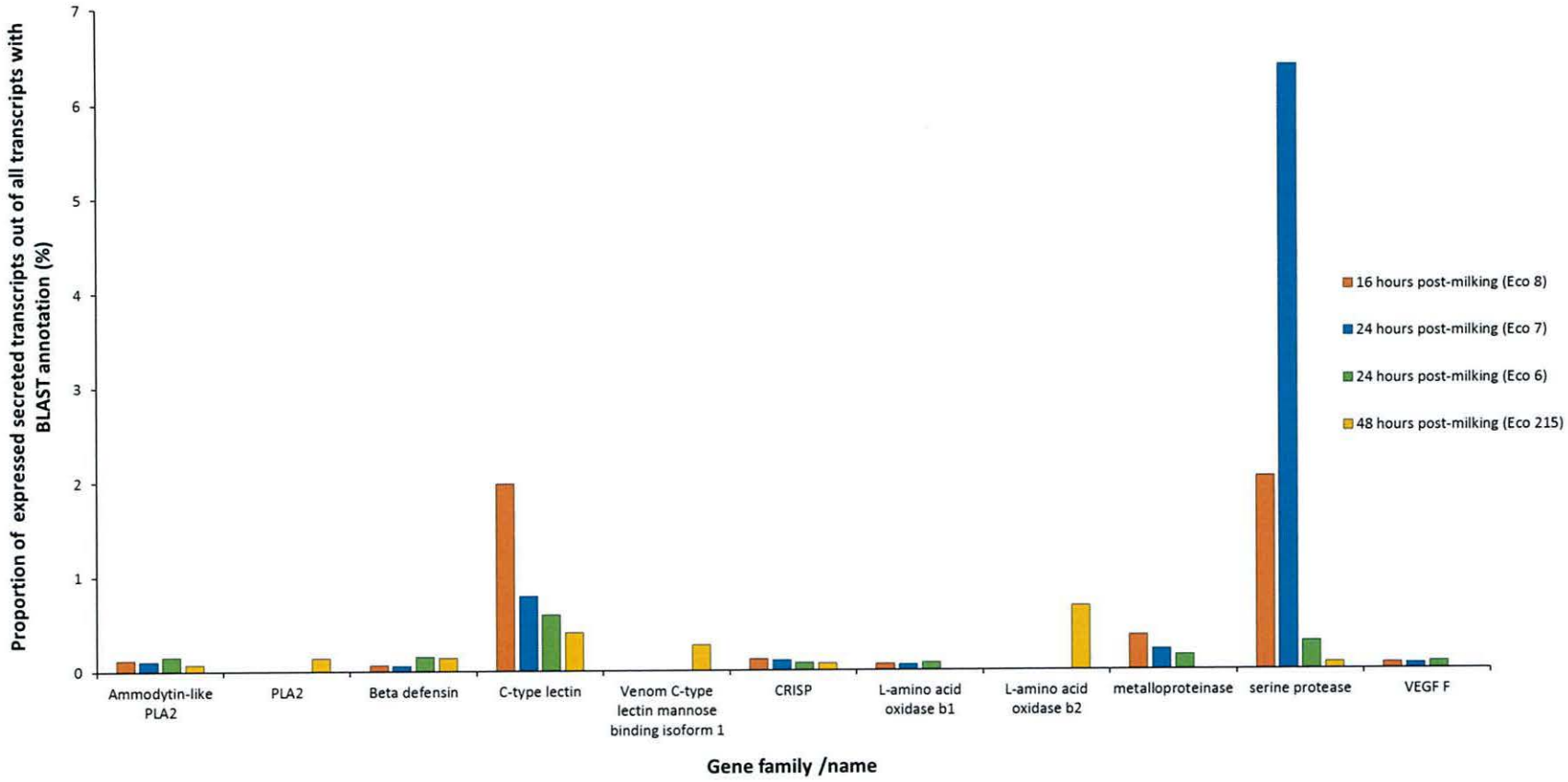


Figure 6.17. Proportion of secreted transcripts belonging to toxin families expressed in the venom gland of *Echis coloratus* at different timepoints post-milking.

When compared to other body tissues, the *E. coloratus* venom gland was found to be enriched for several GO categories including "peptidase activity", "peptidase activity acting on L-amino acid peptides", "serine-type peptidase activity" and "proteolysis" compared to brain; and for "protein modification by small protein conjugation or removal", "protein ubiquitination" and "ligase activity" when compared to brain, ovary and scent gland. There were no GO terms significantly enriched in the venom gland compared to the remaining body tissues. A more complete comparison of GO terms between the venom gland and other body tissues of this species is represented by Figures 6.4-6.6.

The venom gland of *E. pyramidum* was found to be enriched for "transforming growth factor beta receptor signaling pathway", "negative regulation of macroautophagy", "hedgehog receptor activity", "serine-type endopeptidase activity" and "hormone secretion" compared to all salivary gland secretomes and the venom gland secretome of *E. coloratus*. When compared to the 48 hours post-milking sample of *E. coloratus*, it was also enriched for "protein glycosylation", "metalloexopeptidase activity" and "cellular response to vascular endothelial growth factor stimulus".

In comparison to the venom gland secretomes of both *E. coloratus* and *E. pyramidum*, the venom gland secretome of the eastern coral snake (*Micrurus fulvius*) was found to be enriched for "arachidonic acid secretion" (arachidonic acid is a fatty acid released from a phospholipid molecule following hydrolysis by a phospholipase A_2 (Balsinde et al. 2002). It is a precursor of the eicosanoids (such as prostaglandins) which can exert a diverse array of effects such as platelet aggregation, vasodilation and smooth muscle relaxation (Harizi et al. 2008)), "calcium-dependent phospholipase A_2 activity", "activation of phospholipase A_2 activity" and "positive regulation of protein secretion". This result is complementary to the finding that the venom gland transcriptome of this species consists predominantly of PLA_2 transcripts (Margres et al. 2013) and that the toxicity of its venom is mainly due to PLA_2s (Vergara et al. 2014). No GO terms were found to be significantly enriched for the venom gland secretomes of king cobra or eastern diamondback rattlesnake when compared to the venom gland secretome of either *Echis* species.

6.3.3 Transcription factors

KEGG orthology (KO) analysis (Kanehisa et al. 2012) of the assembled transcriptomes identified between 255 and 358 transcription factors in the venom or salivary gland of the six

study species, with lower numbers (between 40 and 143) in the smaller king cobra, eastern coral snake and eastern diamondback rattlesnake venom gland datasets (Table 6.2).

Table 6.2. Numbers of transcription factors and components of signaling pathways found to be expressed in the venom (VG) or salivary gland (SAL). For completeness, data from Vonk et al. (2013) for the king cobra accessory gland (AG) was also included in the analysis.

	Transcription factors	Ca^{2+} signalling	MAPK	NFκB	TGFβ	VEGF
Echis coloratus (VG)	307	74	139	67	51	31
Echis pyramidum (VG)	342	93	153	73	56	32
Corn snake (SAL)	308	60	137	58	51	30
Rough green snake (SAL)	255	59	125	49	44	31
Royal python (SAL)	278	65	138	55	48	31
Leopard gecko (SAL)	358	80	147	64	53	31
Coral snake (VG)	40	12	43	7	14	15
Rattlesnake (VG)	143	32	79	23	31	22
King cobra (VG)	40	21	26	7	14	11
King cobra (AG)	182	24	34	8	14	11

197 transcription factors were conserved across the venom and salivary glands of all six study species, 77 of which are also found in all six additional *Echis coloratus* body tissues (ovary, liver, kidney, brain, cloacal scent gland and skin) and 71 of which are in five of the six tissues (Figure 6.18).

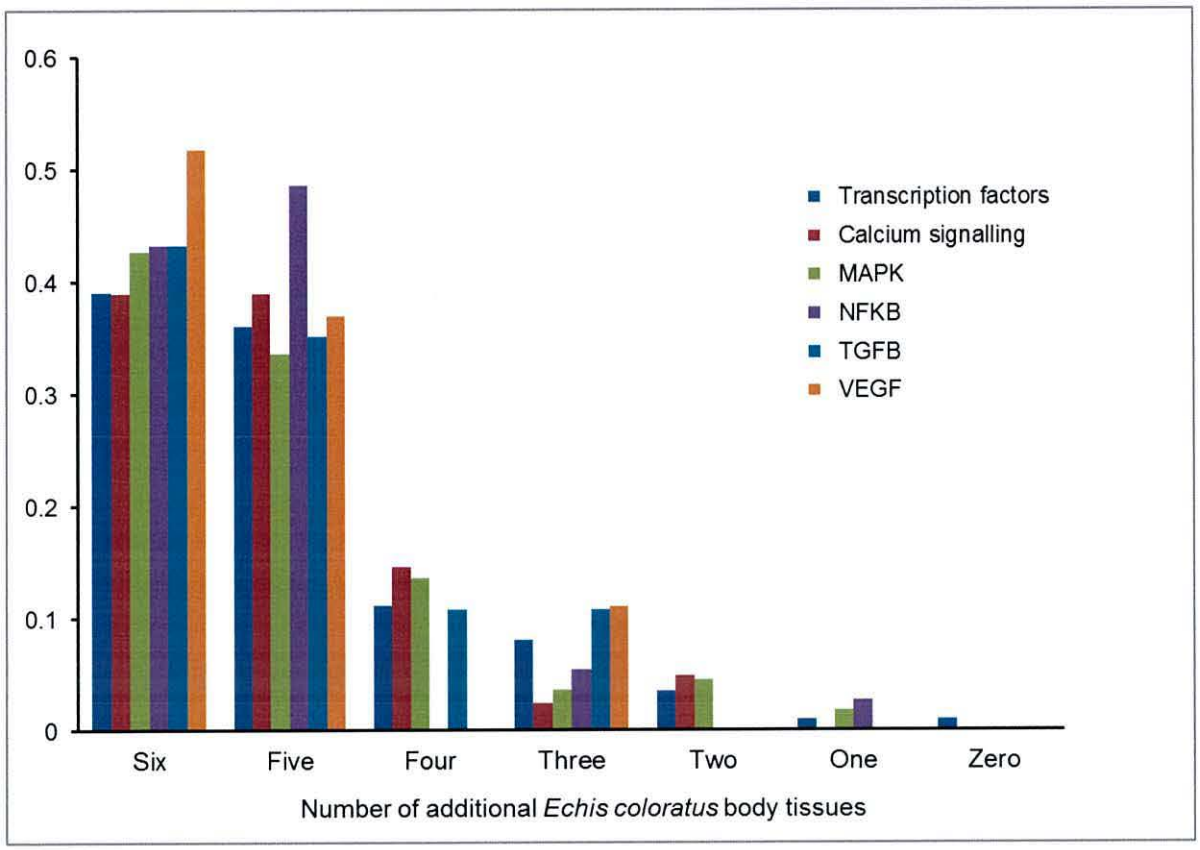


Figure 6.18. Proportions of transcription factors and members of signalling pathways found in the *Echis coloratus* venom gland that are also found in other body tissues.

These 148 widely-distributed transcription factors are therefore likely to represent members of the basal transcription machinery common to all cells, although they may of course still play a role in the regulation of venom production. Interestingly, there were only two transcription factors which appear to be unique to the venom and salivary glands based on the available data; *SAM pointed domain-containing Ets transcription factor* (*Spdef*) and *Forkhead box A1/hepatocyte nuclear factor 3a* (*FoxA1/Hnf3a*) and of these *Spdef* is also found in the venom glands of the king cobra, eastern coral snake and eastern diamondback rattlesnake, whilst *FoxA1/Hnf3a* is also present in the eastern diamondback venom gland. Two transcription factors were found only in the two *Echis* species; the homeodomain transcription factor *Ladybird* (*Lbx*) and *T-cell acute lymphocytic leukemia protein 1* (*Tal1*). Interestingly, no transcription factors were common to all venomous species and the non-venomous colubrids (corn snake and rough green snake), suggesting that the loss of venom in these species may be the result of the loss of gene regulatory components. The king cobra accessory gland transcriptome appears to contain a much higher number of transcription factors than the venom

gland (Table 6.2), with only 12 genes in common between the two. Only a single transcription factor appears to be unique to the venom gland of venomous snakes - the T-box transcription factor *Tbx3*, present in the venom gland of *Echis coloratus*, *Echis pyramidum*, eastern diamondback rattlesnake, king cobra and eastern coral snake, but not in the salivary gland of the leopard gecko, royal python, corn snake or rough green snake (transcripts encoding this gene are also found in four painted saw-scaled viper body tissues (kidney, ovary, scent gland and skin), but not king cobra pooled tissue or accessory gland).

It has previously been suggested that the transcription factors NF1, NFκB and AP-1 are involved in the regulation of snake venom production (Luna et al. 2009). Of the four known members of the nuclear factor 1 (Nf1) gene family (NFIA, NFIB, NFIC and NFIX (Kruse et al. 1991; Rupp et al. 1990)), only NFIA and NFIB were found to be expressed in the venom or salivary gland of all six study species. The NFκB family comprises five related transcription factors (RelA (p65), RelB, c-Rel, p50 (p105 precursor), p52 (p100 precursor)) that are able to form homo- or heterodimers via the shared Rel homology region (Hayden and Ghosh 2012; Napetschnig and Wu 2013). Of these, only RelA (p65), RelB, c-Rel encode a transactivation domain at their C-terminus and so are able to activate transcription as homodimers. The activity of NFκB is regulated via binding to the members of the IκB inhibitor protein family and sequestration of the resulting complex in the cytoplasm (Hayden and Ghosh 2012; Napetschnig and Wu 2013; Sen and Baltimore 1986). Transcripts encoding RelA (p65), RelB, p50 (p105 precursor), p52 (p100 precursor) are expressed in the salivary or venom gland of all six study species and c-Rel in only Egyptian saw-scaled viper venom gland and corn snake salivary gland. Transcripts encoding IκBα, IκBβ and IκBε inhibitors were found in all species. Activator protein 1 (AP-1) typically functions as a heterodimer of members of the Jun and Fos families of basic leucine zipper transcription factors (comprising c-Jun, JunB and JunD and c-Fos, FosB, FOS-like antigen 1 (FOSL1) and FOS-like antigen 2 (FOSL2)) (Curran and Franza Jr 1988; Karin et al. 1997). c-jun, junB and JunD and c-Fos and FOSL2 were detected in the venom or salivary gland of all six study species. Finally, *Sp1* binding sites have previously been identified upstream of the South American rattlesnake *crotamine* gene (Rádis-Baptista et al. 2003) and transcripts encoding this gene were also found in the venom or salivary gland of all six study species.

6.3.4 Signaling

The link between AP-1, NFκB and MAP kinase (ERK1/2) signalling has long been known (Hommes et al. 2003; Karin et al. 1997) and components of these pathways have previously been implicated in the regulation of venom production following milking in *Bothrops jararaca* (Kerchove et al. 2008; Luna et al. 2009), as has a role for calcium signalling (Kerchove et al. 2008), which is also known to be involved in salivary secretion (Melvin et al. 2005; Putney Jr 1986). The Sp1 transcription factor is known to interact with the Transforming Growth Factor-β (TGF-β) signalling pathway and members of the vascular endothelial growth factor (VEGF) gene family have been characterised from the venom gland of a diversity of reptile species (Aird et al. 2013; Francischetti et al. 2004; Margres et al. 2013; Yamazaki et al. 2009). Analysis of members of these pathways in the salivary and venom gland transcriptomes (Table 6.2) revealed 110 transcripts involved in MAP kinase (ERK1/2) signalling that were conserved across the venom or salivary gland of all study species, 37 transcripts involved in NFκB signaling and an identical number involved in the TGF-β pathway, 41 transcripts involved in Ca^{2+} signaling and 27 in VEGF signaling. Total numbers of transcripts involved in each pathway were broadly similar across study species, and slightly lower in the king cobra, eastern diamondback and eastern coral snake (Table 6.2). Interestingly, there was no evidence for venom-gland specific members of these signaling pathways, although RelB (involved in both the NFκB and MAPK pathways) and mitogen-activated protein kinase kinase kinase 2 (MAP3K2) from the MAPK pathway are expressed in only one of the six additional painted saw-scaled viper body tissues, the scent gland (Figure 6.18).

6.3.5 Adrenoceptor signaling

It has previously been suggested that α- and β-adrenoceptors (adrenergic receptors) may have a role in the activation of venom production following milking in *Bothrops jararaca* (Kerchove et al. 2004; Kerchove et al. 2008; Yamanouye et al. 2000) and a transcriptomic survey of the related *Bothrops alternatus* venom gland during venom synthesis three days post-milking suggested the presence of a conserved α_{1D} adrenoceptor (clone BACCGV4069B12, accession GW581578). However, re-analysis of the *B. alternatus* data shows that this sequence in fact encodes a different G protein coupled receptor, most likely a chemokine-like receptor. Nevertheless, several different adrenoceptors were identified in the generated transcriptomic data (Table 6.3), although there was no evidence to support a conserved adrenoceptor in the

venom gland of venomous snakes. No adrenoceptor-like sequences were detected in the venom glands of the king cobra, rattlesnake or coral snake and this, together with the lack of these sequences in previously published transcriptomes, suggests that a high sequencing depth is required to detect these transcripts. The apparent paucity of adrenoceptor transcripts in the newly generated data does not rule out their role in the regulation of venom production however, as it is possible (indeed likely) that these receptors are transcribed and translated prior to the synthesis of venom and therefore earlier than the time points included in previous studies and this analysis.

Table 6.3. Presence of α and β adrenoceptor (adrenergic receptor) transcripts in reptile venom (VG) and salivary (SAL) glands. Eco, painted saw-scaled viper (*Echis coloratus*); Pgu, corn snake (*Pantherophis guttatus*); Oae, rough green snake (*Opheodrys aestivus*); royal python (Python regius); Ema, leopard gecko (*Eublepharis macularius*). Presence is denoted by a "+", absence by a "-".

	Species				
	Eco (VG)	Pgu (SAL)	Oae (SAL)	Pre (SAL)	Ema (SAL)
α1a	+	+	-	-	+
α1b	-	-	-	-	-
α1c	-	-	-	-	-
α1d	-	-	-	-	-
α2a	-	-	-	+	+
α2b	+	-	+	-	-
α2c	-	-	-	-	-
α2d	-	-	-	-	-
β1	+	-	-	+	+
β2	-	+	-	-	+
β4c	-	+	+	+	-

6.3.6 Transcription factor binding sites analysis

The genome sequence for *E. coloratus* (Full assembly metrics available in Appendix 30) was interrogated for genes known to encode venom toxins and the upstream regions of these genes (where data were available) were compared with other members of the same gene family within this species and to related genes in other species using previously published data.

Snake venom metalloproteinases

The *Echis coloratus* genome assembly contains 15 scaffolds encoding the first exon of one of the many snake venom metalloproteinase gene variants (SVMP, a group of metalloproteinases thought to be most closely related to the A Disintegrin And Metalloproteinase family of peptidases (Casewell 2012; Jia et al. 1996)). Of these, eight had typically only a few hundred base pairs of sequence upstream of the start codon and a longer sequence had a string of N's adjacent to the start codon. These 9 sequences were eliminated form subsequent analyses. The remaining seven contigs had ≥1000bp of sequence upstream of the start codon and provided 2,533bp of aligned sequence (based on the length of the shortest aligned sequence) for transcription factor binding site analysis. Although present in all sequences, little conservation of predicted binding sites for candidate transcription factors such as AP-1, AP-2, Sp1, NF1, NFκB, Tbx3, Hnf3a is found across most of these regions, although a roughly 500bp region immediately upstream of the start codon does contain sites conserved across several (NF1 and NFκB) or all (Nkx2.5, Tef1, Barbie, ETS) sequences (Figure 6.19), suggesting that this proximal region may be most important in controlling gene expression.

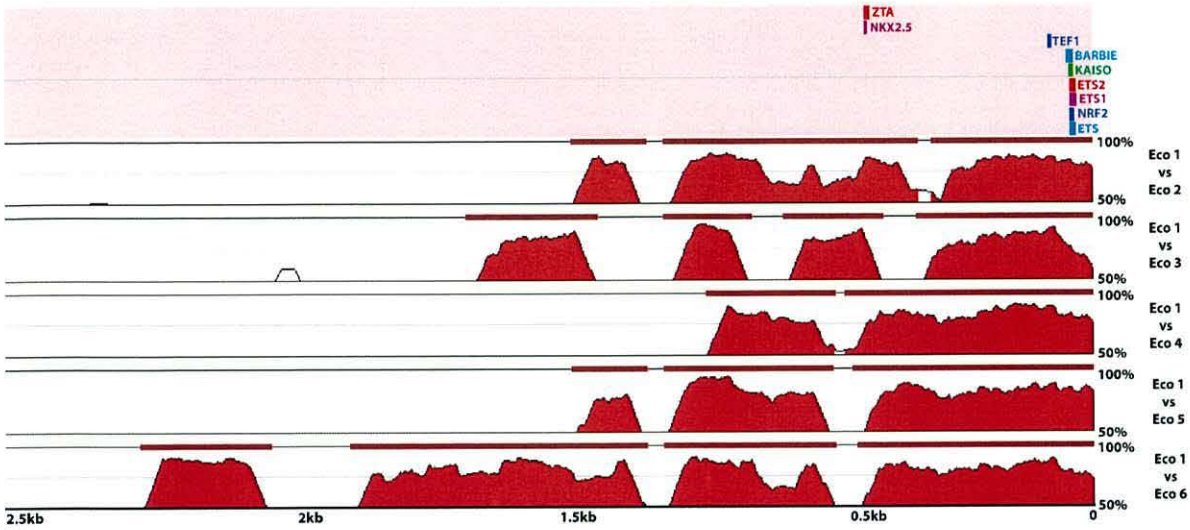


Figure 6.19. Transcription factor binding site analysis of 2,533bp of upstream sequence of six *Echis coloratus* (Eco) snake venom metalloproteinase (SVMP) genes. Position 0 denotes the first base of the start codon. The locations and approximate length of the nine transcription factor binding sites found to be conserved across all six sequences are shown in the top panel and pairwise sequence similarity plots are shown in the bottom panels, with peaks over 75% conserved colored red. Conserved transcription factor binding sites and the highest levels of sequence conservation are seemingly restricted to a region proximal to the transcription start site.

Serine proteases

There are eight scaffolds in the *E. coloratus* genome assembly that encode the first exon of a serine protease gene and it was possible to generate an alignment of at least 1500bp for six of these. A number of conserved transcription factor binding sites were found within approximately 500bp of the start codon of all of these genes (Figure 6.20), although not for any of the candidate transcription factors.

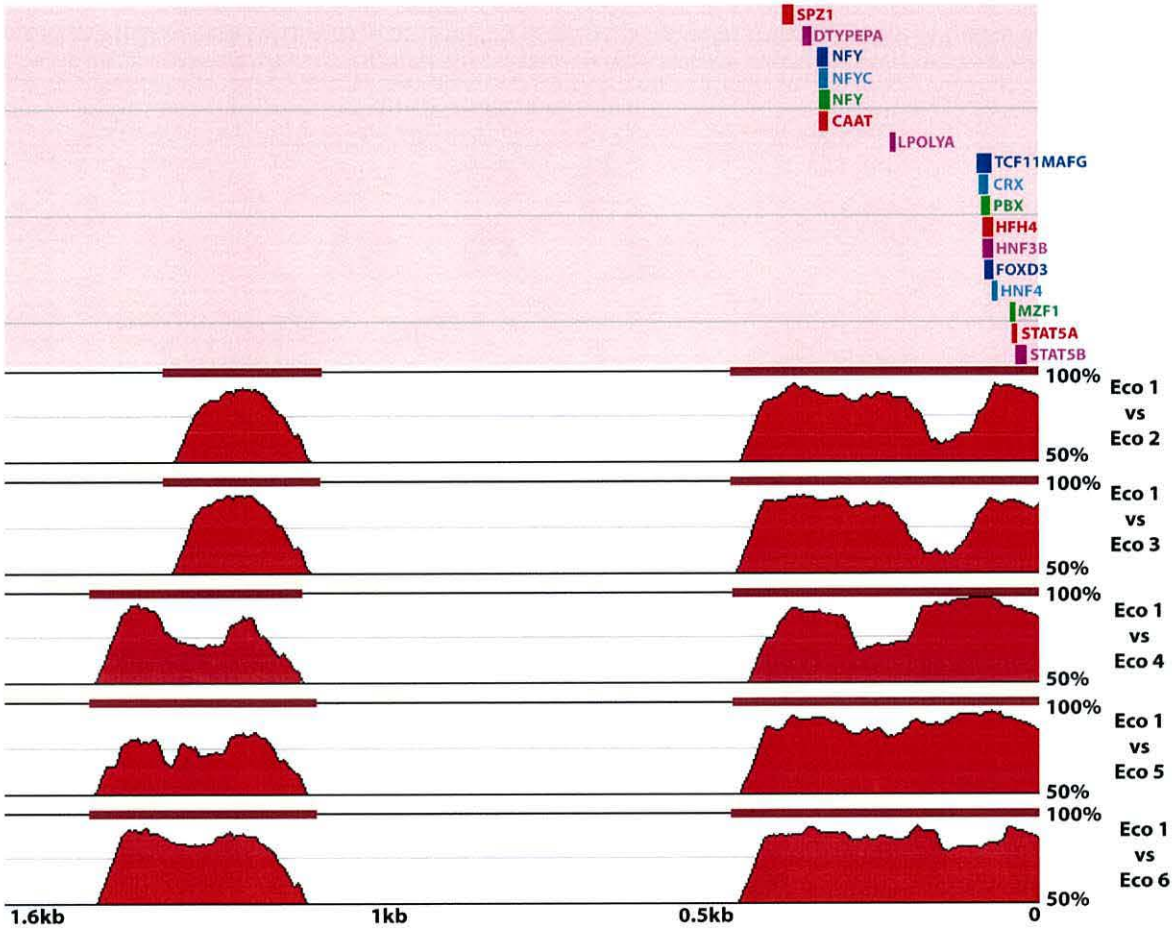


Figure 6.20. Transcription factor binding site analysis of 1,624bp of upstream sequence of six *Echis coloratus* (Eco) serine protease genes. Position 0 denotes the first base of the start codon. The locations of the seventeen transcription factor binding sites found to be conserved across all six sequences are shown in the top panel and pairwise sequence similarity plots are shown in the bottom panels, with peaks over 75% conserved colored red. As for SVMP genes (Figure 6.19), conserved transcription factor binding sites are seemingly restricted to a region proximal to the transcription start site.

C-type lectins

The *E. coloratus* genome assembly contains thirteen scaffolds that encode the first exon of a C-type lectin gene. Six of these scaffolds contain less than 150bp of sequence upstream of the start codon and one contained a long string of N's immediately adjacent to the start codon and these seven sequences were therefore excluded from further analysis. The remaining six sequences provided 402bp of aligned sequence and although each contains putative binding sites for candidate transcription factors there are no conserved sites shared by all loci. However,

analysis of mRNA sequences reveals that transcripts expressed in the venom gland have a 12 nucleotide insertion in their 5' UTR which is not present in the C-type lectin transcript expressed in the scent gland of this species. No transcription factor binding sites were found within this region, although the first 3 nucleotides of the insertion have led to the formation of a binding site for the transcriptional activator MYB.

Other venom genes

Two scaffolds in the *E. coloratus* genome assembly encode Group IIA Phospholipase A2 (PLA_2) genes and these provided 292bp of aligned sequence. The upstream regions of these two genes are highly identical and contained 56 conserved transcription factor binding sites, including sites for Spz1, Tbx, NF1, Mzf1, Sp1, HFH4. Two scaffolds in our assembly encode CRISP genes and these provided 2.2kb of upstream sequence for analysis, resulting in the identification of 136 conserved transcription factor binding sites. It seems likely that the level of sequence conservation between both PLA_2 and CRISP paralogs is too high for the accurate prediction of putative transcription factor binding sites in this species. Finally, the *E. coloratus* genome sequence contains an 8.3kb scaffold that encodes the full coding sequence of the *vascular endothelial growth factor f* (*vegf-f*) gene (a viper-specific paralog which appears to be venom gland-specific (Hargreaves et al. 2014a, Chapter 4)), including 2.4kb of sequence upstream of the transcription start site. This upstream region contains putative transcription factor binding sites for AP-1, AP-2, Tbx and NFκB (among others), although a more detailed comparison is not currently possibly due to a lack of sequence for the upstream regions of other *vegf* genes in this species or others.

Multi-species comparisons

A 25,026bp cosmid sequence encoding a cluster of three functional PLA_2 genes and two non-functional pseudogenes has previously been published for the Okinawan habu (*Protobothrops flavoviridis*) (Ikeda et al. 2010). The upstream regions of each of the functional genes (defined as the region upstream of the start (ATG) codon to the stop codon of the preceding gene or pseudogene, comprising 1,829bp upstream of PfPLA-2; 4,045bp upstream of PfPLA-5 and 5,444bp upstream of PfPLA-4) were searched and 168 putative transcription factor binding sites were identified to be conserved across the three sequences (an unsurprising finding given the high level of sequence identity between the three regions). These included conserved sites for candidate transcription factors such as AP-1, AP-2, Sp1, NFκB, Tbx3 and Hnf3a and, most

interestingly, a conserved Tef1 binding site, similar to that seen in *E. coloratus* SVMP loci (Figure 6.19). Comparison of these *Protobothrops* sequences to those from the *E. coloratus* genome identified 9 conserved transcription factor binding sites (Figure 6.21), all located within a few hundred base pairs of the start codon.

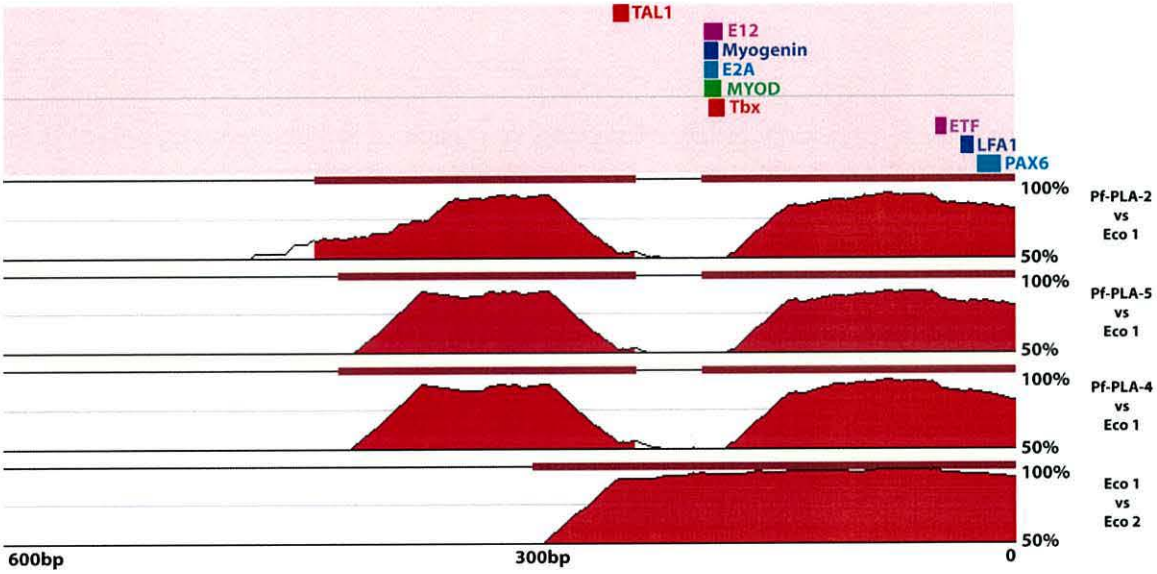


Figure 6.21. Transcription factor binding site analysis of the upstream regions of *Echis coloratus* (Eco) and *Protobothrops flavoviridis* (Pfl) phospholipase A_2 (PLA_2) genes. Position 0 denotes the first base of the start codon. The locations of the nine transcription factor binding sites found to be conserved across all sequences are shown in the top panel and pairwise sequence similarity plots are shown in the bottom panels, with peaks over 75% conserved colored red.

The full coding regions and approximately 300bp of upstream sequence for 19 PLA_2 genes from four North American rattlesnakes (*Sistrurus catenatus edwardsi*, *S. c. termgeminus*, *S. c. catenatus* and *Sistrurus miliarius barbouri*) have previously been published (Gibbs and Rossiter 2008). The upstream regions of these genes are highly similar, although there has been a 42bp insertion in one of the *S. c. edwardsi* sequences (edw4, accession EU369751) which has been annotated as a pseudogene. Comparing the remaining 18 genes to the upstream regions of the two *E. coloratus* and three *Protobothrops* PLA_2 sequences identified three conserved transcription factor binding sites, for E2A, myogenin and Tbx.

Although the level of sequence conservation between the *E. coloratus* CRISP paralogs was too high to accurately predict conserved transcription factor binding sites, a comparison incorporating an upstream region of a king cobra CRISP gene (on scaffold 7136, accession AZIM01007132) resulted in a much smaller list of just 35 conserved TF binding sites, all located within 2kb of the start of the gene (Figure 6.22).

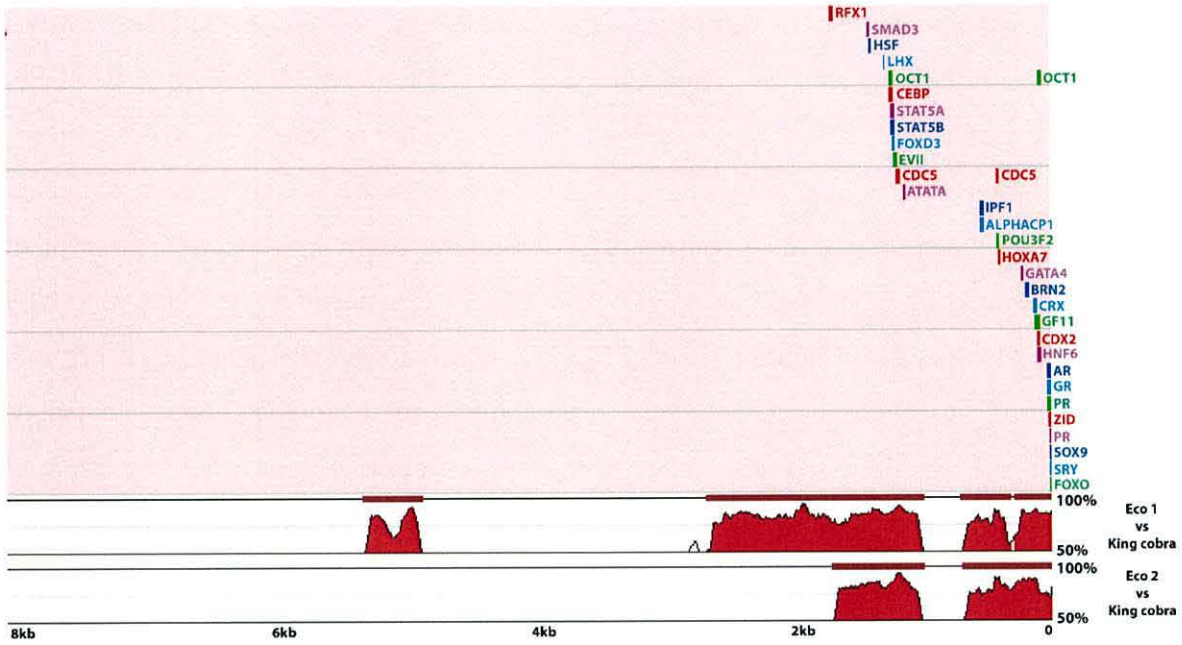


Figure 6.22. Transcription factor binding site analysis of 7,272bp of upstream sequence of *Echis coloratus* (Eco) and king cobra (*Ophiophagus hannah*) cysteine-rich secretory protein (CRISP) genes. Position 0 denotes the first base of the start codon. The locations of the 33 transcription factor binding sites found to be conserved across all sequences are shown in the top panel and pairwise sequence similarity plots are shown in the bottom panels, with peaks over 75% conserved colored red.

The available genomic scaffolds for nerve growth factor (*ngf*, accessions AZIM01002615 and AZIM01012844) from the king cobra (*Ophiophagus hannah*) (Vonk et al. 2013) were also analysed as it has previously been suggested to have undergone a gene duplication within the Elapid lineage (Hargreaves et al. 2014a; Hargreaves et al. 2014b; Sunagar et al. 2013, and also Chapters 4 and 5). Based upon phylogenetic analysis (Hargreaves et al. 2014a, Chapter 4) these have been designated *ngfa* and *ngfb*, with *ngfb* being the putatively toxic version due to its selective expression in the venom gland. *ngfa* sequences for the painted saw-scaled viper, the

corn snake, the Burmese python (Castoe et al. 2013) (accession KE954116) and the boa constrictor (Bradnam et al. 2013) (scaffold SNAKE00002822) were also extracted to aid in the identification of unique transcription factor binding sites. All six upstream regions were aligned to give 528bp of sequence upstream from the start codon for transcription factor binding site analysis. The total number of transcription factor binding sites was found to vary between species with 98 and 88 in king cobra *ngfa* and *ngfb* respectively, 163 in the painted saw-scaled viper, 101 in corn snake, 96 in Burmese python and 88 in boa constrictor. Just twelve transcription factor binding sites were found to be conserved between all species (CBF, USF, EBOX, BEL1, CMAF, PAX9, NRSE, NRSF and 2 sites for both YY1 and MYCMAX) (Figure 6.23). In the king cobra, *ngfa* and *ngfb* have 57 transcription factor binding sites in common. 28 additional sites were found to be conserved between king cobra *ngfa* and *ngfa* from other species, but are missing from king cobra *ngfb*. A total of 13 binding sites were unique to king cobra *ngfa*, most noticeably 3 binding sites for SMADs which are not present in any other *ngfa* upstream region from other species, or in *ngfb*. The putatively toxic *ngfb* has 70 sites conserved between itself and *ngfa* in either king cobra or at least one other snake species. *ngfb* has 13 novel transcription factor binding sites including one for hepatocyte nuclear factor (HNF4), one site for TATA-binding protein (TBP) and one site for transcription factor IIA (TFIIA) and there has also been an 8 nucleotide deletion 409bp upstream from the start codon.

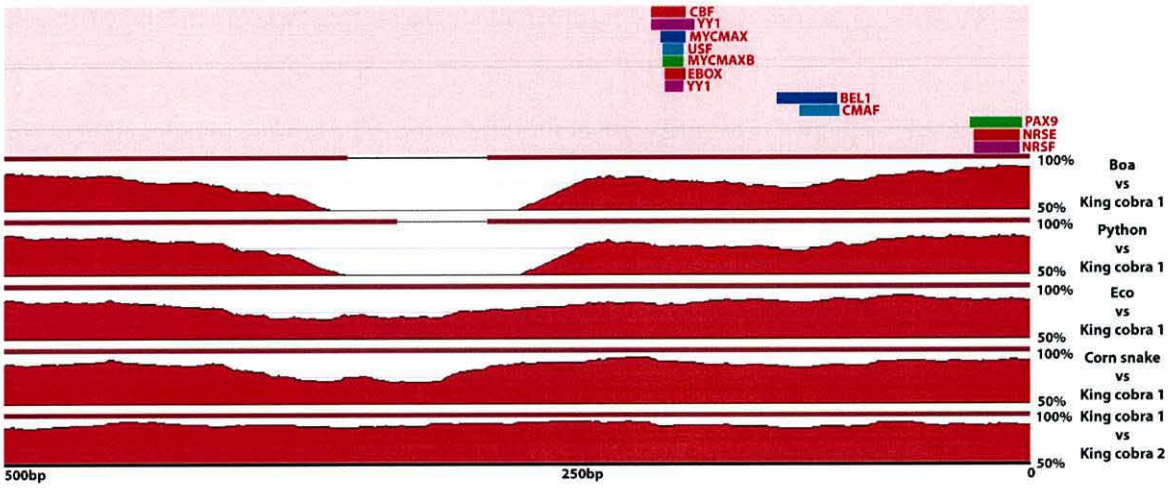


Figure 6.23. Transcription factor binding site analysis of 528bp of upstream sequence of *Echis coloratus*, king cobra (*Ophiophagus hannah*), corn snake (*Pantherophis guttatus*), Burmese python (*Python molurus bivittatus*) and Boa constrictor (*Boa constrictor constrictor*) nerve growth factor (*ngf*) genes. Position 0 denotes the first base of the start codon. The locations of

the 12 transcription factor binding sites found to be conserved across all sequences are shown in the top panel and pairwise sequence similarity plots are shown in the bottom panels, with peaks over 75% conserved colored red.

An insertion upstream of the transcription start site of the *coagulation factor X* gene in the rough-scaled snake (*Tropidechis carinatus*) and Eastern brown snake (*Pseudonaja textilis*) has been claimed to be responsible for the increased expression level of a venom gland-specific paralog following gene duplication (Kwong et al. 2007; Reza et al. 2007), and accordingly we find no evidence for either gene duplication or this insertion in *E. coloratus*, corn snake, king cobra or Burmese python. The upstream regions of the duplicated factor X genes (pseutarin C and trocarin D (Rao et al. 2004; Reza et al. 2007)) are extremely similar and possess 147 conserved transcription factor binding sites.

6.4 Discussion

Comparative transcriptomic and genomic analysis provides evidence that a diversity of gene regulatory networks may be involved in the transcriptional regulation of genes encoding snake venom toxins. Analyses of different time points during venom synthesis following manual extraction of venom reveals an apparent shift from a focus on transcription to translation between 16 and 24 hours post milking, indicating that activation of genes encoding venom toxins is a relatively rapid event. Only a single transcription factor (*Tbx3*) appears to be unique to the venom glands of venomous snakes and absent from the salivary glands of non-venomous species. Whilst more work is needed to establish the distribution of this gene in additional species, its presence in the venom gland is intriguing, since TBX3 is a known transcriptional repressor and is known to interact with diverse gene regulatory networks in multiple tissues in a context-dependent manner (Washkowitz et al. 2012). If TBX3 is carrying out a similar repressive function in the venom gland, then it is possible that the initiation of venom production following expenditure is facilitated by the removal of transcriptional repression (i.e. negative regulation), rather than by transcriptional activation. A similar situation is known to occur during embryonic development, where RNA polymerases are 'paused' on gene promoters and in this way offer a rate-limiting mechanism for transcription (Core and Lis 2008; Krumm et al. 1995). Such a system also enables rapid initiation of gene expression, as the transcriptional machinery is already assembled and in place on the gene itself (Margaritis and Holstege 2008). A gene regulatory mechanism such as this obviously has important

implications for the rapid initiation of venom replenishment, although the lack of conserved TBX binding sites in the upstream regions of venom genes would suggest that TBX3 is acting higher up in the venom gene regulatory network.

The presence of NFκB and its inhibitors in the venom glands of all species, together with the previous identification of a role for these transcription factors in the regulation of venom production (Luna et al. 2009) also supports a mechanism of rapid initiation of venom replenishment following expenditure, as NFκB dimers are held inactive in the cytoplasm via association with inhibitor proteins and can be rapidly activated by the removal and degradation of these inhibitors (Karin 1999).

If different gene families are regulated by distinct gene regulatory networks (or sub-components of the same network) then differences in the temporal expression of these genes should be expected. The results from this study indicate that this may be the case, with evidence for the expression of some gene families declining quite rapidly after 16 hours, and some increasing quite dramatically (such as *laao-b2*) after 48 hours. Not only is this suggestive of a number of gene regulatory networks being involved, it also has implications for the importance of respective venom toxins, with those essential to the functional efficacy of the venom being expressed and replaced almost immediately following expenditure.

The draft painted saw-scaled viper (*Echis coloratus*) genome sequence enabled comparative analyses of the upstream regions of the gene families that make the most substantial contribution to the venom of this and closely related species (SVMPs, Phospholipase A_2s, C-type lectins and Serine Proteases) (Casewell et al. 2009; Casewell et al. 2014; Wagstaff et al. 2009). These analyses identified a number of conserved predicted transcription factor binding sites between members of gene families, but not between members of different gene families, supporting the proposal that multiple gene regulatory networks may be acting within the snake venom gland, each working to activate the expression of typically one gene family. This situation is likely a reflection of multiple restriction events at different times during the evolution of venomous snakes, where a gene encoding a salivary protein has been duplicated and the expression of one of the copies restricted to the venom gland (Hargreaves et al. 2014b). It is also likely that the different transcriptional networks acting on the different gene families contribute to observed differences in venom composition within and between species (Casewell et al. 2014; Chippaux et al. 1991). In all cases, the conserved transcription factor binding sites were located within a few hundred base pairs of the start of the gene and this, together with the observed short intergenic distances in the previously characterized Okinawan habu

(*Protobothrops flavoviridis*) PLA_2 gene cluster (Ikeda et al. 2010), supports the initial hypothesis and explains the apparent ease with which functional copies of existing venom genes are produced via gene duplication (Kordiš and Gubenšek 2000; Wong and Belov 2012). It remains to be seen however if the short regulatory regions of these genes are an ancestral condition, in which case they may be considered to be pre-adapted or exapted (Gould and Vrba 1982) to repeated gene duplication, or if only partial regulatory regions were initially duplicated, in which case these genes may have been subfunctionalised and restricted to the venom gland by default.

Finally, comparative transcriptomic analysis of the *E. coloratus* venom gland with other body tissues in this species, and with the venom and salivary glands of other species, has resulted in some intriguing findings. Not only does the venom gland express the lowest number of unique transcripts of any of the seven tissues investigated, but it also does not appear to be particularly outstanding with respect to the number and type of secreted products. Whilst snake venom itself may represent an evolutionary innovation, with extensive intra- and inter-specific variation in venom composition resulting from a range of genomic, transcriptional, translational and post-translational mechanisms, the venom gland itself may in fact be a simple tissue expressing genes whose regulation has undergone only slight modification.

Chapter 7
General discussion

7.1 Principal findings

The use of venom is widespread amongst a diverse array of taxa, and represents a unique way of using modified proteins as a prey capture strategy. Venomous snakes have created both fear and fascination in equal measure for centuries, and the understanding of them is still a widely pursued area of biological research. From a human point of view, bites from venomous snakes cause a significant amount of mortality and morbidity each year, predominantly in the developing world. Conversely, snake venom represents a potential source of novel compounds with pharmaceutical applications.

As discussed in the preceding chapters venom has been, for many years, stated as being "highly complex" (Casewell et al. 2013), with the genes encoding venom toxins being "recruited" from multiple body tissues into the venom gland where they subsequently undergo mutation to become toxic (Fry 2005). Venom has been proposed to have evolved once at the base of squamate reptiles, with many secondary losses in favour of alternative prey capture methods such as constriction (Fry et al. 2006). Therefore, from the outset snake venom appears to be a fascinating model to study.

However, these theories have lacked the enormous amount of data provided by current DNA and RNA sequencing technologies. Despite these technologies being around for some time, at least in principle (the technologies themselves have evolved considerably since their first emergence), their use in the study of reptile venom is still moderately low although admittedly is on the rise (for example (Rokyta et al. 2012; Vonk et al. 2013; Margres et al. 2013)).

This work aimed to use 2nd generation sequencing technology to achieve several goals. Firstly, to sequence low-coverage draft genome sequences of three species of snake, in an order to fill the void of available snake genomic resources (Chapter 2). The recent sequencing of draft whole genome sequences for the Burmese python, king cobra and boa constrictor alongside these is surely only the beginning of a large influx of available reptile genome sequences.

Secondly, a large amount of tissue transcriptomes were sequenced for a range of reptile species and tissues (Chapter 3). This was done mostly for use in the analysis of venom gene expression,

but also to provide encouragement and assurance of the utility of this approach to other researchers, especially as traditional EST sequencing is still being used, for example (Casewell et al. 2014).

Lastly, these resources were used to evaluate several hypotheses of snake venom evolution, namely the Toxicofera hypothesis (Fry et al. 2006; Fry et al. 2009a; Fry et al. 2012b; Fry et al. 2013) (Chapter 4) and the theory of venom gene recruitment (Fry 2005) (and to a lesser extent the theory of reverse recruitment (Casewell et al. 2012)) (Chapter 5). Additionally the predominantly neglected area of venom gene regulation was also investigated (Chapter 6).

The results of this thesis can be split into two broad areas: bioinformatics and evolutionary genetics:

Bioinformatics

Performing a "mini Assemblathon" using newly generated genomic sequence data and three publicly available snake genome sequences was extremely informative in a number of regards. This analysis identified several considerations for genomic library sequencing and assembly which consistently improved the resulting output sequence, namely the use of paired-end reads sequenced to a high sequencing coverage. Alterations in other parameters had varying effects which seem to be dependent on the dataset itself. Out of the two assemblers CLC produced assemblies with the best metrics in terms of contig length and N50 values, but had the worst performance in terms of genome completeness based on CEGMA analysis. It became apparent that genome assemblies cannot be assessed based solely on basic metrics and that multiple metrics should be used when evaluating the quality of an assembly. If there is one take-home message from this analysis, it is that there is currently not one sole "correct" methodology to assemble a genome, and as such multiple methods should be attempted and evaluated for each genome assembly project.

Assembling multiple transcriptomes using several different methods was also revealing. For *de novo* transcriptome assembly Trinity greatly outperformed SOAPdenovo-Trans for most datasets. The exception to this was the assembly of transcriptomes from the available king cobra RNA-seq data, which were greatly improved by first mapping the reads to the genome of this species and then carrying out a genome-guided assembly using the Tuxedo suite. Other genome-guided assemblies in other species were not so improved, perhaps suggesting that the longer scaffold length of the king cobra genome was a key to this method, as more exons will

be located on a single scaffold, allowing the assembly of full length transcripts. Additionally the short, single end reads for the king cobra tissues may have caused issues in the De Bruijn graph construction utilised by the *de novo* assemblers. The quantitative nature of RNA-seq data adds to its appeal, and was demonstrated by analysing the expression of several reference genes used for qPCR experiments across a number of tissues and in the venom gland at different time points following milking. Again this study confirmed that GAPDH and β actin are unsuitable for use as reference genes.

Comparison to the previously generated venom gland transcriptomes from *Echis* species showed the utility of RNA-seq in this area, being more sensitive in detecting lower expressed transcripts and also splice variants of several genes. The quantitative qualities of RNA-seq data also solidified its utility in estimating the abundance of venom gene transcripts.

Finally, conducting sub-assemblies of the venom gland data identified that many transcript sequences for proposed venom genes were successfully reconstructed using only 2 million 100bp paired-end reads as initial input to Trinity. As a conservative estimate it was decided that 8 million paired-end reads should be the minimum required, as the majority of transcripts were present using this amount of data whilst also being similar in length to the query sequence (assembled from considerably more data) and also high in similarity to it (suggesting a minimum amount of error had been incorporated into the assembly).

Evolutionary genetics

Using these newly generated genomic and transcriptomic resources, and incorporating a considerable amount of publicly available transcriptomic data from a range of reptile species and tissues, I set out to evaluate several hypotheses of venom evolution in reptiles. As stated in the introduction (Chapter 1) the hypothesis for a single, early origin of venom in reptiles (the Toxicofera hypothesis) has been propagated and expanded for many years, but has lacked the insight provided by the addition of non-venom gland body tissues. A quantitative, comparative transcriptomic assessment elucidated that the previous claims of the Toxicofera hypothesis are likely due to incomplete tissue sampling and the incorrect interpretation of phylogenetic trees. It appears that all of the genes used in support of the Toxicofera hypothesis are expressed in multiple tissues with no evidence of a venom gland-specific splice variant or an elevated expression in the venom gland, and so most likely encode housekeeping or maintenance genes. This highlights that the transcriptomic analysis of solely venom gland is not sufficient to assign

a toxic role to a gene, or infer its true evolutionary history, especially in support of a hypothesis encapsulating an entire lineage. Several problematic issues were identified such as the classification of extremely similar sequences as different genes and the misidentification or incorrect annotation of others. The inclusion of genes which have never been functionally characterised as being toxic is also especially confusing (for example cystatin which has never been shown to show toxicity, even when it was first discovered (Ritonja et al. 1987)). Phylogenetic analyses carried out in the "reverse recruitment" study (Casewell et al. 2012) found no support for the majority of Toxicoferan toxins when non-venom gland derived sequences were included. As a result the incidence of reverse recruitment was proposed, with toxin genes being recruited back into the body to once again perform a physiological role. However, the more parsimonious conclusion is that these genes were never toxins, and as such represent housekeeping or maintenance genes with a wide expression pattern. Snake venom appears to be a simple mixture whose active components are encoded by a small number of gene families which have expanded through gene duplication. The results of the comparative and quantitative approach used to identify putative toxin-encoding transcripts appears to accord well with previous proteomic analyses of snake venom, making it a useful methodology for future transcriptomic studies.

Analysis of gene expression in multiple body tissues detected the expression of members of venom gene families in an array of tissues, including the salivary gland of non-venomous reptiles. Re-analysis of the original proposal of venom gene recruitment (Fry 2005) revealed that there was no evidence to support this hypothesis from the outset, despite being widely accepted. A comparative transcriptomic analysis found that many genes belonging to putative venom gene families are expressed in a number of tissues including the salivary gland of non-venomous reptiles, and therefore venom genes have been restricted to the venom gland following gene duplication. This process is much more parsimonious than the previous suggestion that venom genes undergo numerous rounds of neofunctionalisation, requiring the evolution of novel gene regulatory networks to lead to venom gland-specific expression. In the case of restriction, genes are already expressed in the venom/salivary gland prior to developing toxicity, and so there is no requirement for a new gene regulatory network.

The gene regulatory mechanisms active in the venom gland do not appear to be particularly outstanding compared to the salivary gland of non-venomous species, with both tissues having very similar transcription factors and signalling pathways active, and both having similar numbers of secreted products. Only one transcription factor was found to be expressed in the

venom gland and not the salivary gland, *Tbx3*, which along with the previously suggested role for NFκB in venom production suggests negative regulation, where the removal of repression rather than activation is the key first step in gene transcription. Although previously suggested to play a role in venom production, there was no evidence of conserved adrenergic receptors in snake venom glands. Finally, transcription factor binding sites analysis revealed that binding sites were conserved between members of the same gene family but not between gene families suggesting that multiple gene regulatory networks are involved in venom gene expression. The close proximity of binding sites to the start of genes is indicative of an arrangement which facilitates repeated gene duplication with the necessary regulatory regions for expression.

7.2 Implications

These results taken together have several implications for the current understanding of venom evolution in reptiles. The finding that many proposed venom toxin genes are in fact more likely to encode housekeeping genes suggests that much of the research effort investigating the evolution of these genes and the properties of the proteins they encode will not be directly applicable to the development of the next generation of antivenom treatments. Perhaps the most significant finding of this work is that the Toxicofera hypothesis is unsupported (Chapter 4), prompting a move back to the previously held view that venom evolved twice within the reptile lineage, once in snakes and once in venomous lizards such as the Gila monster (the classification of Varanid lizards as venomous is uncertain). The complexity of snake venom composition appears to have been consistently overestimated by previous authors. Indeed, the low number of products in venom makes perfect sense as (1) a complex proteinaceous mixture would be metabolically expensive to produce and (2) natural selection will act to streamline the venom to be tailored to the snakes prey items. In short, a simple venom is efficient; a complex venom is overkill. Interestingly it appears that some putative toxin genes appear to be lineage specific, with evidence of gene duplication and restriction of *nerve growth factor b* and *complement c3b* in Elapids/cobras. *vascular endothelial growth factor f* also appears to be specific to viperid snakes. These findings have greater evolutionary implications, and may suggest the adoption of specific venom components in order to exploit a new ecological niche or to expand the variety of potential prey items.

In Chapter 5 it was found that venom genes do not evolve via recruitment to the venom gland, but rather restriction to it following gene duplication. The expression of members of toxin gene

families in multiple tissues, most notably the salivary gland of non-venomous species, is suggestive that venom is simply a modified version of saliva. It is possible that the initial toxicity of a venom component is dosage dependent, where the two duplicate genes produce twice the amount of the same protein product (their temporal and spatial expression patterns will be identical immediately following duplication). If this increased dosage provides an advantage in prey capture, selection will act to maintain its expression in the venom gland, where it can subsequently be modified by mutation to develop toxicity, whilst being lost from other tissues.

Chapter 6 identified that there is little difference in the transcription factors and signalling pathways active in the venom gland and the salivary gland of non-venomous reptiles, suggesting that the venom gland is not as specialised as it may seem. Indeed, it expresses the lowest number of unique transcripts compared to 6 other body tissues. The presence of only a single transcription factor expressed in the venom gland but not the salivary gland is interesting, particularly as *Tbx3* is a transcriptional repressor. The conservation of transcription factor binding sites between genes in the same gene family but not between different gene families is indicative that multiple gene regulatory networks are active in venom gene expression, and the evidence of temporal variation in expression following venom extraction would also support this. This would imply that multiple restriction events have led to the formation of the venom gene repertoire, with each family being subject to its own regulatory network. The variation in venom composition, especially the obvious variation between vipers and elapids (with vipers predominantly expressing SVMPs and serine proteases and elapids predominantly expressing 3 finger toxins and PLA_2s), may be caused by this differential gene regulation. The close proximity of transcription factor binding sites to the start of venom genes may suggest that they are exapted (Gould and Vrba 1982) to repeated rounds of gene duplication, whilst carrying with them the necessary *Cis*-regulatory architecture needed for their continued venom-gland specific expression and regulation

7.3 Future work

Proposals for future work relevant to each chapter of this thesis are outlined below. There are several possible strategies to improve the overall quality and completeness of the 3 newly generated low-coverage snake genome sequences. An increase in paired-end sequencing will ultimately increase sequencing depth, which is especially needed for the lower coverage

genomes of the corn snake and *Echis pyramidum*. However, due to the repeat content of genomes and compositional biases in genomic regions (for example GC-rich regions), an increased in paired-end sequencing is highly unlikely to generate a highly improved genome assembly by itself.

The inclusion of mate-pair sequencing library data in the boa constrictor, Burmese python and king cobra genomes appears to have greatly aided the formation of longer genomic scaffolds, and it is likely that the addition of this to the current *Echis* and corn snake paired-end data would provide the long-range positional information needed to extend scaffold length considerably. It is worth noting that although the three published genomes are superior to the ones generated in this study, they are still incomplete and have relatively conservative scaffold lengths and N50s when compared to other published genome sequences (for example the Western painted turtle, *Chrysemys picta*, has a scaffold N50 of 5.2Mb (Shaffer et al. 2013)).

Simply carrying out more and more sequencing will inevitably lead to a "sequencing plateau" where the overall coverage (and therefore reliability) of the genome sequence will increase, but the positional information needed to correctly place scaffolds will ultimately be missing. Ideally more traditional mapping techniques could be used to improve genome assemblies such as the generation of a genetic linkage map to orientate the positions of genes relative to each other in the genome or the use of BAC (bacterial artificial chromosome) libraries to sequence long (>100Kbp) sequences of genomic regions. In this way the assembly process can move up from contigs and scaffolds, to potentially assembling entire chromosomes.

Finally, the dawn of long-read sequencing such as that offered by Pacific Biosciences or Oxford Nanopore is a potentially paradigm-shifting event, where the generation of sequencing reads of up to 20kb (potentially much longer using nanopore sensing) could potentially lead to the complete *de novo* sequencing of an entire genome in a few sequencing experiments. The very long read length offered by these technologies also means that issues with troublesome repetitive and GC-rich regions which are difficult to sequence and assemble will be negated. Whilst these technologies are still very much in their infancy, it has been shown that a small amount of long-read sequencing, coupled with Illumina data, can produce very encouraging results (Koren et al. 2012; Ganapathy et al. 2014). The construction of much longer scaffolds would make allow the mechanism of gene duplication responsible for the propagation of venom genes to be elucidated, for example if they are clustered within the genome and flanked by

repetitive elements if would suggest that they have been duplicated by unequal crossing over during homologous recombination.

The variation in venom composition between members of the *Echis* genus is intriguing, and worth further pursuit. It has recently been suggested that several post-genomic mechanisms such as post-transcriptional and post-translational modification, are responsible for the variation in venom composition between members of the same genus (Casewell et al. 2014). However, there has to date been no comprehensive study on the genomic aspect of this variation, where the presence or absence of genes (or indeed occurrences of pseudogenisation within some species but not others) could be responsible for venom variation. Only with further genome sequencing and improvement can this area be fully investigated.

Finally, the whole genome sequence of the corn snake has been used only briefly in this study. The genome of this species will be useful in studies seeking to understand pigmentation patterning in reptiles, both in terms of neural crest cell differentiation and migration and the biosynthetic pathways producing pigments such as melanin. As a member of the Colubridae this species is also a useful addition for comparative genomic studies within Serpentes alongside the Boidae (boa constrictor, Burmese python), Viperidae (painted and Egyptian saw-scaled vipers) and Elapidae (king cobra).

Trinity appears to be the best and most useful approach taken in this work, both for the quality and sensitivity of the *de novo* transcriptome sequences it produces and the versatility of its downstream applications. *De novo* assembly using short reads for venom toxin transcripts appears to have issues with highly similar members of the same gene family, with on the whole only partial sequences being assembled. This is most likely due to the ambiguity of highly similar k-mers during assembly, with such a small amount of sequence variation between fragments being interpreted as sequencing errors. Therefore these k-mers are likely removed during the assembly process. The use of long read sequencing technology offered by companies such as Oxford nanopore would be a logical step in overcoming this problem, and allowing the sequencing of full-length mRNA transcripts. It would then be possible to map short-read data to the full length transcripts, both to aid in their assembly and also for transcript abundance estimation. This is particularly useful as mapping reads to short fragments of a transcript may result in a low FPKM values, thus leading to the conclusion that the transcript is not expressed, when in fact mapping to a full-length transcript may result in the opposite conclusion. A recent study used the Oxford nanopore MinION nanopore sensing device to sequence cDNA derived

from the venom gland of the Okinawan habu, *Protobothrops flavoviridis* (Mikheyev and Tin 2014). The findings of this paper were somewhat controversial in that the authors claimed that "the current iteration of the MinION is not ready for routine use". It is likely that this is due to the use of an early iteration of the library preparation kit and software, and that the performance of this technology is now greatly improved, signifying an exciting and potentially limitless tool for DNA and RNA sequencing.

Whilst this work casts doubt on the Toxicofera hypothesis, further analysis is required in order to fully refute it. More specifically an increased number of body tissues (from the study species used and also a wider range of reptile taxa) need to be incorporated to gain a clearer picture of the true extent of the expression of proposed Toxicoferan venom genes. The inclusion of the green anole lizard, *Anolis carolinensis*, stands out as a logical and easy example to incorporate due to its whole genome sequence (Alföldi et al. 2011) and several body tissue transcriptomes (Eckalbar et al. 2013) (including brain, heart, liver and ovary) already being available. The phylogenetic position of *Anolis* as members of the Iguania at the base of the Toxicofera clade means they are one of the closest relatives of the proposed venomous squamate ancestor, and so represent ideal candidates for study.

The expression of putatively toxic transcripts (2 SVMPs and *crisp-b*) in the corn snake salivary gland is of great interest, and may suggest that an ancestor of the Caenophidia (advanced snakes) may have been venomous. Further investigation via the sampling of other colubrid snakes including opisthoglyphous species known to be venomous (such as the boomslang, *Dispholidus typus*) and those which rely on constriction (such as the corn snake) may shed light on this. Finally, a transcriptomic approach identifies putative toxin-encoding transcripts, but proteomic and functional characterisation are needed in order to validate whether those transcripts represent active venom components.

It is unlikely that the expression of venom genes in the snake venom gland arose through the recruitment of non-toxic physiological or housekeeping genes from multiple body tissues. Instead it appears that venom toxin genes were ancestrally expressed in a variety of tissues including the venom/salivary gland prior to developing toxicity. The addition of more species and more tissue samples will add further support to this hypothesis. The finding of lowly expressed genes in the venom gland which are members of venom gene families (such as SVMPs) in Chapter 4 suggests that perhaps not all genes belonging to these families are toxic, which again is suggestive that members of these gene families are expressed in the venom or

salivary gland prior to gene duplication and the development of toxicity. An expansion in the number of species sampled (the addition of the most basal species of viper, *Azemiops feae*, would be especially interesting) would enable the reconstruction of the true evolutionary history of these gene families, perhaps indicating the timings of gene duplications (as was demonstrated for *nerve growth factor*, *complement c3* and *crisp* in Chapter 4) to give a clearer picture of their origins.

Improved, more complete, genome sequences will allow a more thorough analysis of venom gene promoter regions which, when compared to the gene promoters of non-venomous species, may indicate whether venom genes have been restricted to the venom gland through the gradual mutation and degeneration of transcription factor binding sites, or if these genes have been partially duplicated with binding sites missing, leading to them being only expressed in the venom gland by default.

As with all aspects of this study, increased sampling from a range of species and tissues will greatly benefit further investigation of the genetic regulatory mechanisms underpinning venom production. Venom gene expression appears to be fully underway at 16 hours post-milking, and so earlier sampling timepoints are needed in order to gain a comprehensive picture of gene expression following venom expenditure. The seemingly sudden expression of *laao-b2* after 48 hours also indicates that sampling at later timepoints is also required to gain a full understanding of this process.

A genome sequence with improved venom gene promoter regions, paired with the sequencing of more venom gland samples milked at an increased range of timepoints could potentially allow the reconstruction of the venom gene regulatory networks active in the venom gland post extraction (this approach was taken in a study predicting the gene regulatory networks active during soybean nodulation (Zhu et al. 2013)). The proteomic analysis of extracted venom at different timepoints following an initial milking may also indicate whether there is a temporal difference in venom gene expression by the presence or absence of venom proteins at different stages.

Finally, Chromatin immunoprecipitation sequencing (ChIP-seq) would help identify the binding sites of transcription factors, and could shed light on those involved in the activation (or the removal of inhibition) of venom gene expression.

7.4 Perspectives

This work has highlighted several broad-ranging issues to consider alongside its scientific conclusions. Firstly it has identified several cases where more stringency and scrutiny were required, especially when characterising and annotating sequences as different genes based on either minimal differences in sequence or based on BLAST searches. Furthermore the problems associated with a lack of standardised gene nomenclature became apparent, and a plea for a more rational nomenclatural system for snake venom toxins has been addressed to the editor of Toxicon (the official journal of the International Society on Toxinology) (Hargreaves and Mulley 2014). It is hoped that this will encourage the adoption of this system, making the evolutionary history of genes encoding snake venom toxins clear to see and current research more accessible to scientists from outside the relatively niche area of snake venom.

This study was in some ways limited by the unavailability of the genomic sequencing reads of the recently published Burmese python (Castoe et al. 2013) and king cobra (Vonk et al. 2013). More specifically, the king cobra genome paper claims to have shown that several gene families (snake venom metalloproteinases, cysteine-rich secretory proteins and lectins) are clustered in the genome of this species, which would add support to venom genes being arranged in clusters in the genome, having implications for how venom gene families expand through gene duplication. However, the scaffolds and gene models which purportedly show this clustering are not publicly available, despite several requests to the authors. Therefore, the actual clustered arrangement of these genes within the genome are in doubt. As reproducibility and verification are key to strengthening (or conversely refuting) a hypothesis, and are "...integral foundations of the scientific method." (Kardong 2012), it is paramount that data is made available to the scientific community.

These analyses have brought to light several occurrences of "just-so stories" in evolutionary biology. The theory of venom gene "recruitment" and the Toxicofera hypothesis have been widely, and unquestionably, accepted for nearly a decade. In his seminal work "*The structure of scientific revolutions*", Thomas Kuhn states: "Scientists work from models acquired through education and through subsequent exposure to the literature often without quite knowing or needing to know what characteristics have given these models the status of community paradigms" (Kuhn 1962). Indeed this appears to have been the case, and is a cautionary tale that hypotheses, especially those which are drastically different to previous long-standing ones, should constantly be re-assessed and re-tested.

Considering these findings in a wider context, they also present several examples of high impact papers (one in Genome research (Fry 2005) and one in Nature (Fry et al. 2006)) which have been, to some extent, refuted by this study. The timing of these findings is timely, considering the recent retraction of two publications in Nature proposing a method of creating stimulus-triggered acquisition of pluripotency (STAP) cells (Obokata et al. 2014a; Obokata et al. 2014b), another example of high impact papers being rebutted through re-analysis (albeit in a much quicker timeframe). Herein lies the question of whether the appeal of these papers is an overall factor in their publication, being relatively controversial, without having irrefutable evidence to support their claims. Additionally, it also raises the question of whether pre-publication peer review is a stringent enough process for studies with conclusions contrary to the status quo. The recent emergence and expansion of pre-publication servers (such as BioRxiv) and Open Access (OA) journals (such as PeerJ) is indicative that post-publication peer-review, where reviews are visible and open to all, is on the increase. The OA publishing policy also means that the data used in a study is available to all for use in other studies or to test the repeatability of a particular finding. Coupled with the rise in "altmetrics" (internet-based alternatives to the commonly used journal impact factor) (Roemer and Borchardt 2012) this also suggests that attitudes and approaches to publishing scientific research, and more generally the way scientific knowledge is communicated and disseminated (both amongst the scientific community and the general public), appears to be changing.

In summary, this thesis work has provided evidence to cast doubt on several hypotheses of venom evolution in reptiles. The widespread expression of many genes proposed to encode Toxicoferan venom toxins suggests that snake (and possibly lizard) venoms are not as complex as they have been made out to be, with a large number of genes likely to encode housekeeping or maintenance genes. This finding, alongside little evidence of the generation of venom-specific transcripts via alternative splicing or instances of pleiotropy via elevated expression in the venom gland, is indicative that venom is a relatively simple mixture containing a small number of proteins from a few gene families. The original findings of the Toxicofera hypothesis appear to primarily be a consequence of incomplete sampling, and the evolution of venom once at the base of squamate reptiles with multiple subsequent secondary losses is not supported.

Gene duplication has been suggested for many years to be a major driving force in the creation of new genes with novel functions. Whilst this process is an important part of venom gene diversification, venom genes have not originated through repeated gene duplication and

neofunctionalisation. Instead it appears that venom genes have an ancestral expression pattern in a range of tissues which includes the venom gland and salivary gland of non-venomous reptiles. Hence venom genes have been subfunctionalised, or restricted, to the venom gland following gene duplication, a much simpler and parsimonious process.

Analysis of the gene regulatory mechanisms active in the venom gland and salivary gland of non-venomous species revealed little difference, with only a single transcription factor (*Tbx3*) being found in the venom gland but not the salivary gland. Both the venom and salivary gland appear to have similar numbers of secreted products, and the venom gland has the lowest amount of unique expressed transcripts out of the 7 tissues surveyed. The venom gland appears to be a relatively simple secretory glandular tissue, which produces and stores proteins whose encoding genes are subject to slightly modified regulation by pre-existing gene regulatory networks.

In conclusion, the evolution of venom in reptiles appears to be much less complex than previously believed but then again as Edsger Wybe Dijkstra, the Turing award winning computer scientist, alludes to in the quote at the beginning of this thesis: "complexity sells better".

Appendix 1. Contig assembly metrics for painted saw-scaled viper (*Echis coloratus*) genome assemblies.

SE/PE	Left/Right	Library insert size	Trimmed (Y/N)	Assembler	k-mer length	no. of contigs	total length	max contig length	contig N50
SE	Left	300	N	ABySS	20	52,882,988	1,610,627,662	993	27
SE	Left	300	Y	ABySS	20	48,549,549	1,486,824,137	1,132	28
SE	Left	300	N	ABySS	31	23,097,699	1,359,978,735	4,445	66
SE	Left	300	Y	ABySS	31	21,554,630	1,258,405,847	4,428	67
SE	Left	300	N	ABySS	40	16,712,851	1,303,886,440	4,433	99
SE	Left	300	Y	ABySS	40	15,547,230	1,190,670,477	3,880	92
SE	Left	300	N	ABySS	60	7,781,002	871,811,997	2,969	134
SE	Left	300	Y	ABySS	60	6,164,254	661,078,412	2,964	127
SE	Right	300	N	ABySS	20	52,034,844	1,591,731,376	1,496	28
SE	Right	300	Y	ABySS	20	47,060,869	1,445,670,816	1,434	28
SE	Right	300	N	ABySS	31	22,920,656	1,353,333,552	5,376	67
SE	Right	300	Y	ABySS	31	21,205,914	1,232,421,614	4,028	67
SE	Right	300	N	ABySS	40	16,690,349	1,303,273,745	3,645	100
SE	Right	300	Y	ABySS	40	15,318,533	1,163,925,023	3,507	90
SE	Right	300	N	ABySS	60	7,756,161	872,498,661	2,902	134
SE	Right	300	Y	ABySS	60	5,728,952	606,883,928	2,873	126
SE	Left	600	N	ABySS	20	98,440,069	2,955,533,611	1,599	27
SE	Left	600	Y	ABySS	20	94,669,392	2,863,427,880	1,599	28
SE	Left	600	N	ABySS	31	26,064,782	2,024,274,880	15,322	129
SE	Left	600	Y	ABySS	31	24,748,036	1,948,662,552	12,281	142
SE	Left	600	N	ABySS	40	17,743,010	1,972,017,089	15,508	259
SE	Left	600	Y	ABySS	40	17,100,594	1,903,435,778	10,968	245
SE	Left	600	N	ABySS	60	10,786,376	1,902,979,994	6,750	301
SE	Left	600	Y	ABySS	60	10,293,087	1,673,279,519	3,820	222
SE	Right	600	N	ABySS	20	98,642,093	2,950,928,219	1,258	27
SE	Right	600	Y	ABySS	20	94,449,577	2,847,504,620	1,311	28
SE	Right	600	N	ABySS	31	25,919,755	2,007,541,631	13,772	130

SE	Right	600	Y	ABySS	31	25,024,212	1,955,113,802	13,058	136
SE	Right	600	N	ABySS	40	17,910,730	1,974,511,873	14,943	247
SE	Right	600	Y	ABySS	40	16,761,647	1,866,627,177	11,665	242
SE	Right	600	N	ABySS	60	10,671,479	1,859,182,498	7,417	286
SE	Right	600	Y	ABySS	60	10,070,117	1,608,837,841	4,823	213
SE	Left	300 600	N	ABySS	20	101,765,390	3,041,901,910	1,258	27
SE	Left	300 600	Y	ABySS	20	103,986,533	3,081,773,307	1,258	27
SE	Left	300 600	N	ABySS	31	29,483,749	2,162,803,892	18,425	100
SE	Left	300 600	Y	ABySS	31	28,542,094	2,112,135,397	21,409	105
SE	Left	300 600	N	ABySS	40	20,737,248	2,146,083,623	29,257	191
SE	Left	300 600	Y	ABySS	40	20,398,133	2,114,457,968	18,834	191
SE	Left	300 600	N	ABySS	60	12,018,342	2,033,068,268	12,464	318
SE	Left	300 600	Y	ABySS	60	13,055,595	2,032,642,770	5,810	229
SE	Right	300 600	N	ABySS	20	101,553,627	3,031,839,570	1,326	27
SE	Right	300 600	Y	ABySS	20	104,228,817	3,080,209,441	1,258	27
SE	Right	300 600	N	ABySS	31	29,708,623	2,168,301,275	18,809	98
SE	Right	300 600	Y	ABySS	31	28,763,301	2,116,164,603	18,364	103
SE	Right	300 600	N	ABySS	40	20,916,539	2,150,737,056	19,750	185
SE	Right	300 600	Y	ABySS	40	20,211,681	2,091,859,274	16,673	191
SE	Right	300 600	N	ABySS	60	11,912,369	2,001,342,166	9,284	307
SE	Right	300 600	Y	ABySS	60	12,902,625	1,984,853,250	4,599	221
PE	BOTH	300 600	?	CLC	Default	5,836	23,998,012	70,197	7,887
PE	BOTH	300	N	ABySS	20	2.95E+07		2,561	605
PE	BOTH	300	Y	ABySS	20	2.65E+07		3,118	606
PE	BOTH	300	N	ABySS	31	1.21E+07		7,798	707
PE	BOTH	300	Y	ABySS	31	1.10E+07		7,092	693
PE	BOTH	300	N	ABySS	40	7.99E+06		7,787	726
PE	BOTH	300	Y	ABySS	40	7.11E+06		6,441	701
PE	BOTH	300	N	ABySS	60	3.39E+06		5,690	711
PE	BOTH	300	Y	ABySS	60	2.68E+06		4,061	680
PE	BOTH	600	N	ABySS	31				

PE	BOTH	600	Y	ABySS	31	1.60E+07		26,352	1,855
PE	BOTH	600	N	ABySS	40	1.00E+07		45,908	2,602
PE	BOTH	600	Y	ABySS	40	9.13E+06		43,181	2,497
PE	BOTH	600	N	ABySS	60	3.41E+06		58,907	3,557
PE	BOTH	600	Y	ABySS	60	2.83E+06		33,543	2,611
PE	BOTH	300 600	N	ABySS	31	1.90E+07	1.85E+09	37,410	1,921
PE	BOTH	300 600	Y	ABySS	31	1.85E+07	1.82E+09	32,238	1,895
PE	BOTH	300 600	N	ABySS	40	1.20E+07	1.83E+09	49,059	2,672
PE	BOTH	300 600	Y	ABySS	40	1.14E+07	1.79E+09	49,048	2,597
PE	BOTH	300 600	N	ABySS	60	4.97E+06	1.72E+09	63,379	3,857
PE	BOTH	300 600	Y	ABySS	60	4.49E+06	1.57E+09	47,508	3,244

Appendix 2. Scaffold assembly metrics for painted saw-scaled viper (*Echis coloratus*) genome assemblies.

SE/PE	Left/Right	Library insert size	Trimmed (Y/N)	Assembler	k-mer length	no. of scaffolds	total length (scaffs)	max scaffold length	Scaffold N50
PE	BOTH	300	N	ABySS	20	29,462,223	1,005,733,431	3,135	643
PE	BOTH	300	Y	ABySS	20	26,478,259	899,132,704	3,118	641
PE	BOTH	300	N	ABySS	31	12,070,363	843,221,240	7,798	718
PE	BOTH	300	Y	ABySS	31	10,998,215	748,185,677	7,092	702
PE	BOTH	300	N	ABySS	40	7,982,329	778,625,210	7,787	738
PE	BOTH	300	Y	ABySS	40	7,102,388	690,810,317	6,441	711
PE	BOTH	300	N	ABySS	60	3,372,482	560,252,124	5,691	730
PE	BOTH	300	Y	ABySS	60	2,671,144	407,332,236	4,748	702
PE	BOTH	600	N	ABySS	31	16,628,861	1,774,374,007	59,662	3,493
PE	BOTH	600	Y	ABySS	31	15,706,059	1,717,811,860	58,936	3,070
PE	BOTH	600	N	ABySS	40	9,773,504	1,706,903,884	63,029	4,095
PE	BOTH	600	Y	ABySS	40	8,894,776	1,622,224,753	61,058	3,874
PE	BOTH	600	N	ABySS	60	3,233,470	1,441,087,604	74,799	5,121
PE	BOTH	600	Y	ABySS	60	2,692,689	1,189,492,149	47,664	3,333
PE	BOTH	300 600	N	ABySS	31	18,700,353	1,876,280,157	102,856	3,283
PE	BOTH	300 600	Y	ABySS	31	18,155,538	1,845,273,171	82,168	3,150
PE	BOTH	300 600	N	ABySS	40	11,716,689	1,845,952,193	78,528	4,216
PE	BOTH	300 600	Y	ABySS	40	11,165,290	1,800,827,153	70,543	4,010
PE	BOTH	300 600	N	ABySS	60	4,790,800	1,727,283,171	84,548	5,576
PE	BOTH	300 600	Y	ABySS	60	4,307,776	1,576,438,475	65,261	4,524

Appendix 3. Contig assembly metrics for Egyptian saw-scaled viper (*Echis pyramidum*) genome assemblies.

SE/PE	Left/Right	Library insert size	Read length(s)	Trimmed?	Assembler	k-mer length	no. of contigs	total length (contigs)	max contig length	contig N50
PE	BOTH	~400	2x150	Default	CLC	Default	850298	326,820,222	11,529	393
PE	BOTH	~400	2x250	Default	CLC	Default	1670965	796,582,058	11,530	517
PE	BOTH	~400	2x150 2x250	Default	CLC	Default	1651433	944,133,977	11,528	674
PE	BOTH	~400	2x150 2x250	Default	CLC	31	1705426	1,032,996,853	12,522	711
PE	BOTH	~400	2x150	N	ABySS	20	5,127,500	159,188,986	3,724	744
PE	BOTH	~400	2x150	Y	ABySS	20	3,943,161	116,451,504	3,724	798
PE	BOTH	~400	2x150	N	ABySS	31	1,453,426	72,631,759	6,255	610
PE	BOTH	~400	2x150	Y	ABySS	31	1,083,186	50,541,142	6,255	744
PE	BOTH	~400	2x150	N	ABySS	40	730,819	47,054,491	7,811	681
PE	BOTH	~400	2x150	Y	ABySS	40	528,220	31,840,149	7,820	799
PE	BOTH	~400	2x150	N	ABySS	60	208,100	19,481,773	5,553	835
PE	BOTH	~400	2x150	Y	ABySS	60	141,635	12,724,200	11,604	847
PE	BOTH	~400	2x250	N	ABySS	20	2.30E+07	7.72E+08	1,320	548
PE	BOTH	~400	2x250	Y	ABySS	20	1.36E+07	4.71E+08	1,236	556
PE	BOTH	~400	2x250	N	ABySS	31	6,590,424	464520881	5,071	570
PE	BOTH	~400	2x250	Y	ABySS	31	3,719,322	245076639	5,268	588
PE	BOTH	~400	2x250	N	ABySS	40	3,780,891	361563617	5,080	575
PE	BOTH	~400	2x250	Y	ABySS	40	1,939,781	166910631	5,080	591
PE	BOTH	~400	2x250	N	ABySS	60	1,331,000	184143606	6,210	597
PE	BOTH	~400	2x250	Y	ABySS	60	530,090	61862348	6,677	590
PE	BOTH	~400	2x150 2x250	N	ABySS	20	3.64E+07	1.25E+09	990	539
PE	BOTH	~400	2x150 2x250	Y	ABySS	20	2.54E+07		989	530
PE	BOTH	~400	2x150 2x250	N	ABySS	31	9,467,861	799656496	5,424	623
PE	BOTH	~400	2x150 2x250	Y	ABySS	31	6,770,362	548333241	5,104	643
PE	BOTH	~400	2x150 2x250	N	ABySS	40	6,126,136	697274790	5,113	633
PE	BOTH	~400	2x150 2x250	Y	ABySS	40	3,787,285	416017048	4,013	653
PE	BOTH	~400	2x150 2x250	N	ABySS	60	2,382,270	428055199	5,627	669
PE	BOTH	~400	2x150 2x250	Y	ABySS	60	1,098,400	164158578	5,739	637

Appendix 4. Scaffold assembly metrics for Egyptian saw-scaled viper (*Echis pyramidum*) genome assemblies.

SE/PE	Left/Right	Library insert size	Read length(s)	Trimmed?	Assembler	k-mer length	no. of scaffolds	total length (scaffolds)	max scaffold length	Scaffold N50
PE	BOTH	~400	2x150	Default	CLC	Default	N/A	N/A	N/A	N/A
PE	BOTH	~400	2x250	Default	CLC	Default	N/A	N/A	N/A	N/A
PE	BOTH	~400	2x150 2x250	Default	CLC	Default	N/A	N/A	N/A	N/A
PE	BOTH	~400	2x150 2x250	Default	CLC	31	N/A	N/A	N/A	N/A
PE	BOTH	~400	2x150	N	ABySS	20	5,127,469	159,189,643	10,433	782
PE	BOTH	~400	2x150	Y	ABySS	20	3,943,117	116,452,228	10,661	900
PE	BOTH	~400	2x150	N	ABySS	31	1,453,356	72,633,295	11,442	627
PE	BOTH	~400	2x150	Y	ABySS	31	1,083,102	50,541,757	11,251	815
PE	BOTH	~400	2x150	N	ABySS	40	730,724	47,054,450	11,292	719
PE	BOTH	~400	2x150	Y	ABySS	40	528,106	31,841,130	11,301	824
PE	BOTH	~400	2x150	N	ABySS	60	207,879	19,480,417	11,537	883
PE	BOTH	~400	2x150	Y	ABySS	60	141,490	12,723,023	11,604	940
PE	BOTH	~400	2x250	N	ABySS	20	23,015,384	772,312,910	2,323	549
PE	BOTH	~400	2x250	Y	ABySS	20	13,604,775	471,023,416	1,398	569
PE	BOTH	~400	2x250	N	ABySS	31	6,590,403	464,521,175	5,071	570
PE	BOTH	~400	2x250	Y	ABySS	31	3,719,252	245,076,215	5,702	588
PE	BOTH	~400	2x250	N	ABySS	40	3,780,833	361,564,188	5,080	575
PE	BOTH	~400	2x250	Y	ABySS	40	1,939,736	166,910,958	10,646	591
PE	BOTH	~400	2x250	N	ABySS	60	1,330,929	184,143,173	10,615	598
PE	BOTH	~400	2x250	Y	ABySS	60	530,031	61,863,274	11,790	592
PE	BOTH	~400	2x150 2x250	N	ABySS	20	36,416,080	1,246,640,912	990	539
PE	BOTH	~400	2x150 2x250	Y	ABySS	20	25,418,820	896,680,831	989	531
PE	BOTH	~400	2x150 2x250	N	ABySS	31	9,467,831	799,657,757	5,424	623
PE	BOTH	~400	2x150 2x250	Y	ABySS	31	6,770,337	548333766	5,104	643
PE	BOTH	~400	2x150 2x250	N	ABySS	40	6,126,115	697,275,751	5,113	633
PE	BOTH	~400	2x150 2x250	Y	ABySS	40	3,787,238	416017375	6,310	653
PE	BOTH	~400	2x150 2x250	N	ABySS	60	2,382,233	428,055,378	5,627	669
PE	BOTH	~400	2x150 2x250	Y	ABySS	60	1,098,378	164159351	10,997	638

Appendix 5. Contig assembly metrics for corn snake (*Pantherophis guttatus*) genome assemblies.

SE/PE	Left/Right	Library insert size	Trimmed?	Assembler	k-mer length	no. of contigs	max contig length	contig N50
PE	BOTH	300 600	Y	CLC	Default	45,659	13,799	671
PE	BOTH	300	N	ABySS	20	2.66E+07	2,064	602
PE	BOTH	300	Y	ABySS	20	2.09E+07	2,542	612
PE	BOTH	300	N	ABySS	31	1.14E+07	7,321	697
PE	BOTH	300	Y	ABySS	31	8.69E+06	5,377	678
PE	BOTH	300	N	ABySS	40	6.90E+06	6,813	713
PE	BOTH	300	Y	ABySS	40	4.89E+06	6,017	686
PE	BOTH	300	N	ABySS	60	2.71E+06	6,830	705
PE	BOTH	300	Y	ABySS	60	1.58E+06	9,551	665
PE	BOTH	600	N	ABySS	20	1.19E+07	2,972	715
PE	BOTH	600	Y	ABySS	20	1.02E+07	3,642	706
PE	BOTH	600	N	ABySS	31	5.28E+06	6,605	858
PE	BOTH	600	Y	ABySS	31	4.40E+06	11,583	847
PE	BOTH	600	N	ABySS	40	3.24E+06	8,306	903
PE	BOTH	600	Y	ABySS	40	2.62E+06	7,493	887
PE	BOTH	600	N	ABySS	60	1.17E+06	9,271	932
PE	BOTH	600	Y	ABySS	60	7.37E+05	8,985	891
PE	BOTH	300 600	N	ABySS	20	3.58E+07	2,564	629
PE	BOTH	300 600	Y	ABySS	20	2.94E+07	2,209	645
PE	BOTH	300 600	N	ABySS	31	1.52E+07	17,487	797
PE	BOTH	300 600	Y	ABySS	31	1.25E+07	11,950	798
PE	BOTH	300 600	N	ABySS	40	9.50E+06	16,167	851
PE	BOTH	300 600	Y	ABySS	40	7.41E+06	12,054	843
PE	BOTH	300 600	N	ABySS	60	3.83E+06	11,307	886
PE	BOTH	300 600	Y	ABySS	60	2.34E+06	6,384	835

Appendix 6. Scaffold assembly metrics for corn snake (*Pantherophis guttatus*) genome assemblies.

SE/PE	Left/Right	Library insert size	Trimmed?	Assembler	k-mer length	no. of scaffolds	total length (scaffs)	max scaffold length	Scaffold N50
PE	BOTH	300	N	ABySS	20	26,637,717	856,461,685	2,624	646
PE	BOTH	300	Y	ABySS	20	20,873,817	656,572,980	4,549	654
PE	BOTH	300	N	ABySS	31	11,420,030	702,003,844	7,321	714
PE	BOTH	300	Y	ABySS	31	8,684,719	507,186,397	9,724	695
PE	BOTH	300	N	ABySS	40	6,887,091	603,041,887	9,577	736
PE	BOTH	300	Y	ABySS	40	4,886,482	399,916,909	9,791	710
PE	BOTH	300	N	ABySS	60	2,690,369	376,202,120	7,687	744
PE	BOTH	300	Y	ABySS	60	1,570,391	180,798,387	9,551	716
PE	BOTH	600	N	ABySS	20	11,879,969	394,463,940	4,558	788
PE	BOTH	600	Y	ABySS	20	10,175,319	333,721,582	4,787	795
PE	BOTH	600	N	ABySS	31	5,269,707	326,896,213	7,946	904
PE	BOTH	600	Y	ABySS	31	4,390,230	266,590,526	11,901	907
PE	BOTH	600	N	ABySS	40	3,218,955	276,257,647	11,577	958
PE	BOTH	600	Y	ABySS	40	2,612,179	214,051,728	11,228	950
PE	BOTH	600	N	ABySS	60	1,159,057	141,114,734	9,271	1,005
PE	BOTH	600	Y	ABySS	60	731,164	79,136,359	12,373	962
PE	BOTH	300 600	N	ABySS	20	35,822,776	1,166,737,592	3,613	692
PE	BOTH	300 600	Y	ABySS	20	29,406,742	949,296,563	4,016	707
PE	BOTH	300 600	N	ABySS	31	15,141,582	981,684,755	17,487	858
PE	BOTH	300 600	Y	ABySS	31	12,493,389	798,761,897	14,633	849
PE	BOTH	300 600	N	ABySS	40	9,442,705	876,105,284	16,167	921
PE	BOTH	300 600	Y	ABySS	40	7,370,167	667,420,081	12,827	901
PE	BOTH	300 600	N	ABySS	60	3,776,446	590,713,310	12,407	951
PE	BOTH	300 600	Y	ABySS	60	2,314,725	306,095,596	8,630	897

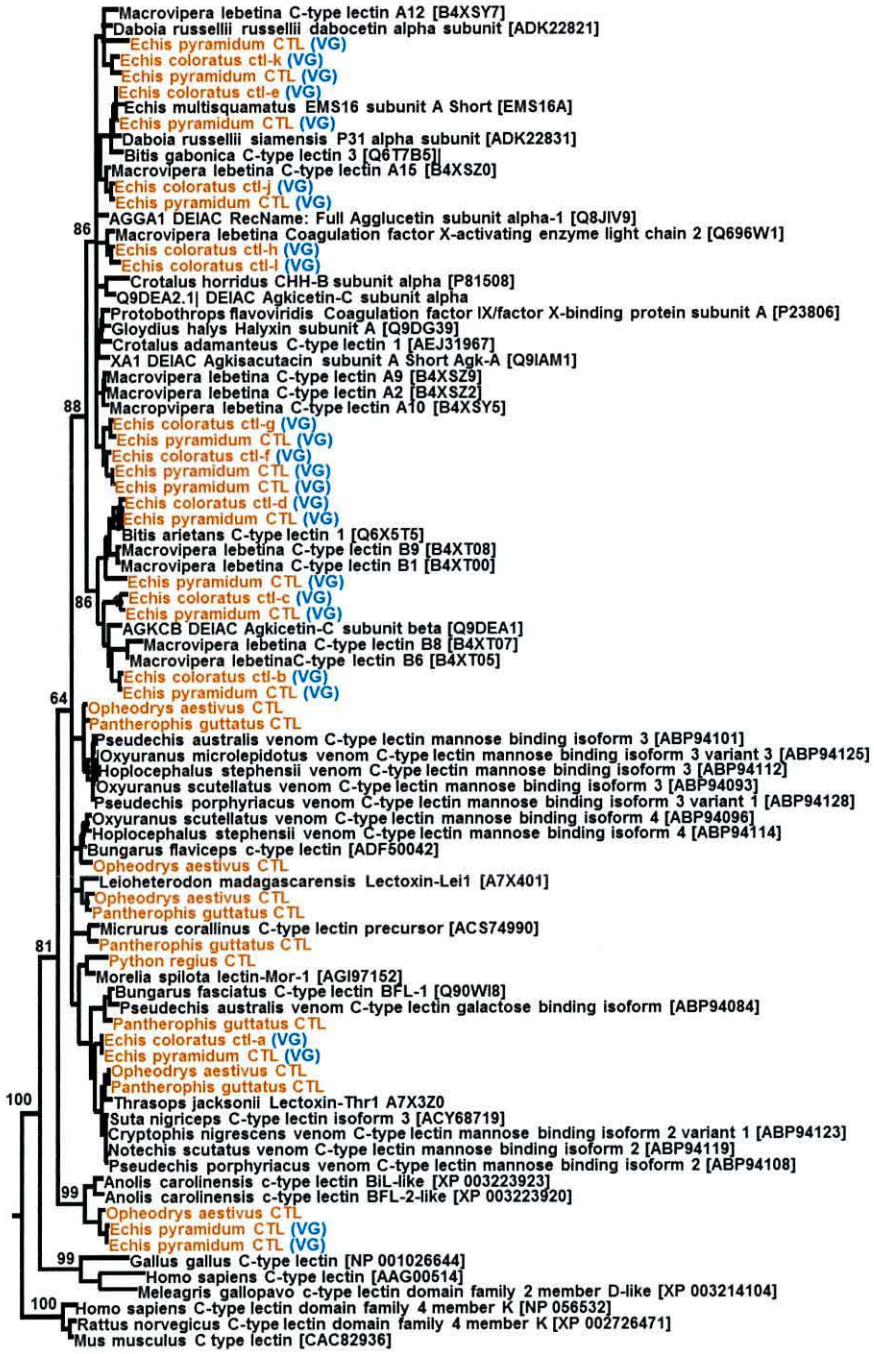
Appendix 7. CEGMA analysis results for all newly generated snake whole genome sequences.

Species	Library	Assembler	k-mer size	No. complete CEGs	Complete % completeness	No. partial CEGs	Partial % completeness
Eco	300 600	CLC	def	6	2.42	10	4.03
Eco	300	ABySS	20	1	0.4	12	4.84
Eco	300	ABySS	20	0	0	9	3.63
Eco	300	ABySS	31	6	2.42	37	14.92
Eco	300	ABySS	31	4	1.61	30	12.1
Eco	300	ABySS	40	5	2.02	34	13.71
Eco	300	ABySS	40	3	1.21	31	12.5
Eco	300	ABySS	60	5	2.02	31	12.5
Eco	300	ABySS	60	1	0.4	17	6.85
Eco	600	ABySS	31	65	26.21	177	71.37
Eco	600	ABySS	31	67	27.02	175	70.56
Eco	600	ABySS	40	77	31.05	186	75
Eco	600	ABySS	40	67	27.02	187	74.4
Eco	600	ABySS	60	63	25.40	180	72.58
Eco	600	ABySS	60	33	13.31	135	54.44
Eco	300 600	ABySS	31	70	28.23	185	74.6
Eco	300 600	ABySS	31	72	29.03	184	74.19
Eco	300 600	ABySS	40	84	33.87	192	77.42
Eco	300 600	ABySS	40	79	31.85	194	78.23
Eco	300 600	ABySS	60	81	32.66	195	78.63
Eco	300 600	ABySS	60	67	27.01	181	72.98
Epy	2x150 2x250	ABySS	31	0	0	8	3.23
Epy	2x150 2x250	ABySS	31	1	0.4	4	1.61
Epy	2x150 2x250	ABySS	40	1	0.4	7	2.82
Epy	2x150 2x250	ABySS	40	0	0	5	2.02
Epy	2x150 2x250	ABySS	60	0	0	5	2.02
Epy	2x150 2x250	ABySS	60				
Pgu	300 600	CLC	DEF	0	0	3	1.21
Pgu	300	ABySS	20	1	0.4	11	4.44
Pgu	300	ABySS	20	0	0	10	4.03

Pgu	300	ABySS	31	8	3.23	21	8.47
Pgu	300	ABySS	31	4	1.61	19	7.66
Pgu	300	ABySS	40	7	2.82	23	9.27
Pgu	300	ABySS	40	3	1.21	12	4.84
Pgu	300	ABySS	60	3	1.21	13	5.24
Pgu	300	ABySS	60				
Pgu	600	ABySS	20	1	0.4	6	2.42
Pgu	600	ABySS	20	0	0	5	2.02
Pgu	600	ABySS	31	2	0.81	10	4.03
Pgu	600	ABySS	31	1	0.4	7	2.82
Pgu	600	ABySS	40	2	0.81	8	3.23
Pgu	600	ABySS	40	1	0.4	7	2.82
Pgu	600	ABySS	60	0	0	3	1.21
Pgu	600	ABySS	60				
Pgu	300 600	ABySS	20	0	0	3	1.21
Pgu	300 600	ABySS	20	1	0.4	11	4.44
Pgu	300 600	ABySS	31	11	4.44	32	12.9
Pgu	300 600	ABySS	31	9	3.63	30	12.1
Pgu	300 600	ABySS	40	12	4.84	34	13.71
Pgu	300 600	ABySS	40	9	3.63	26	10.48
Pgu	300 600	ABySS	60	9	3.63	25	10.08
Pgu	300 600	ABySS	60	1	0.4	10	4.03

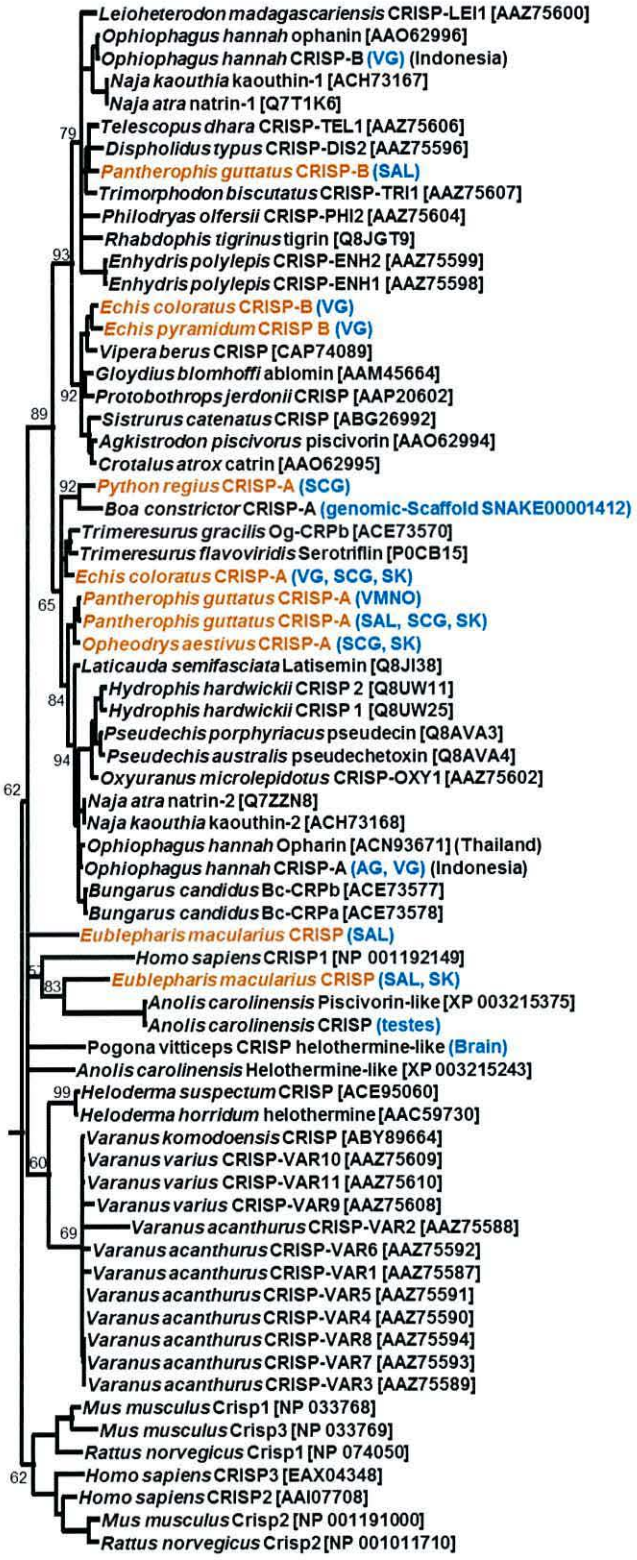
Appendix 8. Maximum likelihood tree of snake venom metalloproteinase (*svmp*) sequences, including *Echis pyramidum* sequences.

Appendix 9. Maximum likelihood tree of C-type lectin (*ctl*) sequences, including *Echis pyramidum* sequences.

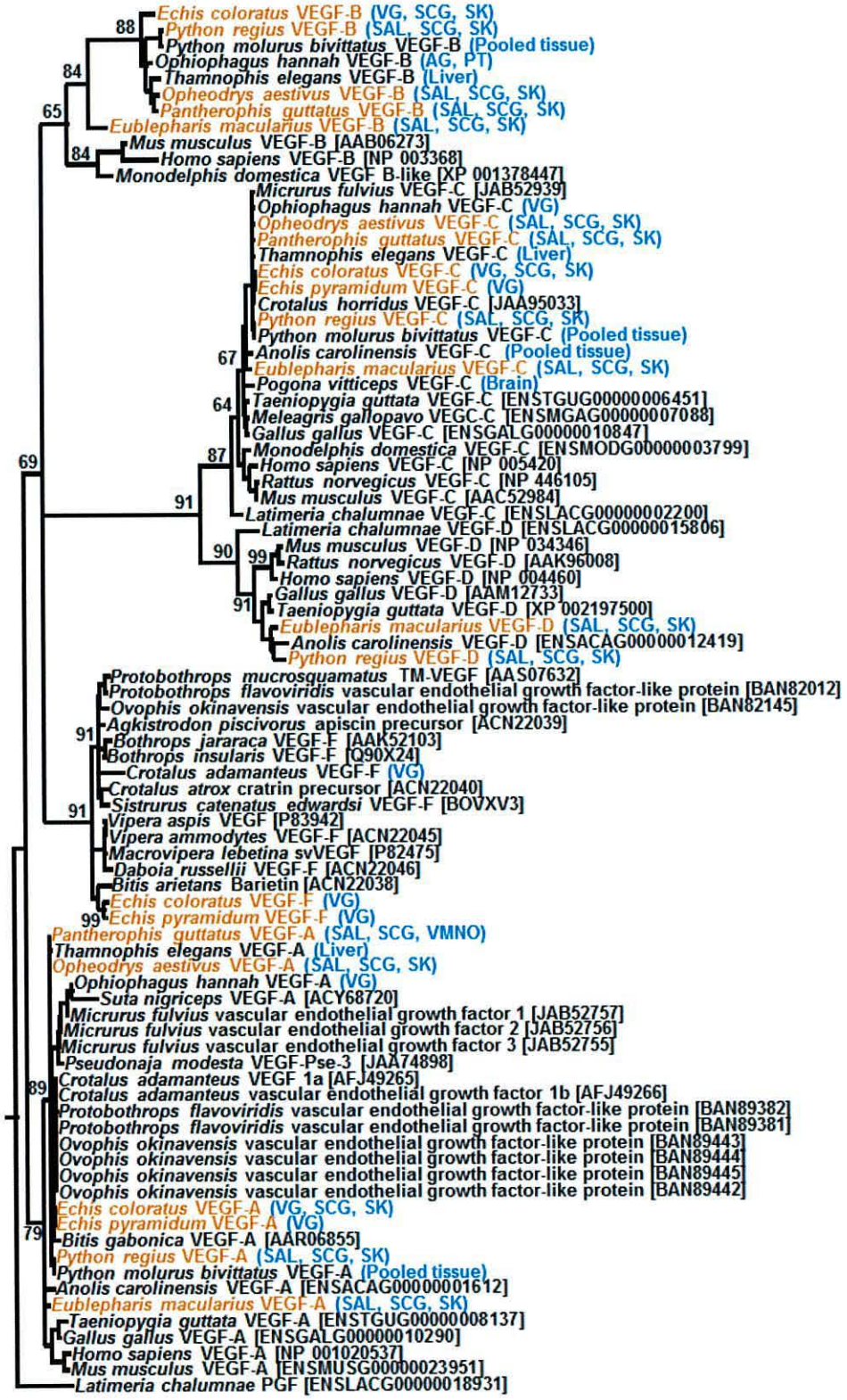


Appendix 10. Maximum likelihood tree of serine protease (*sp*) sequences, including *Echis pyramidum* sequences.

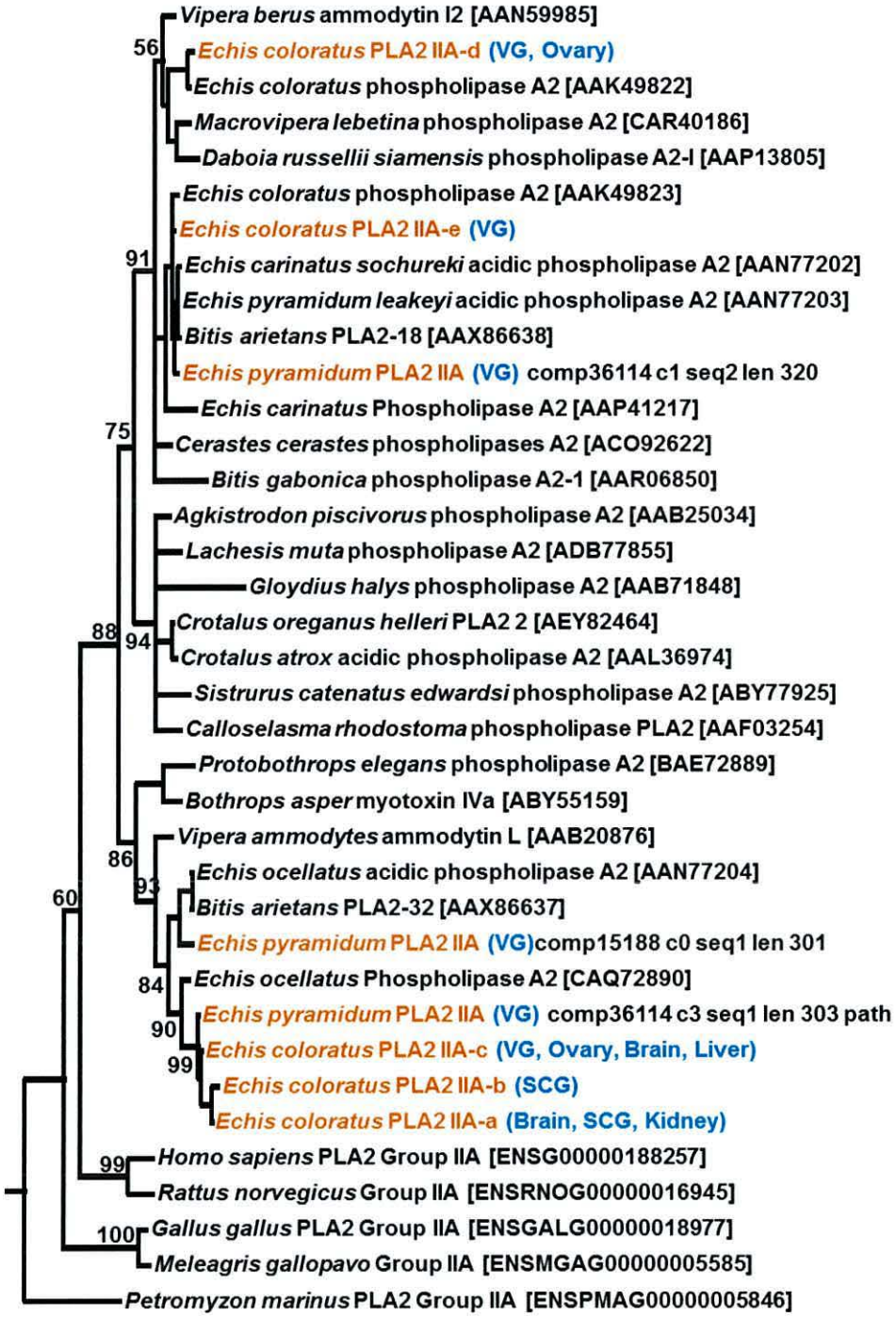
Appendix 11. Maximum likelihood tree of cysteine-rich secretory protein (*crisp*) sequences, including *Echis pyramidum* sequences.



Appendix 12. Maximum likelihood tree of vascular endothelial growth factor (*vegf*) sequences, including *Echis pyramidum* sequences.



Appendix 13. Maximum likelihood tree of Group IIA phospholipase A_2 (PLA_2 group IIA) sequences, including *Echis pyramidum* sequences.



Appendix 14. Sequencing and assembly metrics for tissue assemblies, based on two individuals per tissue for all samples except Corn snake skin (see methods). Eco, Painted saw-scaled viper (*Echis coloratus*); Pgu, Corn snake (*Pantherophis guttatus*); Oae, Rough green snake (*Opheodrys aestivus*); Pre, Royal python (*Python regius*); Ema, Leopard gecko (*Eublepharis macularius*). VG, venom gland (pooled 24 and 48 hours post-milking); SAL, salivary gland; SCG, scent gland; SK, skin.

Species	Tissue	Total number of PE reads	Total number of bases	Number of contigs	Number of contigs >300bp	Contig N50 (bp)	Max contig (bp)
Eco	VG	26,642,227	5,328,445,400	84,846	56,805	1,684	15,819
	SCG	27,206,987	5,441,397,400	138,852	87,389	2,444	36,612
	SK	14,166,420	2,833,284,000	77,402	50,860	1,725	30,610
Pgu	SAL	25,655,661	5,131,132,200	64,595	43,565	1,916	17,102
	SCG	25,982,209	5,196,441,800	110,016	70,265	2,345	17,065
	SK	7,862,371	1,572,474,200	51,199	35,969	1,329	19,348
Oae	SAL	24,959,242	4,991,848,400	65,393	42,558	1,780	17,524
	SCG	28,136,146	5,627,229,200	126,321	77,954	2,064	17,135
	SK	16,398,925	3,279,785,000	92,597	57,242	1,918	33,155
Pre	SAL	28,035,045	5,607,009,000	73,492	48,727	2,265	33,655
	SCG	24,575,003	4,915,000,600	163,065	104,329	3,156	20,966
	SK	15,643,272	3,128,654,400	67,200	47,819	1,864	25,879
Ema	SAL	29,882,110	5,976,422,000	111,345	73,027	2,439	24,285
	SCG	25,502,521	5,100,504,200	129,951	85,014	2,330	29,392
	SK	14,675,568	2,935,113,600	92,506	61,456	2,057	27,091

Appendix 15. Sequencing metrics for additional Painted saw-scaled viper (*Echis coloratus*) venom gland samples used for RSEM abundance estimation.

Time post-milking	Total number of PE reads	Total number of bases
16 hours	38,711,180	7,819,658,360
24 hours	44,678,609	9,025,079,018

Appendix 16. Sequence and assembly metrics for King cobra (*Ophiophagus hannah*) venom gland, accessory gland and pooled tissue (heart, lung, spleen, brain, testes, gall bladder, pancreas, small intestine, kidney, liver, eye, tongue and stomach) data from Vonk et al. (2013).

Tissue	Total number of SE reads	Total number of bases	Number of contigs	Number of contigs >300bp	Contig N50 (bp)	Max contig (bp)
Venom gland	15,166,590	834,162,450	6123	2,925	424	4,585
Accessory gland	11,209,677	616,532,235	9046	4,113	377	3,740
Pooled tissue	17,858,289	910,772,739	8877	4,135	413	5,733

Appendix 17. Sequencing and assembly metrics for reference transcriptome assemblies used for transcript abundance estimation.

Reference assembly	Total number of assembled bases	Number of contigs	Number of contigs >300bp	Contig N50 (bp)	Max contig (bp)
EcoTissueRef	228,063,624	206,147	149,821	2,445	38,875
PguTissueRef	166,312,211	152,359	112,327	2,315	23,951
OaeTissueRef	210,451,256	204,942	147,597	2,200	52,645
PreTissueRef	301,328,500	219,070	166,578	3,407	34,165
EmaTissueRef	266,096,501	228,645	167,623	2,746	33,255

Appendix 18. Transcript abundance estimation values given in FPKM for each Leopard gecko (*Eublepharis macularius*) tissue.

Gene	Salivary gland			Scent gland			Skin		
	1	2	Mean	1	2	Mean	1	2	Mean
ache transcript 1	9.08	8.18	8.63	6.62	2.36	4.49	5.48	2.65	4.07
complement c3	121.32	0.06	60.69	53.85	22.52	38.19	7.85	82.24	45.05
cystatin-e/m	14.97	11.57	13.27	37.6	57.06	47.33	20.15	25.77	22.96
cystatin-f	1.21	1.79	1.50	0.36	1.1	0.73	0.66	4.26	2.46
laao-a	0	4.78	2.39	8.48	5.58	7.03	15.47	50.14	32.81
laao-a	109.19	16.39	62.79	259.98	3.31	131.65	1.71	0.26	0.99
esp-e1	47.71	45.91	46.81	26.36	50.32	38.34	35.12	38.27	36.70
ficolin	7.07	0.9	3.99	5.16	5.43	5.30	0.21	0.95	0.58
kallikrein	1.53	0.12	0.83	4.28	1.83	3.06	-	-	-
kunitz 1	2.06	0.24	1.15	0	3.87	1.94	0	1.11	0.56
kunitz 2	103.1	102	102.55	140.59	192.61	166.6	111.37	111.52	111.45
lipa-a	5.69	6.31	6.00	6.57	8.84	7.71	14.05	25.33	19.69
ngf	2.67	2.16	2.42	1.34	2.08	1.71	1.87	3.06	2.47
PLA_2 group IIE	9.41	22.27	15.84	-	-	-	-	-	-
plb	4.68	6.25	5.47	44.19	53.66	48.93	72.02	91.27	81.65
renin	-	-	-	1	7.01	4.01	-	-	-
vegf-a	2.34	0.82	1.58	5.38	1.24	3.31	2.89	2.19	2.54
vegf-b	1.61	0.97	1.29	1.27	1.63	1.45	1.86	1.53	1.70
vegf-c	1.83	1.81	1.82	1.35	2.15	1.75	0.39	1.2	0.80
vespryn	-	-	-	7.29	10.72	9.01	1.42	29.91	15.67
waprin	74.86	81.24	78.05	22.71	25.55	24.13	4.25	2.52	3.39

Appendix 19. Transcript abundance estimation values given in FPKM for each Royal python (*Python regius*) tissue.

Gene	Salivary gland			Scent gland			Skin		
	1	2	Mean	1	2	Mean	1	2	Mean
3ftx-a	6.16	2.14	4.15	-	-	-	-	-	-
ache transcript 1	0	1.47	0.74	0	0	0	2.74	1.25	2.00
ache transcript 2	-	-	-	0.82	1.12	0.97	-	-	-
complement c3	26.23	1.61	13.92	1.2	2.2	1.70	-	-	-
crisp-a	-	-	-	6.39	4.92	5.66	-	-	-
cystatin-e/m	270.35	135.03	202.69	861.58	681.05	771.32	35.3	49.76	42.53
cystatin-f	1.61	0.8	1.21	5.1	5.7	5.40	0.73	0.37	0.55
dpp3	6.84	5.16	6.00	28.07	28.3	28.19	25.51	25.75	25.63
dpp4	124.71	33.41	79.06	20.01	24.63	22.32	2.72	4.23	3.48
esp-e1	2.54	1.16	1.85	9.3	10.85	10.08	2.99	5.4	4.20
ficolin	3.76	0	1.88	-	-	-	-	-	-
kallikrein	0.34	0	0.17	1.35	2.08	1.72	0	16.9	8.45
kunitz 1	5.36	4.49	4.93	13.45	16.6	15.03	4.45	3	3.73
kunitz 2	35.72	17.57	26.65	50.1	53	51.55	44.08	60.63	52.36
laao-b1	-	-	-	1,066.83	1,242.33	1,154.58	-	-	-
lipa-a	0.95	0.92	0.94	3.49	2.41	2.95	32.37	41.21	36.79
lipa-b	-	-	-	4.13	3.88	4.01	-	-	-
ngf	-	-	-	-	-	-	5.03	2.15	3.59
PLA_2 group IIE	138.12	111.14	124.63	-	-	-	-	-	-
plb	0	2.56	1.28	16.8	12.24	14.52	16.33	34.61	25.47
vegf-a	0.82	0.09	0.46	2.98	2.5	2.74	0.36	0.15	0.26
vegf-b	3.93	3.97	3.95	1.95	0.85	1.40	0	0.86	0.43
vegf-c	0.94	0.58	0.76	17.38	10.43	13.91	0.46	0	0.23
vespryn	-	-	-	15.3	4.24	9.77	99.86	141.84	120.85

Appendix 20. Transcript abundance estimation values given in FPKM for each Rough green snake (*Opheodrys aestivus*) tissue.

	Salivary gland			Scent gland			Skin		
Gene	1	2	Mean	1	2	Mean	1	2	Mean
3ftx-a	120.53	0.36	60.45	-	-	-	1.67	0	0.84
3ftx-b	-	-	-	0	10.59	5.30	-	-	-
ache transcript 1	2.43	0	1.22	8.48	9.7	9.09	-	-	-
ache transcript 2	-	-	-	8.48	9.7	9.09	-	-	-
AVIT	-	-	-	0	0.91	0.46	-	-	-
complement c3	18.29	1.63	9.96	102.28	52.24	77.26	0.81	8.77	4.79
crisp-a	-	-	-	0	5.3	2.65	0	0	0
cystatin-e/m	168.58	213.23	190.91	25.12	18.03	21.58	23.94	12.78	18.36
cystatin-f	0.59	0	0.30	1.75	2.27	2.01	0.14	0.47	0.31
dpp3	0	6.34	3.17	0	3.37	1.69	0	6.43	3.22
dpp4	4.33	5.61	4.97	9.61	9.64	9.63	2.59	2.27	2.43
esp-e1	20.13	11.16	15.65	76.01	84.29	80.15	17.44	20.41	18.93
ficolin	17.32	1.93	9.63	269.16	1.6	135.38	-	-	-
kallikrein	-	-	-	1.34	8.11	4.73	-	-	-
kunitz 1	17.35	26.32	21.84	22.93	60.7	41.82	2.26	2.91	2.59
kunitz 2	60.5	79.52	70.01	144.05	309.27	226.66	18.91	36.53	27.72
laao-b1	-	-	-	2.44	0.44	1.44	-	-	-
lipa-a	7.4	2.87	5.14	26.2	44.7	35.45	11.74	6.66	9.20
ngf	-	-	-	2.8	0.74	1.77	1.15	1.11	1.13
PLA_2 group IIE	8.28	24.89	16.59	-	-	-	-	-	-
plb	-	-	-	9.11	34.64	21.88	7.38	42.55	24.97
vegf-a	0.25	3.57	1.91	0.6	4.44	2.52	0.7	0.78	0.74
vegf-b	1.17	0.72	0.95	1.06	0.92	0.99	0.59	0	0.29
vegf-c	1.17	1.42	1.30	3.85	2.04	2.95	1.21	1.17	1.19
vespryn	-	-	-	0.31	2	1.16	0	0.39	0.19
waprin	41.85	41.26	41.56	277.5	75.5	176.50	3.4	0.58	1.99

Appendix 21. Transcript abundance estimation values given in FPKM for each Corn snake (*Pantherophis guttatus*) tissue.

	Salivary gland			Scent gland			Skin
Gene	1	2	Mean	1	2	Mean	1
3ftx-a	7.00	0.57	3.79	0	1.04	0.52	6.15
ache transcript 1	3.31	2.97	3.14	1.07	1.92	1.50	1.12
complement c3	41.33	76.08	58.71	2.12	3.27	2.70	0.49
crisp-a	0	4.48	2.24	3.11	7.95	5.53	37.27
crisp-b	0.11	13.15	6.63	-	-	-	-
cystatin-e/m	0.85	61.01	30.93	204.32	657.41	430.87	90.14
cystatin-f	1.05	0.77	0.91	1.42	1.4	1.41	1.58
dpp3	2.43	12.31	7.37	9.58	0	4.79	2.49
dpp4	17.67	21.74	19.71	5.06	2.08	3.57	6.98
esp-e1	10.58	13.77	12.18	20.53	68.15	44.34	23.83
ficolin	8.67	39.06	23.87	6.75	13.22	9.99	2.98
kallikrein	-	-	-	4.52	72.05	38.29	-
kunitz 1	12.2	14.89	13.55	46.68	16.63	31.66	2.27
kunitz 2	50.24	112.88	81.56	121.72	79.8	100.76	65.64
laao-b1	-	-	-	97.78	0.26	49.02	-
lipa-a	1.04	0	0.52	19.19	0.1	9.65	6.96
ngf	0.9	0.5	0.70	2.09	0	1.05	-
PLA_2 group IIE	49.24	53.99	51.62	-	-	-	-
Pgu svmp-a	0.42	4.13	2.28	-	-	-	-
Pgu svmp-b	0	7.05	3.53	-	-	-	-
plb	1.06	5.52	3.29	17.96	146.36	82.16	20.76
vegf-a	0.88	0.22	0.55	1.01	0.86	0.94	-
vegf-b	2.37	4.97	3.67	3.84	0.22	2.03	1.88
vegf-c	1.27	1.13	1.20	6.44	0.63	3.54	1.93
vespryn	2.58	1.47	2.03	4.97	5.07	5.02	-
waprin	72.16	34.98	53.57	-	-	-	-

Appendix 22. Transcript abundance estimation values given in FPKM for each Painted saw-scaled viper (*Echis coloratus*) tissue.

	Venom gland					Scent gland			Skin		
Gene	1	2	3	4	Mean	1	2	Mean	1	2	Mean
3ftx-a	6.04	199.4	18.33	1,234.1	364.47	28.68	368.33	198.51	3.61	11.84	7.73
3ftx-b	0.15	3.42	316.94	367.29	171.95	1,491.4	1.21	746.31	-	-	-
ache transcript 1	6.02	10.07	3.61	6.68	6.60	4.75	3.4	4.07	-	-	-
ache transcript 2	-	-	-	-	-	0.52	0.33	0.43	1.16	0.38	0.77
complement c3	1.24	22.41	29.94	180.95	58.64	6.9	495.67	251.29	0.99	0.88	0.94
crisp-a	0.04	0	0.96	1.45	0.61	0.12	72.76	36.44	35.54	0	17.77
crisb-b	7,791.91	1,422.35	4,851.14	14.88	3,520.07	-	-	-	-	-	-
crotamine-like	0	29.66	0	14.15	10.95	-	-	-	-	-	-
ctl-a	1.23	127.05	50.80	741.17	230.06	1.34	168.13	84.73			
ctl-b	25,656.2	555.52	70,262.94	14.97	24,122.41	-	-	-	-	-	-
ctl-c	4,202.8	319.98	10,294.96	7.1	3,706.21	-	-	-	-	-	-
ctl-d	8,795.51	310.09	19,238.23	177.65	7,130.37	-	-	-	-	-	-
ctl-e	2,694.03	393.99	2,814.17	1.33	1,475.88	-	-	-	-	-	-
ctl-f	11,158.35	428.25	15,385	147.73	6,779.83	-	-	-	-	-	-
ctl-g	5,605.99	13.86	14,010.27	13.48	4,910.90	-	-	-	-	-	-
ctl-h	1,203.55	139.1	144.89	1.58	372.28	-	-	-	-	-	-
ctl-i	1.73	0	1.46	0	0.80	-	-	-	-	-	-
ctl-j	21,250.16	725.85	40,076.36	88.99	15,535.34	-	-	-	-	-	-
ctl-k	1.29	1.54	4.85	0	1.92	-	-	-	-	-	-
cystatin-e/m	12.38	158.77	5.11	19.45	48.93	270.94	36.2	153.57	55.64	186.06	120.85
cystatin-f	0.1	2.7	0.38	4.01	1.79	0.22	2.95	1.59	-	-	-
dpp3	15.47	2.06	5.81	0.98	6.08	4.88	3.32	4.10	5.75	1.22	3.49
dpp4	6.74	1.74	3.2	0.1	2.95	8.58	1.53	5.06	1.46	2.46	1.96
esp-e1	3.27	10.52	6.17	16.39	9.09	12.87	14.55	13.71	2.46	14.82	8.64
ficolin	0.15	55.94	10.76	241.81	77.17	4.3	117.87	61.09	4.55	6.4	5.48
kallikrein	0	0.27	0.05	4.32	1.16	23.08	5.6	14.34	79.86	0	39.93

kunitz 1	3.87	1.64	4.28	3.71	3.38	1.83	8.4	5.12	1.26	2.09	1.68
kunitz 2	176.97	58.15	128.89	78.54	110.64	83.65	134.46	109.06	54.04	8.95	31.49
laao-a	2.57	3.24	2.97	4.58	3.34	5.22	6.01	5.62	3.16	2.62	2.89
laao-b1	107.96	22.6	85.47	0	54.01	-	-	-	-	-	-
laao-b2	776.96	20.01	1,710.19	8.18	628.84	-	-	-	-	-	-
lipa-a	5,080.21	942.08	7,259.94	67.08	3,337.33	940.15	28.83	484.49	21.9	23.67	22.79
lipa-b	0	0	10.01	1.83	2.96	6.02	4.55	5.29	-	-	-
ngf	730.19	1.98	1,370.8	0.3	525.82	0.08	0.28	0.18	0.28	0.88	0.58
PLA_2 IIA-a	-	-	-	-	-	0.21	2.48	1.35	-	-	-
PLA_2 IIA-b	-	-	-	-	-	0.21	2.48	1.35	-	-	-
PLA_2 IIA-c	35,965.07	1,285.85	52,775.79	54.91	22,520.41	-	-	-	-	-	-
PLA_2 IIA-d	3,312.33	141.9	3,253.47	0.89	1,677.15	-	-	-	-	-	-
PLA_2 IIA-e	914.26	36.52	787.89	0	434.67	-	-	-	-	-	-
PLA_2 group IIE	27.5	15.45	5.21	0.32	12.12	-	-	-	-	-	-
plb	196.04	32.79	191.94	37.14	114.48	30.18	11.51	20.85	99.02	8.01	53.52
renin	27.39	2.53	17.88	0.5	12.08	0	0	0	0	-	-
sp-a	3,726.15	8.9	17,693.55	0	5,357.15	-	-	-	-	-	-
sp-b	7,471.91	17.79	23,258.4	0	7,687.03	-	-	-	-	-	-
sp-c	2,525.53	131.58	9,646.92	0	3,076.01	-	-	-	-	-	-
sp-d	2,079.25	1,205.26	1,108.08	1.2	1,098.45	-	-	-	-	-	-
sp-e	2,742.78	1,131.13	16,149.18	0	5,005.77	-	-	-	-	-	-
sp-f	39.22	115.65	254.5	0	102.34	-	-	-	-	-	-
svmp-a	7,881.36	2.45	14,258.65	68.91	5,552.84	-	-	-	-	-	-
svmp-b	49,469.53	2,682.3	8,163.04	158.76	15,118.41	-	-	-	-	-	-
svmp-c	149.97	29.9	1.45	2.32	45.91	-	-	-	-	-	-
svmp-d	19,267.95	845.58	3,839.82	84.51	6,009.47	-	-	-	-	-	-
svmp-e	2,772.71	472.08	2,641.5	243.87	1,532.54	-	-	-	-	-	-
svmp-f	0.66	0	1.23	8.04	2.48	-	-	-	-	-	-
svmp-g	2,772.71	472.08	2,641.5	243.87	1,532.54	-	-	-	-	-	-
svmp-h	-	-	-	-	-	0.98	0	0.49			
svmp-i	16,986.24	974.82	20,413.99	705.04	9,770.02	-	-	-	-	-	-

svmp-j	2,772.71	472.08	2,641.5	243.87	1,532.54	-	-	-	-	-	-
svmp-k	2,772.71	472.08	2,641.5	243.87	1,532.54	-	-	-	-	-	-
svmp-l	-	-	-	-	-	-	-	-	5.54	0	2.77
svmp-m	251.21	0	294.07	0	136.32	-	-	-	-	-	-
svmp-n	2,772.71	472.08	2,641.5	243.87	1,532.54	0.06	0	0.03	-	-	-
svmp-o	-	-	-	-	-	0	10.7	5.35	-	-	-
svmp-p	1,273.16	103.84	2,546.07	4.67	981.94	-	-	-	-	-	-
svmp-q	15,356.06	2,247.57	3,951.69	313.12	5,467.11	-	-	-	-	-	-
svmp-r	-	-	-	-	-	0.06	0	0.03	-	-	-
svmp-s	-	-	-	-	-	0.06	0	0.03	-	-	-
svmp-t	4.4	0.68	25.3	1.56	7.99	-	-	-	-	-	-
svmp-u	-	-	-	-	-	0	0.26	0.13	-	-	-
vegf-a	7.64	0.27	4.1	0.96	3.24	2.62	0	1.31	0.57	1.18	0.88
vegf-b	1.05	1.36	0.57	2.12	1.28	4.92	5.83	5.38	0.47	3.21	1.84
vegf-c	1.47	1.74	1.02	1.92	1.54	2.37	2.06	2.22	0.57	1.81	1.19
vegf-f	294.23	362.01	90.07	0.62	186.73	-	-	-	-	-	-
waprin	0.35	19.91	9.82	73.58	25.92	-	-	-	1.59	2.04	1.82

Appendix 23. Sequencing and assembly metrics for tissue assemblies, based on two individuals per tissue for leopard gecko (*Eublepharis macularius*), rough green snake (*Opheodrys aestivus*) and royal python (*Python regius*) skin, scent glands and salivary glands and corn snake scent gland and salivary gland. Only a single corn snake skin sample provided RNA of high enough quality for sequencing. *Echis coloratus* values are derived from four adult individuals for venom gland, two adult individuals for skin, scent gland, kidney and brain and one individual for liver and ovary. SAL, salivary gland; SCG, scent gland; VG, venom gland; KID, kidney; PE, paired-end.

Species	Tissue	Total number of PE reads	Total number of bases	Number of contigs	Number of contigs >300bp	Contig N50 (bp)	Max contig (bp)
Pgu	SAL	25,655,661	5,131,132,200	64,595	43,565	1,916	17,102
	SCG	25,982,209	5,196,441,800	110,016	70,265	2,345	17,065
	Skin	7,862,371	1,572,474,200	51,199	35,969	1,329	19,348
Oae	SAL	24,959,242	4,991,848,400	65,393	42,558	1,780	17,524
	SCG	28,136,146	5,627,229,200	126,321	77,954	2,064	17,135
	Skin	16,398,925	3,279,785,000	92,597	57,242	1,918	33,155
Pre	SAL	28,035,045	5,607,009,000	73,492	48,727	2,265	33,655
	SCG	24,575,003	4,915,000,600	163,065	104,329	3,156	20,966
	Skin	15,643,272	3,128,654,400	67,200	47,819	1,864	25,879
Ema	SAL	29,882,110	5,976,422,000	111,345	73,027	2,439	24,285
	SCG	25,502,521	5,100,504,200	129,951	85,014	2,330	29,392
	Skin	14,675,568	2,935,113,600	92,506	61,456	2,057	27,091
Eco	VG	110,032,016	22,173,182,778	117,125	81,798	2,623	30,131
	SCG	27,206,987	5,441,397,400	138,852	87,389	2,444	36,612
	Skin	14,166,420	2,833,284,000	77,402	50,860	1,725	30,610
	Ovary	18,155,364	3,667,383,528	81,682	52,264	2,023	16,376
	KID	41,148,101	8,311,916,402	120,728	76,660	2,044	12,930
	Brain	31,934,884	6,450,846,568	195,958	134,236	3,552	20,155
	Liver	7,095,517	1,433,294,434	31,205	18,169	950	9,939

Appendix 24. Sequence and assembly metrics for king cobra (*Ophiophagus hannah*) venom gland, accessory gland and pooled tissue (heart, lung, spleen, brain, testes, gall bladder, pancreas, small intestine, kidney, liver, eye, tongue and stomach) data from Vonk et al. (2013).

Tissue	Total number of reads	Total number of bases	Number of contigs	Number of contigs >300bp	Contig N50 (bp)	Max contig (bp)
Venom gland	15,166,590	834,162,450	6123	2,925	424	4,585
Accessory gland	11,209,677	616,532,235	9046	4,113	377	3,740
Pooled tissue	17,858,289	910,772,739	8877	4,135	413	5,733

Appendix 25. Assembly statistics for transcriptomes used in Chapter 6.

Species	Tissue	Number of PE reads	Number of raw sequencing bases	Number of contigs	Number of contigs >300bp	Contig N50 (bp)	Max contig (bp)
Eco	VG	110,032,016	22,173,182,778	117,125	81,798	2,623	30,131
	SK	14,166,420	2,833,284,000	77,402	50,860	1,725	30,610
	SCG	27,206,987	5,441,397,400	138,852	87,389	2,444	36,612
	Liver	7,095,517	1,433,294,434	31,205	18,169	950	9,939
	Brain	31,934,884	6,450,846,568	195,958	134,236	3,552	20,155
	Ovary	18,155,364	3,667,383,528	81,682	52,264	2,023	16,376
	KID	41,148,101	8,311,916,402	120,728	76,660	2,044	12,930
Epy	VG	57,398,601	11,594,517,402	127,519	86,953	2,898	29,658
Pgu	SAL	25,655,661	5,131,132,200	64,595	43,565	1,916	17,102
Oae	SAL	24,959,242	4,991,848,400	65,393	42,558	1,780	17,524
Pre	SAL	28,035,045	5,607,009,000	73,492	48,727	2,265	33,655
Ema	SAL	29,882,110	5,976,422,000	111,345	73,027	2,439	24,285
Oha	VG	15,166,590*	834,162,450	6,123	2,925	424	4,585
	AG	11,209,677*	616,532,235	9,046	4,113	377	3,740
	PT	17,858,289*	910,772,739	8,877	4,135	413	5,733

Species abbreviations:
Eco, painted saw-scaled viper (Echis coloratus); Epy, Egyptian saw-scaled viper (Echis pyramidum); Pgu, corn snake (*Pantherophis guttatus*); Oae, rough green snake (*Opheodrys aestivus*); Pre, royal python (*Python regius*); Ema, leopard gecko (*Eublepharis macularius*); Oha, king cobra (*Ophiophagus hannah*).

Tissue abbreviations:
VG, venom gland; SK, skin; KID, kidney; SAL, salivary gland; AG, accessory gland; PT, pooled tissue.

*These values are for single-end sequencing reads.

Appendix 26. Predicted open reading frame statistics and details of BLAST-based gene ontology (GO) annotation of venom and salivary gland transcriptomes.

	Total contigs	Number of ORFs	Mean ORF length (nt)	Max ORF length (nt)	Number of ORFs with signal peptide	Number of ORFs with signal peptide and BLAST result	Number of ORFs with GO annotation
Eco	56,805	56,761	459	13,642	2,655 (4.68%)	1,341 (2.36%)	896 (1.58%)
Epy	86,953	86,908	699	28,315	4,574 (5.26%)	2,590 (2.98%)	2,020 (2.32%)
Pgu	43,565	43,534	548	12,139	2,197 (5.05%)	1,252 (2.88%)	909 (2.09%)
Oae	42,558	42,534	502	14,314	1,908 (4.49%)	916 (2.15%)	702 (1.65%)
Pre	48,727	48,690	544	33,010	2,247 (4.61%)	1,097 (2.25%)	868 (1.78%)
Ema	73,027	72,980	540	24,064	3,702 (5.07%)	1,856 (2.54%)	1,436 (1.97%)
Oha VG	6,123	6,102	233	2,896	227 (3.72%)	102 (1.67%)	82 (1.34%)
Oha AG	9,046	9,023	234	3,454	353 (3.91%)		
Mfu	2,066	2,066	1,233	2,066	220 (10.64%)	190 (9.19%)	165 (7.98%)
Cad	12,694	12,694	874	11,752	771 (6.07%)	538 (4.23%)	411 (3.23%)

Species abbreviations:
Eco, painted saw-scaled viper (Echis coloratus); Epy, Egyptian saw-scaled viper (Echis pyramidum); Pgu, corn snake (*Pantherophis guttatus*); Oae, rough green snake (*Opheodrys aestivus*); Pre, royal python (*Python regius*); Ema, leopard gecko (*Eublepharis macularius*); Oha, king cobra (*Ophiophagus hannah*); Mfu, Eastern coral snake (*Micrurus fulvius*); Cad, Eastern diamondback rattlesnake (*Crotalus adamanteus*)

Appendix 27. Predicted open reading frame statistics and details of BLAST-based gene ontology (GO) annotation of painted saw-scaled viper (*Echis coloratus*) tissue transcriptomes.

	Total contigs	Number of ORFs	Mean ORF length (nt)	Max ORF length (nt)	Number of ORFs with signal peptide	Number of ORFs with signal peptide and BLAST result	Number of ORFs with GO annotation
VG	44,470	44,445	478	7,705	2,146 (4.83%)	1,367 (3.08%)	984 (2.21%)
SCG	67,857	67,813	557	23,131	3,182 (4.69%)	2,083 (3.07%)	1,581 (2.33%)
SK	44,805	44,760	479	28,480	1,994 (4.45%)	1,299 (2.90%)	972 (2.17%)
Brain	78,074	78,022	615	13,945	3,878 (4.97%)	2,122 (2.72%)	1,694 (2.17%)
Kidney	51,969	51,942	456	13,990	2,257 (4.34%)	1,070 (2.06%)	803 (1.55%)
Ovary	52,264	52,227	584	13,765	2,643 (5.06%)	1,470 (2.81%)	1,159 (2.22%)
Liver	18,169	18,159	346	6,775	881 (4.85%)	450 (2.48%)	368 (2.03%)

Tissue abbreviations:
VG, venom gland; SCG, scent gland; SK, skin.

Appendix 28. Predicted open reading frame statistics and details of BLAST-based gene ontology (GO) annotation of painted saw-scaled viper (*Echis coloratus*) venom gland transcriptomes taken at different timepoints following milking.

	Total contigs	Number of ORFs	Mean ORF length (nt)	Max ORF length (nt)	Number of ORFs with signal peptide	Number of ORFs with signal peptide and BLAST result	Number of ORFs with GO annotation
Eco 8	53,786	53,744	631	15,070	2,629 (4.89%)	1,662 (3.09%)	1,269 (2.36%)
Eco 6	44,470	44,445	478	7,705	2,146 (4.83%)	1,347 (3.03%)	984 (2.21%)
Eco 7	51,505	51,472	624	20,434	2,796 (5.43%)	1,894 (3.68%)	1,412 (2.74%)
Eco 215	48,321	48,284	429	6,622	2,387 (4.94%)	1,483 (3.07%)	1,009 (2.09%)

Appendix 29. Transcriptome metrics and details of BLAST-based gene ontology (GO) annotation of venom gland sequences which are unique to a specific timepoint/sample post-venom extraction.

Venom gland	Time post-milking	Total number of transcripts expressed	Number of unique transcripts in sample	Number of unique transcripts with BLAST result	Number of unique transcripts with GO annotation
Eco 8	16 hr	28,448	5,082	2,203	1,496
Eco 7	24 hr	24,197	1,707	931	641
Eco 6	24 hr	37,834	7,325	2,701	1,727
Eco 215	48 hr	42,662	12,535	3,885	2,355

Appendix 30. Assembly metrics for the genome of the painted saw-scaled viper, *Echis coloratus*

Total PE reads	Raw sequencing bases	Total contigs	Max contig length	Contig N50	Number of scaffolds	Max scaffold length	Scaffold N50
579,767,826	58,202,653,144	4,973,413	63,379	3,857	4,790,800	84,548	5,576

References

Abiko Y, Nishimura M, Kaku T. 2003. Defensins in saliva and the salivary glands. *Med Electron Microsc* **36**:247-252.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252**:1651-1656.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**:2185-2195.

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* **8**:135-141.

Aird SD. 2005. Taxonomic distribution and quantitative analysis of free purine and pyrimidine nucleosides in snake venoms. *Comp Biochem Physiol B Biochem Mol Biol* **140**:109-126.

Aird SD. 2008. Snake venom dipeptidyl peptidase IV: Taxonomic distribution and quantitative variation. *Comp Biochem Physiol B Biochem Mol Biol* **150**:222-228.

Aird D, Ross MG, Chen W, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. *Genome Biol* **12**:R18.

Aird SD, Watanabe Y, Villar-Briones A, Roy MC, Terada K, Mikheyev AS. 2013. Quantitative high-throughput profiling of snake venom gland transcriptomes and proteomes (*Ovophis okinavensis* and *Protobothrops flavoviridis*). *BMC Genomics* **14**:790.

Alape-Girón A, Sanz L, Escolano J, Flores-Díaz M, Madrigal M, Sasa M, Calvete JJ. 2008. Snake venomics of the lancehead pitviper *Bothrops asper*: Geographic, individual, and ontogenetic variations. *J Proteome Res* **7**:3556-3571.

Alföldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, Russell P, Lowe CB, Glor RE, Jaffe JD. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**:587-591.

Alper C, Balavitch D. 1976. Cobra venom factor: Evidence for its being altered cobra C3 (the third component of complement). *Science* **191**:1275-1276.

Andrade DV, Abe AS. 1999. Relationship of venom ontogeny and diet in *Bothrops*. *Herpetologica* **55**:200-204.

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

Angeletti RH. 1970. Nerve growth factor from cobra venom. *Proc Natl Acad Sci U S A* **65**:668-674.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**:1301-1310.

Bairoch A, Boeckmann B. 1991. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* **19**:2247-2249.

Balsinde J, Winstead MV, Dennis EA. 2002. Phospholipase A_2 regulation of arachidonic acid mobilization. *FEBS Lett* **531**:2-6.

Barlow A, Pook CE, Harrison RA, Wüster W. 2009. Coevolution of diet and prey-specific venom activity supports the role of selection in snake venom evolution. *Proc Roy Soc B Biol Sci* **276**:2443-2449.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: Evolution of new genes under continuous selection. *Proc Natl Acad Sci U S A* **104**:17004-17009.

Bernheimer A, Linder R, Weinstein S, Kim K. 1987. Isolation and characterization of a phospholipase B from venom of Collett's snake, *Pseudechis colletti*. *Toxicon* **25**:547-554.

Biardi JE, Coss RG, Smith DG. 2000. California ground squirrel (*Spermophilus beecheyi*) blood sera inhibits crotalid venom proteolytic activity. *Toxicon* **38**:713-721.

Biardi JE, Chien DC, Coss RG. 2006. California ground squirrel (*Spermophilus beecheyi*) defenses against rattlesnake venom digestive and hemostatic toxins. *J Chem Ecol* **32**:137-154.

Biardi JE, Coss RG. 2011. Rock squirrel (*Spermophilus variegatus*) blood sera affects proteolytic and hemolytic activities of rattlesnake venoms. *Toxicon* **57**:323-331.

Bjarnason JB, Fox JW. 1994. Hemorrhagic metalloproteinases from snake venoms. *Pharmacol Ther* **62**:325-372.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**:578-579.

Boldrini-França J, Rodrigues RS, Fonseca FP, Menaldo DL, Ferreira FB, Henrique-Silva F, Soares AM, Hamaguchi A, Rodrigues VM, Otaviano AR. 2009. *Crotalus durissus collilineatus* venom gland transcriptome: Analysis of gene expression profile. *Biochimie* **91**:586-595.

Bolnick DI, Fitzpatrick BM. 2007. Sympatric speciation: Models and empirical evidence. *Annu Rev Ecol Evol Syst* **38**:459-487.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R. 2013. Assemblathon 2: Evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* **2**:1-31.

Bridges CB. 1936. The bar "gene" a duplication. *Science* **83**:210-211.

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: A theory. *Science* **165**:349-357.

Brust A, Sunagar K, Undheim EA, Vetter I, Yang DC, Casewell NR, Jackson TN, Koludarov I, Alewood PF, Hodgson WC et al. 2013. Differential evolution and neofunctionalization of snake venom metalloprotease domains. *Mol Cell Proteomics* **12**:651-663.

Brykczynska U, Tzika AC, Rodriguez I, Milinkovitch MC. 2013. Contrasted evolution of the vomeronasal receptor repertoires in mammals and squamate reptiles. *Genome Biol Evol* **5**:389-401.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: *De novo* assembly of whole-genome shotgun microreads. *Genome Res* **18**:810-820.

Butte AJ, Dzau VJ, Glueck SB. 2002. Further defining housekeeping, or "maintenance," genes: Focus on "A compendium of gene expression in normal human tissues". *Physiol Genomics* **7**:95-95.

Card DC, Schield DR, Reyes-Velasco J, Adams RH, Mackessy SP, Castoe TA. 2014. The genome of the Prairie Rattlesnake (*Crotalus viridis viridis*). In: Conference Abstract from *Biology of the Pitvipers Symposium 2*, 4the7th June 2014. Tulsa, OK, USA.

Calvete J, Moreno-Murciano M, Theakston R, Kisiel D, Marcinkiewicz C. 2003. Snake venom disintegrins: Novel dimeric disintegrins and structural diversification by disulphide bond engineering. *Biochem J* **372**:725-734.

Calvete JJ, Marcinkiewicz C, Sanz L. 2006. Snake venomics of *Bitis gabonica gabonica*. Protein family composition, subunit organization of venom toxins, and characterization of dimeric disintegrins bitisgabonin-1 and bitisgabonin-2. *J Proteome Res* **6**:326-336.

Calvete JJ, Juárez P, Sanz L. 2007a. Snake venomics: strategy and applications. *J Mass Spectrom* **42**:1405-1414.

Calvete JJ, Escolano J, Sanz L. 2007b. Snake venomics of bitis species reveals large intragenus venom toxin composition variation: Application to taxonomy of congeneric taxa. *J Proteome Res* **6**:2732-2745.

Calvete JJ, Fasoli E, Sanz L, Boschetti E, Righetti PG. 2009. Exploring the venom proteome of the western diamondback rattlesnake, *Crotalus atrox*, via snake venomics and combinatorial peptide ligand library approaches. *J Proteome Res* **8**:3055-3067.

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

Casewell NR, Harrison RA, Wuster W, Wagstaff SC. 2009. Comparative venom gland transcriptome surveys of the saw-scaled vipers (viperidae: *Echis*) reveal substantial intra-family gene diversity and novel venom transcripts. *BMC Genomics* **10**:564-2164-10-564.

Casewell NR. 2012. On the ancestral recruitment of metalloproteinases into the venom of snakes. *Toxicon* **60**:449-454.

Casewell NR, Huttley GA, Wüster W. 2012. Dynamic evolution of venom proteins in squamate reptiles. *Nat 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.

Casewell NR, Wagstaff SC, Wuster W, Cook DA, Bolton FM, King SI, Pla D, Sanz L, Calvete JJ, Harrison RA. 2014. Medically important differences in snake venom composition are dictated by distinct postgenomic mechanisms. *Proc Natl Acad Sci U S A* **111**:9205-9210.

Castoe TA, Jiang ZJ, Gu W, Wang ZO, Pollock DD. 2008. Adaptive evolution and functional redesign of core metabolic proteins in snakes. *PLoS One* **3**:e2201.

Castoe TA, Bronikowski AM, Brodie ED,3rd, Edwards SV, Pfrender ME, Shapiro MD, Pollock DD, Warren WC. 2011a. A proposal to sequence the genome of a garter snake (*Thamnophis sirtalis*). *Stand Genomic Sci* **4**:257-270.

Castoe TA, Hall KT, Guibotsy Mboulas ML, Gu W, de Koning AP, Fox SE, Poole AW, Vemulapalli V, Daza JM, Mockler T et al. 2011b. Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. *Genome Biol Evol* **3**:641-653.

Castoe TA, Fox SE, Jason de Koning A, 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 AP, 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.

Chandramohan R, Wu P, Phan JH, Wang MD. 2013. Benchmarking RNA-seq quantification tools. Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE. IEEE, 2013. Pp. 647-650.

Chang L, Lin S, Chung C. 2004. Molecular cloning and evolution of the genes encoding the precursors of Taiwan cobra cardiotoxin and cardiotoxin-like basic protein. *Biochem Genet* **42**:429-440.

Chatrath ST, Chapeaurouge A, Lin Q, Lim TK, Dunstan N, Mirtschin P, Kumar PP, Kini RM. 2011. Identification of novel proteins from the venom of a cryptic snake *Drysdalia coronoides* by a combined transcriptomics and proteomics approach. *J Proteome Res* **10**:739-750.

Chen Y, Liu T, Yu C, Chiang T, Hwang C. 2013. Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PloS One* **8**:e62856.

Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**:520-562.

Chippaux J, Williams V, White J. 1991. Snake venom variability: Methods of study, results and interpretation. *Toxicon* **29**:1279-1303.

Chu Y, Chang L. 2002. The organization of the genes encoding the A chains of β-bungarotoxins: Evidence for the skipping of exon. *Toxicon* **40**:1437-1443.

Chu C, Tsai T, Tsai I, Lin Y, Tu M. 2009. Prey envenomation does not improve digestive performance in Taiwanese pit vipers (*Trimeresurus gracilis* and *T. stejnegeri stejnegeri*). *Comp Biochem Physiol A Mol Integr Physiol* **152**:579-585.

Ciscotto P, Machado de Avila R, Coelho E, Oliveira J, Diniz C, Farías L, De Carvalho M, Maria W, Sanchez E, Borges A. 2009. Antigenic, microbicidal and antiparasitic properties of an l-amino acid oxidase isolated from *Bothrops jararaca* snake venom. *Toxicon* **53**:330-341.

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the solexa/illumina FASTQ variants. *Nucleic Acids Res* **38**:1767-1771.

Cole S, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon S, Eiglmeier K, Gas S, Barry C3. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**:537-544.

Compeau PE, Pevzner PA, Tesler G. 2011. How to apply de bruijn graphs to genome assembly. *Nat Biotechnol* **29**:987-991.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**:3674-3676.

Cook DA, Samarasekara CL, Wagstaff SC, Kinne J, Wernery U, Harrison RA. 2010. Analysis of camelid IgG for antivenom development: Immunoreactivity and preclinical neutralisation of venom-induced pathology by IgG subclasses, and the effect of heat treatment. *Toxicon* **56**:596-603.

Core LJ, Lis JT. 2008. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* **319**:1791-1792.

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

Cousin X, Créminon C, Grassi J, Méflah K, Cornu G, Saliou B, Bon S, Massoulié J, Bon C. 1996a. Acetylcholinesterase from *Bungarus* venom: A monomeric species. *FEBS Lett* **387**:196-200.

Cousin X, Bon S, Massoulie J, Bon C. 1998. Identification of a novel type of alternatively spliced exon from the acetylcholinesterase gene of *Bungarus fasciatus*. Molecular forms of acetylcholinesterase in the snake liver and muscle. *J Biol Chem* **273**:9812-9820.

Cousin X, Bon S, Duval N, Massoulie J, Bon C. 1996b. Cloning and expression of acetylcholinesterase from *Bungarus fasciatus* venom. A new type of COOH-terminal domain;

involvement of a positively charged residue in the peripheral site. *J Biol Chem* **271**:15099-15108.

Curran T, Franza Jr BR. 1988. Fos and jun: The AP-1 connection. *Cell* **55**:395-397.

Currier RB, Calvete JJ, Sanz L, Harrison RA, Rowley PD, Wagstaff SC. 2012. Unusual stability of messenger RNA in snake venom reveals gene expression dynamics of venom replenishment. *PloS One* **7**:e41888.

Daltry JC, Wüster W, Thorpe RS. 1998. Intraspecific variation in the feeding ecology of the crotaline snake *Calloselasma rhodostoma* in Southeast Asia. *J Herpetol* **32**:198-205.

Daltry JC, Wuster W, Thorpe RS. 1996. Diet and snake venom evolution. *Nature* **379**:537-540.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics* **27**:1164-1165.

Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. 2013. An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PloS One* **8**:e85024.

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.

Desmet W. 1981. The nuclear feulgen-DNA content of the vertebrates (especially reptiles), as measured by fluorescence cytophotometry, with notes on the cell and the chromosome size. *Acta Zool Pathol Antverp* **76**:119-167.

Dickinson DP. 2002. Salivary (SD-type) cystatins: Over one billion years in the making--but to what purpose? *Crit Rev Oral Biol Med* **13**:485-508.

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.

Di-Poï N, Montoya-Burgos JI, Miller H, Pourquié O, Milinkovitch MC, Duboule D. 2010. Changes in Hox genes' structure and function during the evolution of the squamate body plan. *Nature* **464**:99-103.

Dollo L. 1893. The laws of evolution. *Bull Soc Bel Geol Paleontol* **7**:164-166.

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES. 1987. A genetic linkage map of the human genome. *Cell* **51**:319-337.

Du X, Clemetson KJ. 2002. Snake venom L-amino acid oxidases. *Toxicon* **40**:659-665.

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.

Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M et al. 2011. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. *Genome Res* **21**:2224-2241.

Eckalbar WL, Hutchins ED, Markov GJ, Allen AN, Corneveaux JJ, Lindblad-Toh K, Di Palma F, Alfoldi J, Huentelman MJ, Kusumi K. 2013. Genome reannotation of the lizard *Anolis carolinensis* based on 14 adult and embryonic deep transcriptomes. *BMC Genomics* **14**:49.

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H, Nadachowska-Brzyska K, Qvarnström A. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**:756-760.

Eppig JT. 2006. Mouse strain and genetic nomenclature: An abbreviated guide. In: Fox, JG et al. (Eds). *The Mouse in Biomedical Research, Vol 1*. Pp. 79-98.

Escoriza D, Metallinou M, Donaire D, Amat F, Carranza S. 2009. Biogeography of the white-bellied carpet viper *Echis leucogaster roman*, 1972 in morocco, a study combining mitochondrial DNA data and ecological niche modeling. *Buttl Soc Cat Herp* **18**: 55-69.

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.

Fabrizio MC. 2001. Use of ecarin clotting time (ECT) with lepirudin therapy in heparin-induced thrombocytopenia and cardiopulmonary bypass. *J Extra Corpor Technol* **33**:117-125.

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.

Feasey N, Wansbrough-Jones M, Mabey DC, Solomon AW. 2010. Neglected tropical diseases. *Br Med Bull* **93**:179-200.

Ferreira S. 1965. A Bradykinin-potentiating factor (BPF) present in the venom of *Bothrops jararaca*. *Brit J Pharmacol* **24**:163-169.

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.

Francischetti I, My-Pham V, Harrison J, Garfield MK, Ribeiro J. 2004. *Bitis gabonica* (gaboon viper) snake venom gland: Toward a catalog for the full-length transcripts (cDNA) and proteins. *Gene* **337**:55-69.

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, Wickramaratna JC, Jones A, Alewood PF, Hodgson WC. 2001. Species and regional variations in the effectiveness of antivenom against the *in vitro* neurotoxicity of death adder (*Acanthophis*) venoms. *Toxicol Appl Pharmacol* **175**:140-148.

Fry BG, Wüster W, Kini RM, Brusic V, Khan A, Venkataraman D, Rooney A. 2003a. Molecular evolution and phylogeny of elapid snake venom three-finger toxins. *J Mol Evol* **57**:110-129.

Fry BG, Winkel KD, Wickramaratna JC, Hodgson WC, Wüster W. 2003b. Effectiveness of snake antivenom: Species and regional venom variation and its clinical impact. *Toxin Rev* **22**:23-34.

Fry BG, Lumsden NG, Wüster W, Wickramaratna JC, Hodgson WC, Kini RM. 2003c. Isolation of a neurotoxin (α-colubritoxin) from a nonvenomous colubrid: Evidence for early origin of venom in snakes. *J Mol Evol* **57**:446-452.

Fry BG, Wüster W, Ramjan R, Fadil S, Jackson T, Martelli P, Kini RM. 2003d. Analysis of colubroidea snake venoms by liquid chromatography with mass spectrometry: Evolutionary and toxinological implications. *Rapid Commun Mass Spectrom* **17**:2047-2062.

Fry BG, Wuster 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 Human Genet* **10**:483-511.

Fry BG, Wroe S, Teeuwisse W, van Osch MJ, Moreno K, Ingle J, McHenry C, Ferrara T, Clausen P, Scheib H et al. 2009c. A central role for venom in predation by *Varanus komodoensis* (komodo dragon) and the extinct giant *Varanus* (*Megalania*) *priscus*. *Proc Natl Acad Sci U S A* **106**:8969-8974.

Fry BG, Winter K, Norman JA, Roelants K, Nabuurs RJ, van Osch MJ, Teeuwisse WM, van der Weerd L, McNaughtan JE, Kwok HF et al. 2010a. Functional and structural diversification of the Anguimorpha lizard venom system. *Mol Cell Proteomics* **9**:2369-2390.

Fry BG, Roelants K, Winter K, Hodgson WC, Griesman L, Kwok HF, Scanlon D, Karas J, Shaw C, Wong L et al. 2010b. Novel venom proteins produced by differential domain-expression strategies in beaded lizards and gila monsters (genus *Heloderma*). *Mol Biol Evol* **27**:395-407.

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.

Fujita MK, Edwards SV, Ponting CP. 2011. The *Anolis* lizard genome: An amniote genome without isochores. *Genome Biol Evol* **3**:974-984.

Furtado MF, Maruyama M, Kamiguti A, Antonio L. 1991. Comparative study of nine *Bothrops* snake venoms from adult female snakes and their offspring. *Toxicon* **29**:219-226.

Ganapathy G, Howard JT, Ward JM, Li J, Li B, Li Y, Xiong Y, Zhang Y, Zhou S, Schwartz DC. 2014. High-coverage sequencing and annotated assemblies of the budgerigar genome. *GigaScience* **3**:1-9.

Georgieva D, Risch M, Kardas A, Buck F, von Bergen M, Betzel C. 2008. Comparative analysis of the venom proteomes of *Vipera ammodytes ammodytes* and *Vipera ammodytes meridionalis*. *J Proteome Res* **7**:866-886.

Gibbs HL, Rossiter W. 2008. Rapid evolution by positive selection and gene gain and loss: PLA_2 venom genes in closely related *Sistrurus* rattlesnakes with divergent diets. *J Mol Evol* **66**:151-166.

Gibbs HL, Mackessy SP. 2009. Functional basis of a molecular adaptation: Prey-specific toxic effects of venom from *Sistrurus* rattlesnakes. *Toxicon* **53**:672-679.

Gibbs HL, Sanz L, Chiucchi JE, Farrell TM, Calvete JJ. 2011. Proteomic analysis of ontogenetic and diet-related changes in venom composition of juvenile and adult dusky pigmy rattlesnakes (*Sistrurus miliarius barbouri*). *J Proteomics* **74**:2169-2179.

Glare EM, Divjak M, Bailey MJ, Walters EH. 2002. Beta-actin and GAPDH housekeeping gene expression in asthmatic airways is variable and not suitable for normalising mRNA levels. *Thorax* **57**:765-770.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108**:1513-1518.

Gomez C, Özbudak EM, Wunderlich J, Baumann D, Lewis J, Pourquié O. 2008. Control of segment number in vertebrate embryos. *Nature* **454**:335-339.

Gomez C, Pourquié O. 2009. Developmental control of segment numbers in vertebrates. *J Exp Zool B Mol Dev Evol* **312**:533-544.

Gould SJ, Vrba ES. 1982. Exaptation-a missing term in the science of form. *Paleobiology* **8**:4-15.

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.

Gregory TR. 2005. Genome size evolution in animals. In: Gregory, TR (Eds). *The Evolution of the Genome.* Elsevier Academic press, Oxford. pp. 4-87.

Gregory S, Barlow K, McLay K, Kaul R, Swarbreck D, Dunham A, Scott C, Howe K, Woodfine K, Spencer C. 2006. The DNA sequence and biological annotation of human chromosome 1. *Nature* **441**:315-321.

Guércio RA, Shevchenko A, Shevchenko A, López-Lozano JL, Paba J, Sousa MV, Ricart CA. 2006. Ontogenetic variations in the venom proteome of the Amazonian snake *Bothrops atrox*. *Proteome Sci* **4**:11.

Guhad F. 2005. Introduction to the 3Rs (refinement, reduction and replacement). *J Am Assoc Lab Anim Sci* **44**:58-59.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**:1072-1075.

Gutiérrez JM, Sanz L, Flores-Díaz M, Figueroa L, Madrigal M, Herrera M, Villalta M, León G, Estrada R, Borges A. 2009. Impact of regional variation in *Bothrops asper* snake venom on the design of antivenoms: Integrating antivenomics and neutralization approaches. *J Proteome Res* **9**:564-577.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat Protocols* **8**:1494-1512.

Hall TA. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. In *Nucleic acids symposium series* **41**:95-98.

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

Hansen KD, Brenner SE, Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**:e131.

Hargreaves AD, Swain MT, Logan DW, Mulley JF. 2014a. Testing the toxicofera: Comparative reptile transcriptomics casts doubt on the single, early evolution of the reptile venom system. *Toxicon* **92**:140-156.

Hargreaves AD, Swain MT, Hegarty MJ, Logan DW, Mulley JF. 2014b. Restriction and recruitment – gene duplication and the origin and evolution of snake venom toxins. *Genome Biol Evol* **6**:2088-2095.

Hargreaves AD, Mulley JF. 2014. A plea for standardized nomenclature of snake venom toxins. *Toxicon* **90**:351-353.

Harizi H, Corcuff J, Gualde N. 2008. Arachidonic-acid-derived eicosanoids: Roles in biology and immunopathology. *Trends Mol Med* **14**:461-469.

Harrison R. 2004. Development of venom toxin-specific antibodies by DNA immunisation: Rationale and strategies to improve therapy of viper envenoming. *Vaccine* **22**:1648-1655.

Harrison RA, Ibison F, Wilbraham D, Wagstaff SC. 2007. Identification of cDNAs encoding viper venom hyaluronidases: Cross-generic sequence conservation of full-length and unusually short variant transcripts. *Gene* **392**:22-33.

Harrison RA, Hargreaves A, Wagstaff SC, Faragher B, Lalloo DG. 2009. Snake envenoming: A disease of poverty. *PLoS Negl Trop Dis* **3**:e569.

Harvey AL. 2001. Twenty years of dendrotoxins. *Toxicon* **39**:15-26.

Haussler D, O'Brien SJ, Ryder OA, Barker FK, Clamp M, Crawford AJ, Hanner R, Hanotte O, Johnson WE, McGuire JA. 2009. Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered* **100**:659-674.

Hawgood BJ. 1992. Pioneers of anti-venomous serotherapy: Dr Vital Brazil (1865–1950). *Toxicon* **30**:573-579.

Hawgood BJ. 1999. Doctor Albert Calmette 1863–1933: Founder of antivenomous serotherapy and of antituberculous BCG vaccination. *Toxicon* **37**:1241-1258.

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.

Hayashi MA, Oliveira EB, Kerkis I, Karpel RL. 2012. Crotamine: A novel cell-penetrating polypeptide nanocarrier with potential anti-cancer and biotechnological applications. In: *Anonymous Nanoparticles in Biology and Medicine*. Springer. pp 337-352.

Hayden MS, Ghosh S. 2012. NF-kappaB, the first quarter-century: Remarkable progress and outstanding questions. *Genes Dev* **26**:203-234.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**:695-716.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**:129-149.

Ho CYL. 2005. Purification of metalloprotease inhibitors that neutralize snake venom toxins in California ground squirrel blood. *Explorations*.

Hommes DW, Peppelenbosch MP, van Deventer SJ. 2003. Mitogen activated protein (MAP) kinase signal transduction pathways and novel anti-inflammatory targets. *Gut* **52**:144-151.

Hughes MK, Hughes AL. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol Biol Evol* **10**:1360-1369.

Hui Ng H, Bird A. 2000. Histone deacetylases: Silencers for hire. *Trends Biochem Sci* **25**:121-126.

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. 2013. REAPR: A universal tool for genome assembly evaluation. *Genome Biol* **14**:R47.

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.

Izidoro LFM, Ribeiro MC, Souza GR, Sant'Ana CD, Hamaguchi A, Homsi-Brandeburgo MI, Goulart LR, Beleboni RO, Nomizo A, Sampaio SV. 2006. Biochemical and functional characterization of an l-amino acid oxidase isolated from *Bothrops pirajai* snake venom. *Bioorg Med Chem* **14**:7034-7043.

Jeyaseelan K, Armugam A, Donghui M, Tan NH. 2000. Structure and phylogeny of the venom group I phospholipase A_2 gene. *Mol Biol Evol* **17**:1010-1021.

Jeyaseelan K, Ma D, Armugam A. 2001. Real-time detection of gene promoter activity: Quantitation of toxin gene transcription. *Nucleic Acids Res* **29**:E58-8.

Jia L, Shimokawa K, Bjarnason JB, Fox JW. 1996. Snake venom metalloproteinases: Structure, function and relationship to the ADAMs family of proteins. *Toxicon* **34**:1269-1276.

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.

John TR, Smith JJ, Kaiser II. 1996. A phospholipase A_2-like pseudogene retaining the highly conserved introns of Mojave toxin and other snake venom group II PLA_2s, but having different exons. *DNA Cell Biol* **15**:661-668.

Junqueira de Azevedo, ILM, Farsky SHP, Oliveira MLS, Ho PL. 2001. Molecular cloning and expression of a functional snake venom vascular endothelium growth factor (VEGF) from the *Bothrops insularis* pit viper: a new member of the VEGF family of proteins. *J Biol Chem* **276**:39836-39842.

Jurgilas PB, Neves-Ferreira AG, Domont GB, Perales J. 2003. PO41, a snake venom metalloproteinase inhibitor isolated from *Philander* opossum serum. *Toxicon* **42**:621-628.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**:D109-14.

Kardong KV. 1980. Evolutionary patterns in advanced snakes. *Am Zool* **20**:269-282.

Kardong K, Weinstein S, Smith T, Mackessy S. 2009. Reptile venom glands: Form, function, and future. In: Mackessy, SP (Eds). *Handbook of Venoms and Toxins of Reptiles*. Pp 65-91.

Kardong KV. 2012. Replies to Fry et al.(toxicon 2012, 60/4, 434–448). part B. properties and biological roles of squamate oral products: The "venomous lifestyle" and preadaptation. *Toxicon* **60**:964-966.

Karin M, Liu Z, Zandi E. 1997. AP-1 function and regulation. *Curr Opin Cell Biol* **9**:240-246.

Karin M. 1999. How NF-kappaB is activated: The role of the IkappaB kinase (IKK) complex. *Oncogene* **18**:6867-6874.

Kasahara M. 2007. The 2R hypothesis: An update. *Curr Opin Immunol* **19**:547-552.

Kasturiratne A, Wickremasinghe AR, de Silva N, Gunawardena NK, Pathmeswaran A, Premaratna R, Savioli L, Lalloo DG, de Silva HJ. 2008. The global burden of snakebite: A literature analysis and modelling based on regional estimates of envenoming and deaths. *PLoS Medicine* **5**:e218.

Kelley DR, Schatz MC, Salzberg SL. 2010. Quake: Quality-aware detection and correction of sequencing errors. *Genome Biol* **11**:R116.

Kemparaju K, Girish K. 2006. Snake venom hyaluronidase: A therapeutic target. *Cell Biochem Funct* **24**:7-12.

Kerchove CM, Carneiro SM, Markus RP, Yamanouye N. 2004. Stimulation of the alpha-adrenoceptor triggers the venom production cycle in the venom gland of *Bothrops jararaca*. *J Exp Biol* **207**:411-416.

Kerchove CM, Luna MS, Zablith MB, Lazari MF, Smaili SS, Yamanouye N. 2008. α_1-adrenoceptors trigger the snake venom production cycle in secretory cells by activating phosphatidylinositol 4, 5-bisphosphate hydrolysis and ERK signaling pathway. *Comp Biochem Physiol A Mol Integr Physiol* **150**:431-437.

Khan MI, Kamal MS. 2014. De bruijn graph based *de novo* genome assembly. *J Softw* **9**:2160-2168.

Kilmon Sr J. 1976. High tolerance to snake venom by the Virginia opossum, *Didelphis virginiana*. *Toxicon* **14**:337-340.

King GF, Gentz MC, Escoubas P, Nicholson GM. 2008. A rational nomenclature for naming peptide toxins from spiders and other venomous animals. *Toxicon* **52**:264-276.

Kini RM, Evans HJ. 1992. Structural domains in venom proteins: Evidence that metalloproteinases and nonenzymatic platelet aggregation inhibitors (disintegrins) from snake venoms are derived by proteolysis from a common precursor. *Toxicon* **30**:265-293.

Kini RM, Chan YM. 1999. Accelerated evolution and molecular surface of venom phospholipase A_2 enzymes. *J Mol Evol* **48**:125-132.

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.

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

Kochva E. 1978. Oral glands of the reptilia. In: *Biology of the Reptilia*. Academic press, New York. Pp. 43-161.

Kochva E, Nakar O, Ovadia M. 1983. Venom toxins: Plausible evolution from digestive enzymes. *Am Zool* **23**:427-430.

Kochva E. 1987. The origin of snakes and evolution of the venom apparatus. *Toxicon* **25**:65-106.

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.

Koludarov I, Sunagar K, Undheim EA, Jackson TN, Ruder T, Whitehead D, Saucedo AC, Mora GR, Alagon AC, King G. 2012. Structural and molecular diversification of the Anguimorpha lizard mandibular venom gland system in the arboreal species *Abronia graminea*. *J Mol Evol* **75**:168-183.

Komori Y, Nikai T, Sugihara H. 1988. Biochemical and physiological studies on a kallikrein-like enzyme from the venom of *Crotalus viridis viridis* (prairie rattlesnake). *Biochim Biophys Acta* **967**:92-102.

Komori Y, Nikai T. 1998. Chemistry and biochemistry of kallikrein-like enzyn from snake venoms. *Toxin Rev* **17**:261-277.

Kordiš D, Gubenšek F. 1997. Bov-B long interspersed repeated DNA (LINE) sequences are present in *Vipera ammodytes* phospholipase A_2 genes and in genomes of viperidae snakes. *Eur J Biochem* **246**:772-779.

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

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* **30**:693-700.

Kostiza T, Meier J. 1996. Nerve growth factors from snake venoms: Chemical properties, mode of action and biological significance. *Toxicon* **34**:787-806.

Kouzarides T. 2007. Chromatin modifications and their function. *Cell* **128**:693-705.

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (G C)-biased genomes. *Nat Methods* **6**:291-295.

Krumm A, Hickey LB, Groudine M. 1995. Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev* **9**:559-572.

Kruse U, Qian F, Sippel AE. 1991. Identification of a fourth nuclear factor I gene in chicken by cDNA cloning: NFI-X. *Nucleic Acids Res* **19**:6641.

Kuhn TS. 1962. The structure of scientific revolutions. TheUniversity of Chicago press, Chicago.

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.

Kuraku S, Meyer A. 2009. The evolution and maintenance of hox gene clusters in vertebrates and the teleost-specific genome duplication. *Int J Dev Biol* **53**:765.

Kusumi K, Kulathinal RJ, Abzhanov A, Boissinot S, Crawford NG, Faircloth BC, Glenn TC, Janes DE, Losos JB, Menke DB et al. 2011. Developing a community-based genetic nomenclature for anole lizards. *BMC Genomics* **12**:554.

Kuwada Y. 1911. Maiosis in the pollen mother cells of zea mays L.(with plate V.). 植物学雑誌 **25**:163-181.

Kwong PD, McDonald NQ, Sigler PB, Hendrickson WA. 1995. Structure of β2-bungarotoxin: Potassium channel binding by kunitz modules and targeted phospholipase action. *Structure* **3**:1109-1119.

Kwong S, Kini RM. Duplication of coagulation factor genes and evolution of snake venom prothrombin activators. In: Friedberg F, (Eds). *Gene duplication*. InTech. p. 257–278.

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**:469–478.

Lachumanan R, Armugam A, Tan C, Jeyaseelan K. 1998. Structure and organization of the cardiotoxin genes in *Naja naja sputatrix*. *FEBS Lett* **433**:119-124.

Lachumanan R, Armugam A, Durairaj P, Gopalakrishnakone P, Tan CH, Jeyaseelan K. 1999. *In situ* hybridization and immunohistochemical analysis of the expression of cardiotoxin and neurotoxin genes in *Naja naja sputatrix*. *J Histochem Cytochem* **47**:551-560.

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

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2**:231-239.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860-921.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**:R25.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. *Nat Methods* **9**:357-359.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al. 2007. Clustal W and clustal X version 2.0. *Bioinformatics* **23**:2947-2948.

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.

Lederberg J, Mccray A. 2001. The scientist:\'Ome sweet\'omics--A genealogical treasury of words. *The Scientist* **17**.

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol Biol Evol* **4**:203-221.

Li M, Fry B, Kini RM. 2005a. Eggs-only diet: Its implications for the toxin profile changes and ecology of the marbled sea snake (*Aipysurus eydouxii*). *J Mol Evol* **60**:81-89.

Li M, Fry BG, Kini RM. 2005b. Putting the brakes on snake venom evolution: The unique molecular evolutionary patterns of *Aipysurus eydouxii* (marbled sea snake) phospholipase A_2 toxins. *Mol Biol Evol* **22**:934-941.

Li W, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends Genet* **21**:602-607.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1,000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**:2078-2079.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y. 2009. The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**:311-317.

Li B, Dewey CN. 2011. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**:323-2105-12-323.

Liang G, Lin JC, Wei V, Yoo C, Cheng JC, Nguyen CT, Weisenberger DJ, Egger G, Takai D, Gonzales FA et al. 2004. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc Natl Acad Sci U S A* **101**:7357-7362.

Liang D, Wu R, Geng J, Wang C, Zhang P. 2011. A general scenario of Hox gene inventory variation among major Sarcopterygian lineages. *BMC Evol Biol* **11**:25.

Ligabue-Braun R, Verli H, Carlini CR. 2012. Venomous mammals: A review. *Toxicon* **59**:680-695.

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.

Lomonte B, Escolano J, Fernández J, Sanz L, Angulo Y, Gutiérrez JM, Calvete JJ. 2008. Snake venomics and antivenomics of the arboreal neotropical pitvipers *Bothriechis lateralis* and *Bothriechis schlegelii*. *J Proteome Res* **7**:2445-2457.

Long M. 2001. Evolution of novel genes. *Curr Opin Genet Dev* **11**:673-680.

López-Lozano JL, de Sousa MV, Ricart CAO, Chávez-Olortegui C, Flores Sanchez E, Muniz EG, Bührnheim PF, Morhy L. 2002. Ontogenetic variation of metalloproteinases and plasma coagulant activity in venoms of wild *Bothrops atrox* specimens from Amazonian rain forest. *Toxicon* **40**:997-1006.

Low DH, Sunagar K, Undheim EA, Ali SA, Alagon AC, Ruder T, Jackson TN, Pineda Gonzalez S, King GF, Jones A. 2013. Dracula's children: Molecular evolution of vampire bat venom. *J Proteomics* **89**:95-111.

Lu Q, Navdaev A, Clemetson JM, Clemetson KJ. 2005. Snake venom C-type lectins interacting with platelet receptors. structure–function relationships and effects on haemostasis. *Toxicon* **45**:1089-1098.

Lumsden NG, Fry BG, Ventura S, Kini RM, Hodgson WC. 2005. Pharmacological characterisation of a neurotoxin from the venom of *Boiga dendrophila* (mangrove catsnake). *Toxicon* **45**:329-334.

Luna MS, Hortencio TM, Ferreira ZS, Yamanouye N. 2009. Sympathetic outflow activates the venom gland of the snake *Bothrops jararaca* by regulating the activation of transcription factors and the synthesis of venom gland proteins. *J Exp Biol* **212**:1535-1543.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y. 2012. SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**:18.

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.

Ma D, Armugam A, Jeyaseelan K. 2001. Expression of cardiotoxin-2 gene. *Eur J Biochem* **268**:1844-1850.

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

Mackessy SP. 1988. Venom ontogeny in the pacific rattlesnakes *Crotalus viridis helleri* and *C. v. oreganus*. *Copeia* **1988**:92-101.

Mackessy SP. 2002. Biochemistry and pharmacology of colubrid snake venoms. *Toxin Rev* **21**:43-83.

Mackessy SP, Williams K, Ashton KG. 2003. Ontogenetic variation in venom composition and diet of *Crotalus oreganus concolor*: A case of venom paedomorphosis? *Copeia* **2003**:769-782.

Mansfield JH. 2013. Cis-regulatory change associated with snake body plan evolution. *Proc Natl Acad Sci U S A* **110**:10473-10474.

Mardis ER. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**:387-402.

Margaritis T, Holstege FC. 2008. Poised RNA polymerase II gives pause for thought. *Cell* **133**:581-584.

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.

Marguerat S, Bähler J. 2010. RNA-seq: From technology to biology. *Cell Mol Life Sci* **67**:569-579.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y, Chen Z. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376-380.

Markland Jr FS, Swenson S. 2013. Snake venom metalloproteinases. *Toxicon* **62**:3-18.

Marshall CR, Raff EC, Raff RA. 1994. Dollo's law and the death and resurrection of genes. *Proc Natl Acad Sci U S A* **91**:12283-12287.

Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. *Nat Rev Genet* **12**:671-682.

Mathews M, Jia HP, Guthmiller JM, Losh G, Graham S, Johnson GK, Tack BF, McCray PB,Jr. 1999. Production of beta-defensin antimicrobial peptides by the oral mucosa and salivary glands. *Infect Immun* **67**:2740-2745.

McCue MD. 2005. Enzyme activities and biological functions of snake venoms. *Appl Herpetol* **2**:109-123.

McCue MD. 2006. Cost of producing venom in three North American pitviper species. *Copeia* **2006**:818-825.

McCue MD. 2007. Prey envenomation does not improve digestive performance in Western diamondback rattlesnakes (*Crotalus atrox*). *J Exp Zool A Ecol Genet Physiol* **307**:568-577.

McGettigan PA. 2013. Transcriptomics in the RNA-seq era. *Curr Opin Chem Biol* **17**:4-11.

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK. 2001. A physical map of the human genome. *Nature* **409**:934-941.

Meisler MH, Ting CN. 1993. The remarkable evolutionary history of the human amylase genes. *Crit Rev Oral Biol Med* **4**:503-509.

Melvin JE, Yule D, Shuttleworth T, Begenisich T. 2005. Regulation of fluid and electrolyte secretion in salivary gland acinar cells. *Annu Rev Physiol* **67**:445-469.

Menchaca JM, Perez JC. 1981. The purification and characterization of an antihemorrhagic factor in opossum (*Didelphis virginiana*) serum. *Toxicon* **19**:623-632.

Menezes MC, Furtado MF, Travaglia-Cardoso SR, Camargo A, Serrano SM. 2006. Sex-based individual variation of snake venom proteome among eighteen *Bothrops jararaca* siblings. *Toxicon* **47**:304-312.

Meyer A, Van de Peer Y. 2005. From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD). *Bioessays* **27**:937-945.

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

Mikheyev AS, Tin MM. 2014. A first look at the oxford nanopore MinION sequencer. *Mol Ecol Resour*, In press.

Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**:315-327.

Morita T. 2005. Structures and functions of snake venom CLPs (C-type lectin-like proteins) with anticoagulant-, procoagulant-, and platelet-modulating activities. *Toxicon* **45**:1099-1114.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35**:W182-5.

Mullins M. 1995. Genetic methods: Conventions for naming zebrafish genes. In: *The Zebrafish Book*. University of Oregon Press, Eugene, Oregon. Pp 7.1-7.4.

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.

Nagaraj SH, Gasser RB, Ranganathan S. 2007. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform* **8**:6-21.

Nair D, Fry B, Alewood P, Kumar P, Kini R. 2007. Antimicrobial activity of omwaprin, a new member of the waprin family of snake venom proteins. *Biochem J* **402**:93-104.

Napetschnig J, Wu H. 2013. Molecular basis of NF-kappaB signaling. *Annu Rev Biophys* **42**:443-468.

Nei M, Gu X, Sitnikova T. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci U S A* **94**:7799-7806.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39**:121-152.

Neiva M, Arraes F, de Souza JV, Rádis-Baptista G, Prieto da Silva, Álvaro RB, Walter MEM, Brigido MdM, Yamane T, López-Lozano JL, Astolfi-Filho S. 2009. Transcriptome analysis of the Amazonian viper *Bothrops atrox* venom gland using expressed sequence tags (ESTs). *Toxicon* **53**:427-436.

Nekaris KA, Moore RS, Rode EJ, Fry BG. 2013. Mad, bad and dangerous to know: The biochemistry, ecology and evolution of slow loris venom. *J Venom Anim Toxins Incl Trop Dis* **19**:21.

Neves-Ferreira AG, Cardinale N, Rocha SL, Perales J, Domont GB. 2000. Isolation and characterization of DM40 and DM43, two snake venom metalloproteinase inhibitors from *Didelphis marsupialis* serum. *Biochim Biophys Acta* **1474**:309-320.

Obokata H, Wakayama T, Sasai Y, Kojima K, Vacanti MP, Niwa H, Yamato M, Vacanti CA. 2014a. Retraction: Stimulus-triggered fate conversion of somatic cells into pluripotency. *Nature* **511**:112-112.

Obokata H, Sasai Y, Niwa H, Kadota M, Andrabi M, Takata N, Tokoro M, Terashita Y, Yonemura S, Vacanti CA. 2014b. Retraction: Bidirectional developmental potential in reprogrammed cells with acquired pluripotency. *Nature* **511**:112-112.

Ogawa T, Chijiwa T, Oda-Ueda N, Ohno M. 2005. Molecular diversity and accelerated evolution of C-type lectin-like proteins from snake venom. *Toxicon* **45**:1-14.

Oguiura N, Boni-Mitake M, Rádis-Baptista G. 2005. New view on crotamine, a small basic polypeptide myotoxin from South American rattlesnake venom. *Toxicon* **46**:363-370.

Ohno S. 1967. Sex chromosomes and sex-linked genes. Springer, Berlin.

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

Oliveira JS, Fuentes-Silva D, King GF. 2012. Development of a rational nomenclature for naming peptide and protein toxins from sea anemones. *Toxicon* **60**:539-550.

OmPraba G, Chapeaurouge A, Doley R, Devi KR, Padmanaban P, Venkatraman C, Velmurugan D, Lin Q, Kini RM. 2010. Identification of a novel family of snake venom proteins veficolins from *Cerberus rynchops* using a venom gland transcriptomics and proteomics approach. *J Proteome Res* **9**:1882-1893.

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

Ovcharenko I, Loots GG, Giardine BM, Hou M, Ma J, Hardison RC, Stubbs L, Miller W. 2005. Mulan: Multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res* **15**:184-194.

Pahari S, Mackessy SP, Kini RM. 2007. The venom gland transcriptome of the desert massasauga rattlesnake (*Sistrurus catenatus edwardsii*): Towards an understanding of venom composition among advanced snakes (superfamily colubroidea). *BMC Mol Biol* **8**:115.

Paine M, Desmond H, Theakston R, Crampton J. 1992. Gene expression in *Echis carinatus* (carpet viper) venom glands following milking. *Toxicon* **30**:379-386.

Parkinson J, Blaxter M. 2004. Expressed sequence tags. In: *Anonymous Parasite Genomics Protocols*. Springer. pp. 93-126.

Parra G, Bradnam K, Korf I. 2007. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**:1061-1067.

Pawlak J, Kini RM. 2008. Unique gene organization of colubrid three-finger toxins: Complete cDNA and gene sequences of denmotoxin, a bird-specific toxin from colubrid snake *Boiga dendrophila* (mangrove catsnake). *Biochimie* **90**:868-877.

Peng Y, Leung HC, Yiu SM, Lv MJ, Zhu XG, Chin FY. 2013. IDBA-tran: A more robust *de novo* de bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics* **29**:i326-34.

Perales J, Munoz R, Moussatche H. 1986. Isolation and partial characterization of a protein fraction from the opossum (*Didelphis marsupialis*) serum, with protecting property against the *Bothrops jararaca* snake venom. *An Acad Bras Cienc* **58**:155-162.

Perez JC, Haws WC, Garcia VE, Jennings III BM. 1978. Resistance of warm-blooded animals to snake venoms. *Toxicon* **16**:375-383.

Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat Methods* **8**:785-786.

Pimenta DC, Prezoto BC, Konno K, Melo RL, Furtado MF, Camargo A, Serrano SM. 2007. Mass spectrometric analysis of the individual variability of *Bothrops jararaca* venom peptide fraction. evidence for sex-based variation among the bradykinin-potentiating peptides. *Rapid Commun Mass Spectrom* **21**:1034-1042.

Pintor AF, Krockenberger AK, Seymour JE. 2010. Costs of venom production in the common death adder (*Acanthophis antarcticus*). *Toxicon* **56**:1035-1042.

Pirkle H. 1998. Thrombin-like enzymes from snake venoms: An updated inventory. *Thromb Haemost* **79**:675-683.

Piskurek O, Austin CC, Okada N. 2006. Sauria SINEs: Novel short interspersed retroposable elements that are widespread in reptile genomes. *J Mol Evol* **62**:630-644.

Pogrel MA, Low MA, Stern R. 2003. Hyaluronan (hyaluronic acid) and its regulation in human saliva by hyaluronidase and its inhibitors. *J Oral Sci* **45**:85-92.

Pook CE, Joger U, Stümpel N, Wüster W. 2009. When continents collide: Phylogeny, historical biogeography and systematics of the medically important viper genus *Echis* (squamata: Serpentes: Viperidae). *Mol Phylogenet Evol* **53**:792-807.

Popper K. 1959. The logic of scientific discovery. Routledge, New York.

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

Pough F, Andrews R, Cadle J, Crump M, Savitzky A, Wells K. 2004. Herpetology as a field of study. In: Pough, FH (Eds). *Herpetology*, 3rd Edn pp. 6-24.

Pung YF, Wong PT, Kumar PP, Hodgson WC, Kini RM. 2005. Ohanin, a novel protein from king cobra venom, induces hypolocomotion and hyperalgesia in mice. *J Biol Chem* **280**:13137-13147.

Pung YF, Kumar SV, Rajagopalan N, Fry BG, Kumar PP, Kini RM. 2006. Ohanin, a novel protein from king cobra venom: Its cDNA and genomic organization. *Gene* **371**:246-256.

Putney Jr J. 1986. Identification of cellular activation mechanisms associated with salivary secretion. *Annu Rev Physiol* **48**:75-88.

Pyron RA, Burbrink FT. 2009. Neogene diversification and taxonomic stability in the snake tribe lampropeltini (serpentes: Colubridae). *Mol Phylogenet Evol* **52**:524-529.

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.

Radis-Baptista G, Kerkis I. 2011. Crotamine, a small basic polypeptide myotoxin from rattlesnake venom with cell-penetrating properties. *Curr Pharm Des* **17**:4351-4361.

Radonić A, Thulke S, Mackay IM, Landt O, Siegert W, Nitsche A. 2004. Guideline to reference gene selection for quantitative real-time PCR. *Biochem Biophys Res Commun* **313**:856-862.

Rao VS, Swarup S, Kini RM. 2004. The catalytic subunit of pseutarin C, a group C prothrombin activator from the venom of *Pseudonaja textilis*, is structurally similar to mammalian blood coagulation factor Xa. *Thromb Haemost* **92**:509-521.

Rehana S, Manjunatha Kini R. 2007. Molecular isoforms of cobra venom factor-like proteins in the venom of *Austrelaps superbus*. *Toxicon* **50**:32-52.

Rehana S, Kini RM. 2008. Complement C3 isoforms in *Austrelaps superbus*. *Toxicon* **51**:864-881.

Ren L, Qingrun Z, Yun Z. 1999. Snake venom serine proteases. *Exploration of nature* **3**:003.

Reza MA, Minh Le T, Swarup S, Kini R. 2006. Molecular evolution caught in action: Gene duplication and evolution of molecular isoforms of prothrombin activators in *Pseudonaja textilis* (brown snake). *J Thromb Haemost* **4**:1346-1353.

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.

Richards R, St Pierre L, Trabi M, Johnson LA, de Jersey J, Masci PP, Lavin MF. 2011. Cloning and characterisation of novel cystatins from elapid snake venom glands. *Biochimie* **93**:659-668.

Richards D, Barlow A, Wüster W. 2012. Venom lethality and diet: Differential responses of natural prey and model organisms to the venom of the saw-scaled vipers (*Echis*). *Toxicon* **59**:110-116.

Ritonja A, Evans HJ, Machleidt W, Barrett AJ. 1987. Amino acid sequence of a cystatin from venom of the African puff adder (*Bitis arietans*). *Biochem J* **246**:799-802.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ. 2010. *De novo* assembly and analysis of RNA-seq data. *Nat Methods* **7**:909-912.

Roemer RC, Borchardt R. 2012. From bibliometrics to altmetrics: A changing scholarly landscape. *College & Research Libraries News* **73**:596-600.

Rokyta DR, Wray KP, Lemmon AR, Lemmon EM, Caudle SB. 2011. A high-throughput venom-gland transcriptome for the Eastern diamondback rattlesnake (*Crotalus adamanteus*) and evidence for pervasive positive selection across toxin classes. *Toxicon* **57**:657-671.

Rokyta DR, Lemmon AR, Margres MJ, Aronow K. 2012. The venom-gland transcriptome of the Eastern diamondback rattlesnake (*Crotalus adamanteus*). *BMC Genomics* **13**:312.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242**:84-89.

Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP. 2005. The DNA sequence of the human X chromosome. *Nature* **434**:325-337.

Rudd S. 2003. Expressed sequence tags: Alternative or complement to whole genome sequences? *Trends Plant Sci* **8**:321-329.

Ruder T, Sunagar K, Undheim EA, Ali SA, Wai T, Low DH, Jackson TN, King GF, Antunes A, Fry BG. 2013. Molecular phylogeny and evolution of the proteins encoded by coleoid (cuttlefish, octopus, and squid) posterior venom glands. *J Mol Evol* **76**:192-204.

Rupp RA, Kruse U, Multhaup G, Gobel U, Beyreuther K, Sippel AE. 1990. Chicken NFI/TGGCA proteins are encoded by at least three independent genes: NFI-A, NFI-B and NFI-C with homologues in mammalian genomes. *Nucleic Acids Res* **18**:2607-2616.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**:5463-5467.

Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**:2012-2018.

Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NT, Schreiber SL, Mellor J, Kouzarides T. 2002. Active genes are tri-methylated at K4 of histone H3. *Nature* **419**:407-411.

Sanz L, Escolano J, Ferretti M, Biscoglio MJ, Rivera E, Crescenti EJ, Angulo Y, Lomonte B, Gutiérrez JM, Calvete JJ. 2008. Snake venomics of the South and Central American bushmasters. Comparison of the toxin composition of *Lachesis muta* gathered from proteomic versus transcriptomic analysis. *J Proteomics* **71**:46-60.

Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T. 2003. Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat Cell Biol* **6**:73-77.

Schubeler D, MacAlpine DM, Scalzo D, Wirbelauer C, Kooperberg C, van Leeuwen F, Gottschling DE, O'Neill LP, Turner BM, Delrow J et al. 2004. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev* **18**:1263-1271.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**:1086-1092.

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.

Secor SM, Diamond J. 1998. A vertebrate model of extreme physiological regulation. *Nature* **395**:659-662.

Secor SM. 2008. Digestive physiology of the Burmese python: Broad regulation of integrated performance. *J Exp Biol* **211**:3767-3774.

Selvey S, Thompson E, Matthaei K, Lea RA, Irving MG, Griffiths LR. 2001. B-actin—an unsuitable internal control for RT-PCR. *Mol Cell Probes* **15**:307-311.

Sen R, Baltimore D. 1986. Inducibility of κ immunoglobulin enhancer-binding protein NF-κB by a posttranslational mechanism. *Cell* **47**:921-928.

Serebrovsky A. 1938. Genes scute and achaete in *Drosophila melanogaster* and a hypothesis of gene divergency. *CR Acad Sci URSS* **19**:77-81.

Serrano SM, Maroun RC. 2005. Snake venom serine proteinases: Sequence homology vs. substrate specificity, a paradox to be solved. *Toxicon* **45**:1115-1132.

Shaffer HB, Minx P, Warren DE, Shedlock AM, Thomson RC, Valenzuela N, Abramyan J, Amemiya CT, Badenhorst D, Biggar KK. 2013. The Western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biol* **14**:R28.

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.

Shows T, Alper C, Bootsma D, Dorf M, Douglas T, Huisman T, Kit S, Klinger H, Kozak C, Lalley P. 1979. International system for human gene nomenclature (1979) ISGN (1979). Cytogenet Genome Res **25**:96-116.

Siang AS, Doley R, Vonk FJ, Kini RM. 2010. Transcriptomic analysis of the venom gland of the red-headed krait (*Bungarus flaviceps*) using expressed sequence tags. *BMC Mol Biol* **11**:24.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Res* **19**:1117-1123.

Simpson JT, Durbin R. 2012. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Res* **22**:549-556.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: Key considerations in genomic analyses. *Nat Rev Genet* **15**:121-132.

Song C, Liu S, Xiao J, He W, Zhou Y, Qin Q, Zhang C, Liu Y. 2012. Polyploid organisms. *Sci China Life Sci* **55**:301-311.

Stephens S. 1951. Possible significance of duplication in evolution. *Adv Genet* **4**:247-265.

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

St John JA, Braun EL, Isberg SR, Miles LG, Chong AY, Gongora J, Dalzell P, Moran C, Bed'hom B, Abzhanov A et al. 2012. Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biol* **13**:415.

Strydom D. 1973. Snake venom toxins: The evolution of some of the toxins found in snake venoms. *Syst Biol* **22**:596-608.

Suhr S, Kim D. 1996. Identification of the snake venom substance that induces apoptosis. *Biochem Biophys Res Commun* **224**:134-139.

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.

Sunagar K, Undheim EA, Scheib H, Gren EC, Cochran C, Person CE, Koludarov I, Kelln W, Hayes WK, King GF. 2014. Intraspecific venom variation in the medically significant Southern pacific rattlesnake (*Crotalus oreganus helleri*): Biodiscovery, clinical and evolutionary implications. *J Proteomics* **99**:68-83.

Swift H. 1950. The constancy of desoxyribose nucleic acid in plant nuclei. *Proc Natl Acad Sci U S A* **36**:643-654.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**:2731-2739.

Taylor JS, Raes J. 2004. Duplication and divergence: The evolution of new genes and old ideas. *Annu Rev Genet* **38**:615-643.

Taylor JS, Raes J. 2005. Small-scale gene duplications. In: Gregory, TR (Eds). *The Evolution of the Genome*. Elsevier Academic press, Oxford. Pp. 289-327.

Thomas Jr C. 1971. The genetic organization of chromosomes. *Annu Rev Genet* **5**:237-256.

Thomas R, Pough FH. 1979. The effect of rattlesnake venom on digestion of prey. *Toxicon* **17**:221-228.

Torres AM, Wong HY, Desai M, Moochhala S, Kuchel PW, Kini RM. 2003. Identification of a novel family of proteins in snake venoms. Purification and structural characterization of nawaprin from *Naja nigricollis* snake venom. *J Biol Chem* **278**:40097-40104.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics* **25**:1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**:511-515.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat Protocols* **7**:562-578.

Tsai IJ, Otto TD, Berriman M. 2010. Method improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* **11**:R41.

Tu AT, Hendon RR. 1983. Characterization of lizard venom hyaluronidase and evidence for its action as a spreading factor. *Comp Biochem Physiol B Comp Biochem* **76**:377-383.

Tytgat J, Chandy KG, Garcia ML, Gutman GA, Martin-Eauclaire M, van der Walt, Jurg J, Possani LD. 1999. A unified nomenclature for short-chain peptides isolated from scorpion venoms: A-KTx molecular subfamilies. *Trends Pharmacol Sci* **20**:444-447.

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.

Undheim EA, King GF. 2011. On the venom system of centipedes (chilopoda), a neglected group of venomous animals. *Toxicon* **57**:512-524.

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.

Van de Peer Y, Meyer A. 2005. Large-scale gene and ancient genome duplications. In: Gregory, TR (Eds). *The Evolution of the Genome*. Elsevier Academic press, Oxford. Pp. 328-368.

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett Jr DE, Hieter P, Vogelstein B, Kinzler KW. 1997. Characterization of the yeast transcriptome. *Cell* **88**:243-251.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al. 2001. The sequence of the human genome. *Science* **291**:1304-1351.

Vergara I, Pedraza-Escalona M, Paniagua D, Restano-Cassulini R, Zamudio F, Batista CV, Possani LD, Alagón A. 2014. Eastern coral snake *Micrurus fulvius* venom toxicity in mice is mainly determined by neurotoxic phospholipases A_2. *J Proteomics* **105**:295-306.

Vetter RS, Visscher PK. 1998. Bites and stings of medically important venomous arthropods. *Int J Dermatol* **37**:481-496.

Vicoso B, Emerson J, Zektser Y, Mahajan S, Bachtrog D. 2013. Comparative sex chromosome genomics in snakes: Differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS Biology* **11**:e1001643.

Vidal N, Hedges SB. 2002. Higher-level relationships of caenophidian snakes inferred from four nuclear and mitochondrial genes. *Comptes Rendus Biologies* **325**:987-995.

Vidal N, Hedges SB. 2004. Molecular evidence for a terrestrial origin of snakes. *Proc Biol Sci* **271**:S226-9.

Vidal N, Hedges SB. 2005. The phylogeny of squamate reptiles (lizards, snakes, and amphisbaenians) inferred from nine nuclear protein-coding genes. *Comptes Rendus Biologies* **328**:1000-1008.

Vikrant S, Verma BS. 2013. Monitor lizard bite-induced acute kidney injury–a case report. *Renal failure* **36**:444-446.

Vonk FJ, Richardson MK. 2008. Developmental biology: Serpent clocks tick faster. *Nature* **454**:282-283.

Vonk FJ, Jackson K, Doley R, Madaras F, Mirtschin PJ, Vidal N. 2011. Snake venom: From fieldwork to the clinic. *Bioessays* **33**:269-279.

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.

von Reumont BM, Blanke A, Richter S, Alvarez F, Bleidorn C, Jenner RA. 2014. The first venomous crustacean revealed by transcriptomics and functional morphology: Remipede venom glands express a unique toxin cocktail dominated by enzymes and a neurotoxin. *Mol Biol Evol* **31**:48-58.

Wagstaff SC, Harrison RA. 2006. Venom gland EST analysis of the saw-scaled viper, *Echis ocellatus,* reveals novel α_9 β_1 integrin-binding motifs in venom metalloproteinases and a new group of putative toxins, renin-like aspartic proteases. *Gene* **377**:21-32.

Wagstaff SC, Laing GD, Theakston RDG, Papaspyridis C, Harrison RA. 2006. Bioinformatics and multiepitope DNA immunization to design rational snake antivenom. *PLoS Med* **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.

Wang Z, Pascual-Anaya J, Zadissa A, Li W, Niimura Y, Huang Z, Li C, White S, Xiong Z, Fang D. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat Genet* **45**:701-706.

Warner TG, Dambach LM, Shin JH, O'Brien JS. 1981. Purification of the lysosomal acid lipase from human liver and its role in lysosomal lipid hydrolysis. *J Biol Chem* **256**:2952-2957.

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

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S. 2010. The genome of a songbird. *Nature* **464**:757-762.

Washkowitz AJ, Gavrilov S, Begum S, Papaioannou VE. 2012. Diverse functional networks of Tbx3 in development and disease. *Wiley Interdiscip Rev Syst Biol Med* **4**:273-283.

Weinstein SA, Keyler DE, White J. 2012. Replies to Fry et al. (toxicon 2012, 60/4, 434–448). part A. analyses of squamate reptile oral glands and their products: A call for caution in formal assignment of terminology designating biological function. *Toxicon* **60**:954-963.

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M. 1992. A second-generation linkage map of the human genome. *Nature* **359**:794-801.

Westhoff G, Tzschätzsch K, Bleckmann H. 2005. The spitting behavior of two species of spitting cobras. *J Comp Physiol A* **191**:873-881.

Whittington C, Belov K. 2007. Platypus venom: A review. *Australian Mammalogy* **29**:57-62.

Whitton JL, Sheng N, Oldstone MB, McKee TA. 1993. A "string-of-beads" vaccine, comprising linked minigenes, confers protection from lethal-dose virus challenge. *J Virol* **67**:348-352.

Wiens JJ, Kuczynski CA, Townsend T, Reeder TW, Mulcahy DG, Sites JW,Jr. 2010. Combining phylogenomics and fossils in higher-level squamate reptile phylogeny: Molecular data change the placement of fossil taxa. *Syst Biol* **59**:674-688.

Wilde CD. 1986. Pseudogenes. *CRC Crit. Rev Biochem* **19**:323-352.

Wilkinson JA, Glenn JL, Straight RC, Sites Jr JW. 1991. Distribution and genetic variation in venom A and B populations of the Mojave rattlesnake (*Crotalus scutulatus scutulatus*) in Arizona. *Herpetologica* **47**:54-68.

Williford A, Demuth JP. 2012. Gene expression levels are correlated with synonymous codon usage, amino acid composition, and gene architecture in the red flour beetle, *Tribolium castaneum*. *Mol Biol Evol* **29**:3755-3766.

Winkler H. 1920. Verbreitung und ursache der parthenogenesis im pflanzen-und tierreiche. Jena, Verlag Fischer.

Woltering JM, Vonk FJ, Müller H, Bardine N, Tuduce IL, de Bakker MA, Knöchel W, Sirbu IO, Durston AJ, Richardson MK. 2009. Axial patterning in snakes and caecilians: Evidence for an alternative interpretation of the Hox code. *Dev Biol* **332**:82-89.

Woltering JM. 2012. From lizard to snake; behind the evolution of an extreme body plan. *Curr. Genomics* **13**:289-299.

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

Wooldridge B, Pineda G, Banuelas-Ornelas J, Dagda R, Gasanov S, Rael E, Lieb C. 2001. Mojave rattlesnakes (*Crotalus scutulatus scutulatus*) lacking the acidic subunit DNA sequence lack Mojave toxin in their venom. *Comp Biochem Physiol B Biochem Mol Biol* **130**:169-179.

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.

Wüster W, Crookes S, Ineich I, Mané Y, Pook CE, Trape J, Broadley DG. 2007. The phylogeny of cobras inferred from mitochondrial DNA sequences: Evolution of venom spitting and the phylogeography of the African spitting cobras (serpentes: Elapidae: *Naja nigricollis* complex). *Mol Phylogenet Evol* **45**:437-453.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S et al. 2014. SOAPdenovo-trans: *De novo* transcriptome assembly with short RNA-seq reads. *Bioinformatics* **30**:1660-1666.

Yamanouye N, Carneiro SM, Scrivano CN, Markus RP. 2000. Characterization of β-adrenoceptors responsible for venom production in the venom gland of the snake *Bothrops jararaca*. *Life Sci.* **67**:217-226.

Yamazaki Y, Hyodo F, Morita T. 2003a. Wide distribution of cysteine-rich secretory proteins in snake venoms: Isolation and cloning of novel snake venom cysteine-rich secretory proteins. *Arch Biochem Biophys* **412**:133-141.

Yamazaki Y, Takani K, Atoda H, Morita T. 2003b. Snake venom vascular endothelial growth factors (VEGFs) exhibit potent activity through their specific recognition of KDR (VEGF receptor 2). *J Biol Chem* **278**:51985-51988.

Yamazaki Y, Morita T. 2004. Structure and function of snake venom cysteine-rich secretory proteins. *Toxicon* **44**:227-231.

Yamazaki Y, Matsunaga Y, Tokunaga Y, Obayashi S, Saito M, Morita T. 2009. Snake venom vascular endothelial growth factors (VEGF-fs) exclusively vary their structures and functions among species. *J Biol Chem* **284**:9885-9891.

Zelanis A, Travaglia-Cardoso SR, de Fátima Domingues Furtado M. 2008. Ontogenetic changes in the venom of *Bothrops insularis* (serpentes: Viperidae) and its biological implication. *S Am J Herpetol* **3**:43-50.

Zelanis A, de Souza Ventura J, Chudzinski-Tavassi AM, de Fátima Domingues Furtado M. 2007. Variability in expression of *Bothrops insularis* snake venom proteases: An ontogenetic approach. *Comp Biochem Physiol C Toxicol Pharmacol* **145**:601-609.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Res* **18**:821-829.

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

Zhu M, Dahmen JL, Stacey G, Cheng J. 2013. Predicting gene regulatory networks of soybean nodulation from RNA-seq transcriptome data. *BMC Bioinformatics* **14**:278.

Župunski V, Kordiš D, Gubenšek F. 2003. Adaptive evolution in the snake venom kunitz/BPTI protein family. *FEBS Lett* **547**:131-136.