# Transcriptomics and venomics: implications for medicinal chemistry

**by Frédéric Ducancel[1]*, Jordi Durban[2] & Marion Verdenaud[3]**

[1]Department of ImmunoVirology, iMETI/DSV/CEA, Bt 02, 18 rue du Panorama
BP 6, 92265, Fontenay-aux-Roses, France.
Tel.: +33.01.46.54.83.18, E-mail: frederic.ducancel@cea.fr
[2]Instituto de Biomedicina de València, Jaume Roig, 11, 46010, Valencia, Spain. E-mail: jordi.durban@gmail.com
[3]Department of Pharmacology and Immunoanalysis, iBiTec-Saclay/DSV/CEA, Bt 152, CEA de Saclay, 91191 Gif sur Yvette Cedex, France
E-mail: marion.verdenaud@cea.fr

***Author for correspondence**

**Background:** Over the last three decades, transcriptomic studies of venom gland cells have continuously evolved, opening up new possibilities for exploring the molecular diversity of animal venoms, a prerequisite for the discovery of new drug candidates and molecular phylogenetics. **Discussion:** The molecular complexity of animal venoms is much greater than initially thought. In this review, we describe the different technologies available for transcriptomic studies of venom, from the original individual cloning approaches to the more recent global Next Generation Sequencing strategies. **Conclusions:** Our understanding of animal venoms is evolving, with the discovery of complex and diverse bio-optimised cocktails of compounds, including mostly peptides and proteins, which are now beginning to be studied by academic and industrial researchers.

## What is the natural venom resource?

Animal venoms are complex cocktails of several hundreds of components, most of which (≈ 90%) are proteins or peptides [1**]. Venom compounds are characterized by their capacity to recognise various targets, such as enzymes, ion channels and receptors. Their interaction with these targets results in direct or indirect effects on cell integrity, the central and/or peripheral nervous system, muscles and blood flow. This potential for biological activity is seen as a consequence of their continual refinement by natural evolution. Venoms are thus natural reservoirs containing many bioactive molecules that have been selected and recruited for their secretion, structural stability, functional plasticity and capacity to engage in precise molecular interactions with their targets (affinity, specificity, selectivity). The remaining 10% of venom components include various organic components, such as sugars, salts, amino acids, biogenic amines (**Defined key term N°1)**, nucleotides, nucleosides and nucleic acids (mRNAs, DNAs).

Recent technological developments in mass spectrometry for venom-based explorations (proteomic studies) and venom-gland gene expression profiling (transcriptomic studies) have revolutionised our perception of the nature of animal venoms. Indeed, the use of these approaches, either separately or together, has led to the discovery of hundreds of peptides and proteins in the venom of single venomous species, with cone snails and spiders appearing to possess the most diverse cocktails or “arsenals” of compounds [2**].

The global animal venom resource can therefore be seen as a collection of more than 40,000,000 compounds, mainly peptides and proteins, of which only ~3,000 are known and have been studied to any extent. Additional levels of venom diversity have also been observed, due to intraspecific variations of the venom components [3-5]. Such variation may occur between specimens of a defined species as a result of changes in biotope, diet, age, development, sex, season or geographic location. Differences have also been observed between dissected and milked (injected) venoms. Finally, one intriguing observation from proteomic studies is the co-existence (within a single venom) of numerous “related-sequences” of the same compound, of different lengths and concentrations. It remains unclear whether these longer and shorter venom molecule “isoforms” (**Defined key term N°2)**, have any biological significance. It has become clear that venom composition varies considerably during the life cycle of venomous animals and that these changes may affect the pharmacology and toxicity of the venom.

These findings have greatly modified the toxinologist’s view of venom gland systems, which now appear to be sophisticated, efficient and highly dynamic structures containing the venom itself, the venom-gland cells and associated tissues (i.e. the salivary glands in cone snails). Together with increases in our capacity to synthesise peptides and recombinant proteins, these changes have opened up new perspectives for the study and use of this huge natural resource by academic researchers and industry.

## Venoms: natural medicinal chemistry libraries and therapeutic potential

The amazing potential of venoms for drug discovery renders them of particular interest to the pharmaceutical industry, which is in dire need of innovation. In recent years, the successful release onto the market of therapeutic antibodies, proteins and peptides has increased interest in “biologics” as potential novel drugs. There are currently 51 therapeutic peptides on the market and more than 100 natural (and a similar number of modified) therapeutic proteins have been approved for clinical use in the European Union and the USA (*Source: 2010 report – Therapeutic Peptides Foundation*). In the domains of clinical and biotechnological applications, venom compounds in general, and peptides/toxins in particular, are naturally **bio-optimised** molecular tools that can be used for studies of their targets, potentially leading to the identification and development of novel candidate therapeutic molecules. Indeed, many toxin targets are involved in various human diseases, such as pain, cancer, neurodegenerative and cardiovascular diseases, diabetes, obesity and depression. Venoms contain stable bioactive molecules of high affinity, target subtype selectivity and large pharmacological spectra, yielding many promising candidates for innovative drug leads. Venom compounds constitute one of the most promising families of compounds for use in the diagnosis and treatment of human disease, due to their functional activities, small size, low immunogenicity and high stability, and the development of powerful strategies for their chemical synthesis or recombinant production. Furthermore, venom peptides and proteins bind to their targets through a large number of interactions, resulting in a decrease in the "capacity" of the targeted system to escape and to "resist" these ligands, through mutations affecting contact areas, for example (as observed in antibiotic resistance). Finally, the scaffold can be engineered for the design of compounds with modified biochemical, functional or biophysical properties, the adaptation of molecules for particular uses and for their labelling for *in-vivo* imaging, for vectorisation, or for the functionalisation of nanoparticles [7-9].

Few peptide drugs of venom origin are currently available, but significant developments are occurring in the fields of pain, infection and cancer [10**-12**]. The analgesic Prialt® (Ziconotide) is a peptide from marine snail (*Conus magus)* that was approved by the FDA in 2004 and used as the last resort in the treatment of severe pain in patients refractory to morphine. Other drugs derived from snake venom proteins are used in the control of hypertension and blood haemostasis. These drugs include the antihypertensive drug Capoten® (captopril) which mimics the action of the angiotensin-converting enzyme inhibitor peptide from a viper venom. On the other hand Integrilin® (eptifibatide) and Aggrastat® (tirofiban) are used in the treatment and prevention of acute coronary syndrome when iodinated chlorotoxin from scorpion venom (TM-601®) has successfully undergone phase II trials for the targeted treatment of glioma, a diffuse form of brain cancer. XEN2174 is an analgesic peptide derived from χ-conotoxin MrIA, which has successfully completed phase IIa clinical trials for cancer-related pain and is currently undergoing testing in phase IIb trials for postoperative pain. Many other venom peptides may one day lead to new drugs and more than 400 patents relating to venom peptides have been filed. Several peptides, mostly conotoxins, have reached preclinical or clinical trial stages. This field of research is very active, with rapid advances driven in particular by Australian groups in particular (University of Queensland and Xenome Ltd). A seminal development has recently occurred with the demonstration of oral activity in a rodent pain model for $\alpha$-conotoxin Vc1.1, a nicotinic receptor antagonist.

Venom proteins are also used for therapeutic applications. Several snake venom proteins with pro-, anti-coagulant or fibrinolytic activities have found uses in medical applications. For instance, ancrod (Viprinex®), from the Malayan pit viper *Calloselasma rhodostoma,* is a defibrinogenating molecule currently under investigation as a possible treatment for stroke. Pefakit® Reptilase®Time (batroxobin) from *Bothrops atrox* is used for investigations of the final phase of blood coagulation. Due to its heparin insensitivity, Reptilase®Time can detect fibrinogen polymerisation disorders even in the presence of heparin. Tirobifan (Aggrastat) is an antiplatelet drug derived from an *Echis carinatus* venom compound. Other proteins with anticoagulant properties from the venoms of the snakes *Bothrops atrox*, *Echis carinatus*, *Oxyuranus scutellatus* and *Daboia russelii* are used to diagnose coagulation disorders and to monitor anticoagulant treatment. Botrocetin™ is a strong platelet-aggregating protein found in *Bothrops atrox* venom. It is used for the diagnosis of several haemorrhagic vascular diseases of genetic origin, such as von Willebrand disease.

## General strategies for venom profiling

Given the tremendous diversity and potential for application of venom compounds, major efforts are required to characterise the molecular and structural diversities of this natural repertoire of bioactive molecules [13]. Traditionally, such characterisation has been based on the fractionation of venoms coupled with bioassay screening, followed by the purification and characterisation of bioactive compounds. This process is time-consuming, limiting the extent of exploration possible and leading to a focus on the most abundant molecules. In addition, “bioassay-guided” drug exploration is hampered by: the availability of material, sample size (most venomous animals are small or very small) and the complexity and variability of venoms.

Sequence-driven approaches, based on mass spectrometry analyses with or without the cloning and sequencing of precursors [14-22], have recently been developed. These approaches were initially based on the transcriptomic exploration of venom glands with expressed sequence tag (EST) technology, but several groups have since successfully applied next-generation sequencing strategies (NGS) [23*] (**Defined key term N°3)**. These highly sensitive and powerful strategies for precursor sequencing have facilitated more global and accurate transcriptomic explorations, paving the way for improvements in our understanding of the true molecular and structural diversity of the cocktails of molecules comprising animal venoms. NGS strategies, in addition to providing information about previously known families of venom compounds, are also, for the first time, providing access to the sequences of low abundance precursors. More importantly, exhaustive transcriptomics combined with annotation tools for bioinformatics is providing researchers with access to the unexplored “Eldorado” of “unknown” precursor sequences (encoding compounds with no equivalent in the available databases), which may account for as many as 20 to 40% of all venom compounds! These pools of precursors will undoubtedly include new sequences with original folding patterns associated with novel biological activities and target specificities.

## Venom-gland cell transcriptomics: from individual cloning to global view

### *Individual cloning of animal toxin precursors*

The group of Prof. Toru Tamiya was the first to clone and sequence a precursor (mRNA) encoding an animal toxin, in 1985 [24]. This mRNA encoded a short-chain neurotoxin, 65 amino acids in length, stabilised by four disulphide bridges. This peptide, erabutoxin a, is synthesised by a sea snake, *Laticauda semifasciata* (figure 1). The next precursor to be studied was that of a phospholipase $A_2$ ($PLA_2$) from the venom of the sea snake *Laticauda laticaudata* [25]. Since these pioneering works, many animal toxin precursors have been cloned and sequenced, revealing both common and distinctive structural elements leading to different precursor organisations (figure 2). Most animal toxin precursors are encoded by monocistronic sequences, with a single venom molecule encoded by a single mRNA. However, a few precursor sequences display a polycistronic organisation, with long open reading frames (ORFs) encoding several toxins [26-30]. The proteins encoded may be isoforms of the same family, as for sarafotoxins [26], when precursors from *Bothrops jararaca* and *Lachesis muta muta*, revealed the presence of seven bradykinin-potentiating peptides, together with one sequence encoding a C-type natriuretic peptide [27,28]. Polycistronic precursors encoding antimicrobial/cytolytic peptides and neurotoxins have also been cloned from the spider *Lachesana tarabaevi* [29] and the sea anemone *Antheopsis maculata* [30], respectively.

### *Towards global transcriptomic studies of venom-gland cells: the EST strategy*

Expressed sequence tag analysis for transcriptomics first appeared towards the end of the 20th century. ESTs are short (10 to 400 base pairs) single DNA sequencing reads obtained from genomic DNA or complementary DNA (cDNA) libraries (**Defined key term N°4)**. They are obtained by the classical Sanger sequencing method and correspond to one-shot partial-sequences of the 5' and/or 3' ends of cloned precursors. ESTs are useful for more global investigations of the gene expression patterns and overall transcriptomic activity of tissues or cells (e.g. venom gland cells).

The first ESTs library for venom gland-cells was described in 2001 [31]. In this pioneering work, the authors aimed to investigate the mechanisms by which cone snail had evolved. They analysed and compared 170 different conopeptide-encoding ESTs from five different species of *Conus* snails. They demonstrated: i) particularly rapid rates of nucleotide substitution within the sequences encoding the signal peptide, propeptide and mature peptide, ii) a bias, with transversions favoured over transitions in nucleotide substitutions (replacement of purines with pyrimidines or *vice versa*), iii) the presence of cysteine codons in specific positions within the hypervariable regions and iv) a preponderance of non-synonymous over synonymous substitutions in the mature peptide (**Defined key term N°5**).

Many EST-based transcriptomic studies have since been carried out on various venomous animals: jellyfish [32], spiders [33-38], snakes [39-48], cone snails [49,50], fish [51], wasps [52], scorpions [53], centipedes [54], and more recently, a venomous ant [55]. In all cases, EST annotation led to the identification of several new isoforms of previously known families of peptide or protein toxins, revealing a predominance of some families of venom components over others. Interestingly, the EST annotation of nucleic acid and protein sequence databases also revealed the existence of two new transverse (found in all venomous animal species) categories of venom-gland components. The first corresponds to sequences that have already been cloned, but the functions of which remain unknown. These sequences account for less than 10% of the annotated ESTs on average and, in many cases, they display similarity to sequences for components found in other tissues and phyla unrelated to venomous animals or to the venom apparatus. This observation has since been confirmed by NGS approaches (see below). These sequences probably correspond either to common cellular components or (in the case of secreted and disulphide-rich sequences) to particular peptide or protein scaffolds preferentially recruited by the venom gland to increase its molecular diversity [1, 2].

Another particularly exciting discovery has been the detection, by EST-based strategies, of venom-apparatus specific cellular and molecular actors displaying no sequence matches, for nucleotide or amino-acid sequences, with any sequence present in current databases. These molecules include compounds contributing to the originality and molecular specificity of the molecular machinery of the cells constituting the venom gland tissues.

Nevertheless, several ESTs are probably components of precursors encoding new peptides or proteins contributing to the richness of the venom arsenal developed by venomous animals. Future challenges in this field will include the development of tools and strategies for the exploration and study of these new sequences. Undoubtedly, these sequences will include some corresponding to new scaffolds and biological activities of potential relevance for new clinical applications.

However, although EST-sequencing constitutes significant progress with respect to the single-precursor cloning approach, it is subject to several limitations. The exploration of gene expression remains partial, with information gleaned mostly about the more abundant families of compounds. Such studies remain time-consuming and expensive, due to the sequencing technology used (Sanger sequencing). Furthermore, for long precursors, such as those encoding venom enzymes, the coverage of cloned sequences by ESTs is far from total, and classical second-step cloning by PCR amplification is required to obtain the complete precursor sequence. However, in such cases, the PCR primers are mostly deduced from 5' and 3' EST sequences.

### *Towards global transcriptomic studies of venom-gland cells: NGS approaches*

In 2005, an alternative DNA-sequencing technology (pyrosequencing) facilitating rapid, large-scale sequencing at low cost was described for the first time [56]. Various other approaches have since emerged and are grouped together under the umbrella term "Next-Generation Sequencing" [57, 58]. NGS uses various strategies to sequence DNA more efficiently than the traditional dideoxynucleotide method pioneered by Sanger [59]. NGS was initially developed for genomic sequencing projects, but was rapidly adopted for use in studies of the gene expression profiles (transcriptomics) of various tissues. For this purpose, mRNAs are extracted from venom gland cells, converted into complementary DNAs, fragmented and sequenced. Whatever be the NGS technology used, the sequenced fragments or "reads" (35 to 300-400 bp long) must be assembled into "contigs", which are ultimately analysed and annotated with bioinformatics labels. NGS techniques are high-throughput, with millions of sequencing reactions carried out in parallel.

Unlike previously described transcriptomic approaches, NGS techniques delivery a much broader view of the cocktail of compounds present in venom, as demonstrated by the increasing number of recent studies involving these approaches [60-82]. Over and above the compounds present in venom, these techniques can also be used to study the molecular machinery specifically used by the venom gland cells in the production of the venom arsenal and to elucidate the post-transcriptional mechanisms at work. In term of candidate drug discovery, deep-transcriptomics studies based on NGS are likely to revolutionise this field in terms of: i) the discovery of new isoforms, ii) the identification of toxin-related genes suggesting convergent recruitment by the venom gland tissues of compounds from diverse taxa [83], iii) the discovery of totally new compounds/scaffolds, and iv) the identification of a molecular signature specific to products of the venom gland cell.

NGS was first used to explore the transcriptomic activity of an animal venom apparatus in 2009. This study concerned the common emperor scorpion, *Pandinus imperator* [60]. Many other venom transcriptomes have since been determined, in snakes, cones snails, spiders, two venomous mammals, one venomous crustacean and one ant. Most of these studies involved i) pyrosequencing with Roche 454 or 454 GS FLX machines, ii) the Illumina GA/HiSeq System, ii) the Illumina system together with 454 technology and, very recently iv)

Ion Torrent technology (see table 1). As expected, these studies have resulted in deeper gene expression (transcriptomic) profiling of venom gland cells. Researchers now have access to a very large range of precursor representations, which was not the case with previous technologies. Thanks to these advances, it is now possible to detect and sequence transcripts present at very low abundance (a few copies only). It has also become possible to generate exhaustive inventories of the molecular diversity characterising families of known toxin peptides. For example, for *Conus consors*, we were able to show the predominance in the venom of three families of conopeptides: superfamilies A, O and M, with 132, 40 and 28 different isoforms, respectively [63]. Within the same transcriptome, precursors encoding conolysins, conantokins, contulakins, conotoxins P, S or T, conodipins and conopressins were detected even when there were fewer than four copies of the precursor present.

Clearly, NGS technologies provide a more complete picture of the composition of the cocktail of chemicals in venom. They have shown that venom gland cells synthesise an average of 200 to 400 different precursors, encoding peptides and proteins that are secreted into the lumen of the venom gland. Levels of transcriptomic activity may vary between studies. Intraspecific and interspecific variations are observed, but gene expression profiles may also differ between specimens of the same defined species of venomous animal.

As pointed out above, venoms appear to have a much broader molecular content than initially suspected, raising questions about the origin of this diversity. Deep transcriptomic analyses, particularly for *Conus*, have suggested a highly dynamic process of sequence diversification. One interesting study on *Conus miles* [71] focused on conopeptides and led to the identification of more than 650 putative conopeptide precursors. Durban *et al*. [71] studied the process driving venom evolution in the snake *Crotalus simus simus* [72], revealing the role played by certain populations of miRNAs (micro RNAs) as modulators of the ontogenetic composition of venoms.

NGS technologies are also used to shed some light on animals not considered truly venomous. Thus, several tick species produce highly paralytic and lethal cocktails of proteinaceous molecules in their salivary glands [84]. Detrimental effects of tick bites, such as paralysis, allergic reactions and pathogen transmission, have been reported in both animals and humans. Tick saliva has clearly evolved to contain a complex cocktail of components counteracting the effects of the host immune system, including anticoagulants, prostaglandins, immunosuppressants, antihistamines, prostacyclin and calreticulins [85]. Furthermore, fatal cases of human envenomation have been reported in Australia. A transcriptome analysis was recently carried out with a combination of 454 GS-FLX and Illumina NGS technologies [86], to improve our understanding of tick saliva. This study established that the saliva gland transcriptome of the Australian paralysis tick, *Ixodes ricinus,* contained housekeeping proteins (≈ 23%), secreted proteins (≈13 %) and unknown compounds (≈ 60 %). The secreted proteins included enzymes, protease inhibitors (basic tail and Kunitz domain families), lipocalins, ixostatins, antimicrobial peptides and 14 other different families of compounds. Together, these results confirm the huge diversity of molecules present in tick saliva, highlighting the importance of studies of this fluid.

Finally, a general feature of these recent NGS transcriptomes is the presence of a large proportion (ranging from 20% to 40%) of reads and contigs corresponding to previously unknown sequences. None of these sequences match any of the sequences present in currently available databases. This discovery, initially made in studies based on ESTs (see above), is clearly one of the most exciting results generated by NGS. The challenge now is for researchers to explore these new "repertoires", which undoubtedly contain precursors encoding venom compounds. Given the huge numbers of new sequences discovered in this way, toxinologists and bioinformaticians need to join forces, combining their expertise, knowledge and capacities in the development of new, efficient exploration algorithms. One possible strategy for identifying the precursors encoding venom compounds would involve making use of the molecular and structural features characteristic of venom compound precursors (see above and figure 2). This approach would involve searching for and identifying i) signal peptides, ii) cysteine-rich sequences, and iii) propeptide-like sequences. For example, the concomitant presence of cysteine residues and a signal sequence within a precursor, with or without N- and/or C-terminal putative propeptide sequences, should be considered as strong indications that the precursor encodes a venom peptide or protein. Cross-referencing with proteomics data will be useful, to confirm this prediction and to identify the start sites of mature sequences more precisely. On the basis of such results, candidate sequences could then be synthesised chemically or produced by recombinant technology, for assessments of their toxicity, biological activity and structure. This is the overall philosophy followed in the ongoing European project “VENOMICS” presented below.

### *VENOMICS: an ambitious European project (2011-2015)*

VENOMICS is a European FP7-Health project dedicated to the exploration of biodiversity for public health (http://www.venomics.eu/). It aims to explore animal venom compounds, with a view to the identification and development of novel biotherapeutics.
Bioassay-guided approaches have classically been used for drug discovery in venoms. However, this low-throughput strategy requires large amounts of venom and focuses principally on compounds abundant in venom. VENOMICS makes use of an innovative “Omics” workflow involving cutting-edge high-throughput transcriptomics, proteomics and peptide production technologies to decipher venom diversity (figure 3).

The core objective of VENOMICS is to recreate *in vitro* collections of venom peptides that can be used as a resource for high-throughput screening, for the more efficient isolation of novel drug leads. This approach is akin to recreating "synthetic venoms" in the laboratory. VENOMICS aims to reach this goal, by:

i) Generating venoms and venom gland biobanks, corresponding to 200 venomous animal species,
ii) Sequencing venom peptides by proteomics and transcriptomics (Illumina technology) approaches, to create a database of 50,000 sequences for mature venom proteins and peptides,
iii) High-throughput *in vitro* chemical synthesis or the recombinant expression of selected peptides, to generate a bank of several thousand peptides,

and

iv) Pharmacological screening of the peptides bank against selected molecular targets and drug lead generation.

This strategy represents a new paradigm for venom-based drug discovery, differentiating the VENOMICS approach from the classical bioassay-guided process. VENOMICS focuses on disulphide bridge-rich venom peptides no longer than about 100 amino acids long. Candidate peptides of less than 40 amino acids in length are generated by chemical synthesis, whereas larger peptides are produced principally by bacteria, with recombinant technologies.

Venom-derived peptide libraries offer the advantage of containing only highly stable natural bioactive and bio-optimised molecules of high affinity and target subtype selectivity. They therefore have a much greater potential for drug discovery than the randomly generated peptide libraries obtained by phage display or combinatorial chemistry techniques. The VENOMICS consortium consists of research laboratories and SMEs from Belgium, Denmark, France, Portugal and Spain.

### *Transcriptomics and molecular phylogenetic studies*

Phylogenetics is used to study or trace the evolutionary history underlying biological diversity in groups of organisms. It has been said that "Nothing in biology makes sense except in the light of evolution" (Theodosius Grygorovych Dobzhansky) [87]. Phylogenetics therefore plays a key role in various areas of biology, including population genetics, ecology and animal behaviour, but also in the clinical and medical contexts, giving rise to what has been called "evolutionary medicine" [88]. Phylogeny has been used to determine the origin and spread of a contagious disease from a molecular epidemiology standpoint [89, 90] and to understand the adaptive evolution of viral pathogens, to facilitate vaccine design [91]. However, one of the most interesting fields in which phylogenetics is proving increasingly valuable is pharmaceutical research for the identification of new drugs or natural products. Wang *et al*. [92] used this approach to identify specific peptides from the tammar wallaby that effectively killed multidrug-resistant bacteria. Phylogenetic analysis could also be used to find genes with a common phylogenetic profile involved in a similar biological pathway or sharing a similar biochemical function. Komatsu *et al.* [93] used this approach to determine which species of *Panax* were most closely related to other medicinal species and might therefore have similar medicinal qualities.

In this context, the FP7-Health project VENOMICS aims to explore venoms from non-model organisms, to identify new therapeutic compounds. As described above, the profiling of those organisms, for which genome sequences are unavailable, has been based on transcriptomic data in particular, adding fuel to the controversy between morphologists and molecular biologists in the field of phylogenetic systematics [94,95], because classical phylogenetic approaches used morphological data to determine taxonomic relationships. Nevertheless, molecular data are becoming a valuable source of information and although taxonomy is still based largely on Linnaean principles and morphological characters, information from DNA and proteins has been used to call into question previous taxonomic classifications based purely on morphological traits.

Studies of the changes in gene expression underlying phenotypic divergence have successfully increased our understanding of transcriptome evolution in several organisms, including mammals [96], fishes [97], mosquitoes [98], molluscs [99], turtles [100] and plants [101]. In systematics, information of this kind can be used to clarify unexpected evolutionary relationships, such as those mentioned above for Colubridae, Viperidae and Elapidae. However, in the context of the VENOMICS project, drug discovery could benefit from phylogenetic studies of the transcriptomes of venomous species [69], because venom toxins probably evolved from proteins with normal physiological function, and comparative phylogenetics might provide clues about disease-related proteins.

Analyses of transcriptome data for non-model organisms for phylogenetic purposes have been driven by the rapid development of sequencing technology. Several phylogenetic surveys [102-106] have been performed on expression sequence tags (ESTs). At this point, it should be pointed out that ESTs obtained from different taxa cannot contain overlapping genes, given the low probability of finding orthologous sequences in a reduced set of sequences. This makes it more difficult to attain comparative phylogenetics goals.

Since Next-Generation Sequencing technologies have become easily affordable due to the drop of per base sequencing costs. As a result, RNA-Seq [107], characterisation of the complete set of transcripts by massive parallel sequencing processes, is becoming a highly valuable tool for exploring the complexity of organisms at the genome-wide scale, giving rise to the field of phylogenomics [108]. This field, at the intersection of evolution and genomics, involves the inference of the phylogenetic history of certain organisms from genome-wide data. Chan *et al*. [109**] coined the term 'Next-generation phylogenomics' for such large-scale phylogenetics sequencing projects, and Lin *et al*. [110] recently performed a phylogenomic analysis of subterranean mammals, using four *de novo* RNA-seq libraries.

Since the first analysis with 454 pyrosequencing methods for phylogenetic purposes performed by Roeding *et al*. [60], this technology has been applied for phylogeographic purposes by McCormack *et al.* [111] and by Rokyta *et al.* [112], for description of the positive selection (reflected by nonsynonymous substitutions) imposed on most venom proteins in *Crotalus adamanteus*. Similarly, Dutertre *et al.* [113**] recently showed that *Conus geographus* has defence-evoked and predation-evoked venoms, and that the conotoxins found in these two different types of venom have evolved rapidly under positive Darwinian selection. However, improvements in the read length, cost per Mb sequenced, total throughput and speed of Illumina technology have made this platform the technology of choice. It has recently been used in several phylogenetic studies [110,114,115].

Given NGS transcriptomic data from a non-model organism, it would be interesting to identify the specific features to be taken into account for phylogenetic surveys. Transcriptome sequences generated with high-throughput techniques have been shown to provide a rich set of characters for phylogenetic studies in eukaryotes. However, the alignment of multiple partial sequences might result in an alignment with large numbers of gaps, potentially compromising the conclusions of any phylogenetic study [116]. It remains unclear to what extent the reliability of tree reconstruction is increased by maximising either the number of taxa or the number of characters studied. McCormack *et al*. [111] recently concluded that the loci identified from 454 pyrosequencing data could be useful for phylogenetics, population genetics or phylogeographic purposes, particularly for closely related species. Despite being incomplete, most of the 454 sequences were useful for determining which toxin protein sequence had been identified in snake venom gland transcriptomes. According to John J. Wiens [117], missing data are less critical than complete characters tied to other taxa in the tree for resolving phylogenetic relationships, and limited taxonomic sampling could be problematic in the downstream phylogenetic analysis. An increase in the number of taxa sampled could, therefore increase confidence levels for the assignment of certain orthologues.

Several features should also be taken into account when carrying out phylogenetic analyses with transcriptome data:

1. Downstream bioinformatics analyses of NGS data cannot currently identify potential pseudogene sequences, relics of the evolution that might lead to incorrect conclusions about phylogenetic relationships (**Defined key term N°6**).

2. Gene trees and species trees are not the same [118]. If there were duplications or polymorphic alleles, then phylogenies for genes will not match those for organisms. In such cases, the phylogenetic reconstruction of orthologous sequences could be useful, to generate the species tree.

3. Conclusions for nuclear genes may conflict with those for mitochondrial genes. Moreover, it seems that the combined use of mitochondrial and nuclear sequences yields better results, without artefacts for nodes for which mitochondrial and nuclear gene datasets used separately generate conflicting topologies [119].

## Future perspective: venom apparatus transcriptomics

Our understanding of animal venoms is clearly evolving, with the revelation that venoms are complex and diverse bio-optimised cocktails of compounds, mostly peptides and proteins, that we are only just beginning to explore. These changes in our perception are closely linked to technological advances and one of the current difficulties facing researchers is establishing ways to analyse, interpret and valorise the huge numbers of sequences being generated.

Efficient data analysis will require the use of bioinformatics knowhow and concepts, together with the development from scratch of new strategies and scripts for the exploration of “Unknown” fractions in particular. One of the difficulties is identifying, as precisely as possible, the true N-terminus of the mature compounds encoded by the totally new precursor sequences emerging from NGS transcriptomic studies. With this goal in mind, the combination of this approach with proteomic analyses of the corresponding venoms may facilitate identification, but only for compounds present at sufficiently high concentrations in the venom studied.

Valorisation will require high-throughput strategies for the synthesis or production of compounds of interest in their native forms. This implies a knowledge of the exact amino-acid sequences of these molecules, the number and location of the disulphide bridges, and the nature and position of post-translational modifications, when present. Again, the combination of these techniques with proteomics appears to be the most appropriate strategy. In the case of a totally new scaffold, one key issue is the native fold adopted by the molecule of interest. If present in sufficiently large amounts in the venom, the new compound could be extracted, purified and studied by X-ray diffraction or NMR. Compounds present at too low a concentration for this approach could be produced by recombinant technologies, by following a strategy allowing the formation of disulphide bridges *in vivo* before the initiation of three-dimensional structure studies.

The development of more systematic and automated strategies for screening for biological activity is also an important issue, for the identification of new biotherapeutic hits. In particular, the development of straightforward screening tools compatible with high-throughput technologies is crucial, to increase the chances of successful exploration. Two principal approaches appear pertinent: i) concentrating the screening of compound diversity on one target of interest, or ii) presenting the libraries of molecules to be tested to libraries of targets. The hits identified in this way then enter a phase of exhaustive exploration of their biochemical, functional, structural and biological activities. These studies will, in many instances, require the mutation, engineering of labelling of compounds of interest.

In conclusion, although the techniques for exploring the resources provided by venomous animals are maturing, these analyses remain time-consuming and must be combined with complementary approaches. Nevertheless, the exploration of **large** and **naturally bio-optimised** libraries of compounds, such as those produced by the venom-gland systems, clearly constitutes a major advance in medicinal biochemistry.

## Executive summary

- Given the existence of several thousand of venomous animal species, the venom resource is huge, opening tremendous perspectives in the fields of fundamental research and development of therapeutics.
- Technological developments such as high-throughput proteomics and transcriptomics, result in a more complex molecular vision of venoms, that appear to be highly diverse cocktails of peptides and proteins mainly, whose composition is susceptible to vary upon different criteria or stimuli.
- Resulting from a long process of molecular evolution and selection, venom peptides and proteins are naturally bio-optimised scaffolds.
- Due their functional activities, reduced sizes, low immunogenicity and their high stability, venom peptides constitute one of the most promising families of compounds for use in the diagnosis and treatment of human diseases.
- Venom peptide-drugs are associated to significant developments occurring in the fields of pain, infection and cancer, when venom-proteins target mainly the human cardio-vascular system.
- The study of the transcriptomic activity of venom-gland cells has started in 1985 with individual cloning of toxin precursors, to reach today more global descriptions of the high activity of synthesis of venom-gland cells.
- Expressed Sequence Tags, then Next Generation Sequencing strategies are responsible for that evolution.
- Aside previously known families of venom peptides and proteins, these more global approaches have revealed new groups of compounds among which several display totally new sequences whose activity and structure are unknown to date! Exploration of these “Unknown” fractions, require the development of new bioinformatics tools to tentatively identify precursors encoding venom components.
- VENOMICS is an ambitious international/European project that aims at applying high-throughput technologies to explore the molecular diversity of venoms with the objective of reproducing artificially a part of it for drug-candidates screening.
- Next Generation Sequencing data requires new strategies of sequence alignments leading to next-generation phylogenomics.

## Defined key terms

1) **Biogenic amines**/ Biogenic amines consist of naturally biologically active enzymatic-products that contain one or more primary amine groups, such as norepinephrine, histamine, and serotonin. They act primarily as neurotransmitters.

2) **Isoforms**/ An isoform is a natural mutant/variant of a peptide/protein of reference, which differ from that "reference" sequence by a minimum of one amino acid. All these mutants/variants/isoforms belong to the same structural and functional family.

3) **NGS**/ Next Generation Sequencing technologies do not use the classical Sanger's strategy of DNA sequencing. To date they correspond mainly to 454-pyrosequencing (Roche), to paired-end (Illumina-Solexa) or to semiconductor (Ion-Torrent$^{TM}$) sequencing strategies.

4) **cDNA**/ Complementary DNA is generated by the reverse transcription of messenger RNAs (mRNAs) extracted from a tissue or pool of cells of interest. Messengers RNAs correspond to copies of genes that are translated *via* the ribosome machinery into polymers of amino acids: peptides or proteins. As such mRNAs are also named "precursors".

5) **Synonymous *versus* nonsynonymous substitutions**/ A synonymous substitution corresponds to a silent mutation of a codon that results in an unchanged amino acid. On the other hand, a nonsynonymous substitution is an amino change within a given sequence that results from a non-silent mutation of the corresponding codon.

6) **Pseudogenes**/ Pseudogenes are genomic ubiquitous and abundant non-functional DNA sequences similar to normal and active genes. They are identified during genome annotation process and contain different type of modifications (mutations/insertions/deletions…) that results in their non-functionality. It is recognized that some of them play essential role in gene regulation of their parent and functional genes.

## Bibliography

Casewell NR, Wüster W, Freek J *et al*. Complex cocktails: the evolutionary novelty of venoms. *Trends in Ecology and Evolution* 28(4), 219-229 (2013).

*** A review that illustrates how, throw the development of "omic" technologies, our perception of the venom systems is evolving.*

Fry BG, Scheib H, van der Weerd L *et al*. Evolution of an arsenal. *Molecular & Cellular Proteomics* 7(2), 215-246 (2008).

*** An illustration of how toxin scaffolds have been recruited and evolved from endogenous families of peptides or proteins.*

Abdel-Rahman MA, Omran MA, Bdel-Nabi IM *et al*. Intraspecific variation in the Egyptian scorpion *Scorpio maurus palmatus* venom collected from different biotopes. *Toxicon* 53, 349-359 (2009).

Creer S, Malhotra A, Thorpe RS *et al*. Genetic and ecological correlates of intra-specific variation in pitviper venom composition detected using matrix-assisted laser desorption time-of-flight mass spectrometry (MALDI-TOF-MS) and isoelectric focusing. *J. Mol. Evolution* 56, 317-329 (2003).

Dutertre S, Biass D, Stöcklin R *et al*. Dramatic intraspecimen variations within the injected venom of *Conus consors*: an unsuspected contribution to venom diversity. *Toxicon* 55, 1453-1462 (2010).

Kola I. The state of innovation in drug development. *Clin Pharmacol Ther* 83(2), 227-230 (2008).

Rennert R, Neundorf I, Beck-Sickinger AG. Synthesis and application of peptides as drug carriers. *Methods Mol. Biol.* 535, 389-403 (2009).

Shaji J, Patole V. Protein and peptide drug delivery: oral approaches. *Indian J of Pharmaceutical Sciences* 70(3), 269-277 (2008).

Hayashi MAF, Ducancel F, Konno K. Natural peptides with potential applications in drug development, diagnosis, and/or biotechnology. *Int. J. Pept* doi: 10.1155/2012/757838 (2012).

Lewis RJ, Garcia ML. Therapeutic potential of venom peptides. *Nature Rev Drug Discovery* 2, 790-802 (2003).

*** A survey of the huge pharmacology diversity of venom peptides and their therapeutic potentials.*

Cossins D. From toxins to therapeutics. *The Scientist* March 19, (2013).

King GF. Venoms as a platform for human drugs: translating toxins into therapeutics. *Expert Opin Biol Ther* 11(11), 1469-1484 (2011).

*** An illustration of the great interest of using selected toxin scaffolds to identify and develop therapeutics.*

Calvete JJ. Snake venomics: from the inventory of toxins to biology. *Toxicon* 75, 44-62 (2013).

Escoubas P, Sollod B, King GF. Venom landscapes: mining the complexicity of spider venoms via a combined cDNA and mass spectrometric approach. *Toxicon* 47(6), 650-663 (2006).

Calvete JJ, Juarez P, Sanz L. Snake venomics. Strategy and applications. *J. Mass Spectrom* 42(11), 1405-1414 (2007).

Escoubas P, King GF. Venomics as a drug discovery platform. *Expert Rev. Proteomics* 6(3), 221-224 (2009).

Calvete JJ. Venomics, what else? *Toxicon* 60(4), 427-433 (2012).

Prashanth JR, Lewis RJ, Dutertre S. Towards an integrated venomics approach for accelerated conopeptide discovery. *Toxicon* 60(4), 470-477 (2012).

Favreau P, Menin L, Michalet S *et al*. Mass spectrometry strategies for venom mapping and peptide sequencing from crude venoms: case applications with single arthropod specimen. *Toxicon* 47, 676-687 (2006).

Rodrigues RS, Boldrini-França J, Fonseca FPP *et al*. Combined snake venomics and venom gland transcriptomic analysis of *Bothropoides pauloensis*. *J. Proteomics* 75, 2707-2730 (2012).

Margres MJ, McGivern JJ, Wray KP *et al*. Linking the trascriptome and proteome to characterize the venom of the eastern diamondback rattlesnake (*Crotalus adamanteus*). *J. of Proteomics* 96, 145-158 (2014).

Calvete JJ. Snake venomics: from the inventory of toxins to biology. *Toxicon* 75, 44-62 (2013).

Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods* 5, 16-18 (2008).

* *A statement of the different technological evolutions that make transcriptomics to enter into a new age.*

Tamiya T, Lamouroux A, Julien JF *et al*. Cloning and sequence analysis of the cDNA encoding a snake neurotoxin precursor. *Biochimie* 67(2), 185-189 (1985).

Guignery-Frelat G, Ducancel F, Ménez A *et al*. Sequence of a cDNA encoding a snake venom phospholipase A2. *Nucleic Acids Res* 15(14), 5892 (1987).

Ducancel F. Endothelin-like peptides. *Cell Mol Life Sci* 62(23), 2828-2839 (2005).

Murayama N, Hayashi MA, Ohi H *et al.* 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 USA* 94(4), 1189-1193 (1997).

Soares MR, Oliveira-Carvalho AL, Wermelinger LS *et al*. Identification of novel bradykinin-potentiating peptides and C-type natriuretic peptides from *Lachesis muta muta* venom. *Toxicon* 46(1), 31-38 (2005).

Kozlov SA, Vassilevski AA, Foefanov AV *et al*. Latarcins, antimicrobial and cytolytic peptides from the venom of the spider *Lachesana tarabaevi* (Zodariidae) that exemplify biomolecular diversity. *J. Biol. Chem* 281(30), 20983-20992 (2006).

Honma T, Hasegawa Y, Ishida M *et al*. Isolation and molecular cloning of novel peptide toxins from the sea-anemone *Antheopsis maculata*. *Toxicon* 45(1), 33-41 (2005).

Conticello SG, Gilad Y, Avidan N *et al*. Mechanisms for evolving hypervariability: the case of conopeptides. *Mol. Biol. Evol* 18(2), 120-131 (2001).

Yang Y, Cun S, Xie X *et al*. EST analysis of gene expression in the tentacle of *Cyanea capillata*. *FEBS Lett* 538(12), 183-191 (2003).

Kozlov S, Malyavka A, McCutchen B *et al*. A novel strategy for the identification of toxin-like structures in spider venom. *Proteins* 59(1), 131-140 (2005).

Jiang L, Peng L, Chen J *et al*. Molecular diversification based on analysis of expressed sequence tags from the venom glands of the Chinese bird spider *Ornithoctonus huwena. Toxicon* 51(8), 1479-1489 (2008).

Jiang L, Liu C, Duan Z *et al*. Transcriptome analysis of venom glands from a single fishing spider *Dolomedes mizhoanus*. *Toxicon* 73, 23-32 (2013).

Chen J, Zhao L, Jiang L *et al*. Transcriptome analysis revealed novel possible venom components and cellular processes of the tarantula *Chilobrachys jingzhao* venom gland. *Toxicon* 52(7), 794-806 (2008).

Tang X, Zhang Y, Hu W *et al*. Molecular diversification of peptide toxins from the tarantula *Haplopelma hainanum* (*Ornithoctonus hainana*) venom based on transcriptomic, peptidomic and genomic analyses. *J. Proteome Res* 9(5), 2550-2564 (2010).

McCowan C, Garb JE. Recruitment and diversification of an ecdysozoan family of neuropeptide hormones for the black widow spider venom expression. *Gene* 536(2), 366-375 (2014).

Wagstaff SC, Harrison RA. Venom gland EST analysis of the saw-scaled viper *Echis ocellatus*, reveals novel alpha9beta1 integrin-binding motifs in venom metalloproteinases and a new group of putative toxins, renin-like aspartic proteases. *Gene* 377, 21-32 (2006).

Junqueira-de-Azevedo IL, Ching AT, Carvalho E *et al*. *Lachesis muta* (Viperidae) cDNAs reveal diverging pit viper molecules and scaffolds typical of cobra (Elapidae) venoms: implications for snake toxin repertoire evolution. *Genetics* 173(2), 877-889 (2006).

Cidade DA, Simão TA, Dávilla AM *et al*. *Bothrops jararaca* venom gland transcriptome: analysis of the gene expression pattern. *Toxicon* 48(4), 437-461 (2006).

Jia Y, Cantu BA, Sánchez EE *et al*. Complementary DNA sequencing and identification of mRNAs from the venomous gland of *Agkistrodon piscivorus leucostoma*. *Toxicon* 51(8), 1457-1466 (2008).

Jiang Y, Li Y, Lee W *et al*. Venom gland transcriptomes of two elapid snakes (*Bungarus multicinctus* and *Naja atra*) and evolution of toxin genes. *BMC Genomics* 12:1. doi: 10.1186/1471-2164-12-1 (2011).

Rodrigues RS, Boldrini-França J, Fonseca FPP *et al*. Combined snake venomics and venom gland transcriptomic analysis of *Bothropoides pauloensis. J. Proteomics* 75(9), 2707-2720 (2012).

Chatrath ST, Chapeurouge A, Lin Q *et al*. Identification of novel proteins from the venom of a cryptic snake *Drysdalia coronoides* by a combined transcriptomics and proteomics approaches. *J. Proteome Res* 10(2), 739-750 (2011).

Zelanis A, Andrade-Silva D, Rocha MM *et al*. A transcriptomic view of the proteome variability of newborn and adult *Bothrops jararaca* snake venoms. *PLoS Negl Trop Dis* 6(3), e1554 (2012).

Ching ATC, Rocha MMT, Paes Leme AF *et al*. Some aspects of the venom proteome of the Colubridae snake *Philodryas olfersii* revealed from a Duvernoy's (venom) gland transcriptome. *FEBS Lett* 580(18), 4417-4422 (2006).

Ching ATC, Paes Leme AF, Zelanis A *et al*. Venomics profiling of *Thamnodynastes strigatus* unveils matrix metalloproteinases and other novel proteins recruited to toxin arsenal of rear-fanged snakes. *J. Proteome Res* 11(2), 1152-1162 (2012).

Pi C, Liu J, Peng C *et al*. Diversity and evolution of conotoxins based on gene expression profiling of *Conus litteratus*. *Genomics* 88(6), 809-819 (2006).

Pi C, Liu J, Peng C *et al*. Analysis of expressed sequence tags from the venom ducts of *Conus striatus*: focusing on the expression profile of conotoxins. *Biochimie* 88(2), 131-140 (2006).

Magalhães GS, Junqueira-de-Azevedo IL, Lopes-Ferreira M *et al*. Transcriptome analysis of expressed sequence tags from the venom glands of the fish *Thalassophryne nattereri. Biochimie* 88(6), 693-699 (2006).

Baek JH, Lee SH. Identification and characterization of venom proteins of two solitary wasps, *Eumenes pomiformis* and *Orancistrocerus drewseni. Toxicon* 56(4), 554-562 (2010).

Abdel-Rahman MA, Quintero-Hernandez V, Possani LD. Venom proteomic and venomous glands transcriptomic analysis of the Egyptian scorpion *Scorpion maurus palmatus* (Arachnida: Scorpionidae). *Toxicon* 74, 193-207 (2013).

Liu ZC, Zhang R, Zhao F *et al*. Venomic and transcriptomic analysis of centipede *Scolopendra subspinipes dehaani*. *J. Proteome Res* 11(12), 6197-6212 (2012).

Bouzid W, Klopp C, Verdenaud M *et al*. Profiling the venom gland transcriptome of *Tetramorium bicarinatum* (Hymenoptera: Formicidae): the first transcriptome analysis of an ant species. *Toxicon* 70, 70-81 (2013).

Margulies M, Egholm M, Altman WE *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(15), 376-380.

Liu L, Li Y, Li S *et al*. Comparison of next-generation sequencing systems. *J. Biomed and Biotech* 2012:251364. doi: 10.1155/2012/251364. Epub 2012 Jul 5

Mardis ER. Next-Generation Sequencing Platforms. *Annual Review of Anal Chem* 6, 287-303 (2013).

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74, 5463-5467 (1977).

Roeding F, Borner J, Kube M *et al*. A 454 sequencing approach for large-scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Mol Phylogen & Evol*. 53, 826-834 (2009).

Rokyta DR, Wray KP, Lemmon AR *et al*. A high-troughput venom-gland transcriptome for the Eastern diamondback rattlesnake (*Crotalus adamanteus*) and evidence for pervasive positive selection across toxin classes. *Toxicon* 57, 657-671 (2011).

Durban J, Juarez P, Angulo Y *et al*. Profiling the venom gland transcriptomes of Costa Rican snakes by 454 pyrosequencing. *BMC Genomics* 12, 259-275 (2011).

Terrat Y, Biass D, Dutertre S *et al*. High-resolution picture of a venom gland transcriptome: case study with the marine snail *Conus consors*. *Toxicon,* 59, 34-46 (2012).

Lluisma AO, Milash BA, Moore B *et al*. Novel venom peptides from the cone snail *Conus pulicarius* discovered through next-generation sequencing of its venom duct transcriptome. *Marine Genomics* 5, 43-51 (2012).

Hu H, Bandyopadhyay PD, Olivera BM *et al*. Elucidation of the molecular envenomation strategy of the cone snail *Conus geographus* through transcriptome sequencing of its venom duct. *BMC Genomics* 13, 284-296 (2012).

Rokyta DR, Lemmon AR, Margres MJ *et al*. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). *BMC Genomics* 13, 312 (2012).

Rendón-Anaya M, Delaye L, Possani LD *et al*. Global transcriptome analysis of the scorpion *Centruroides noxius*: new toxin families and evolutionary insights from an ancestral scorpion species. *PLoS One* 7(8): e43331 (2012).

von Reumont BJ, Blanke A, Richter S *et al*. 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* doi: 10. 1093 Nov 7 [Epub ahead of print] (2013).

Terrat Y, Sunagar K, Fry BG *et al*. *Atractaspis aterrima* toxins: the first insight into the molecular evolution of venom in side-stabbers. *Toxins (Basel)* 5(11): 1948-64. doi: 10.3390/toxins5111948 (2013).

Rokyta DR, Wray KP, Margres MJ. The genesis of an exceptionally lethal venom in the timber rattlesnake (*Crotalus horridus*) revealed through comparative venom-gland transcriptomics. *BMC Genomics* 14, 394 (2013).

Jin AH, Dutertre S, Kaas Q *et al*. Transcriptomic messiness in the venom duct of *Conus miles* contributes to conotoxin diversity. *Molecular & Cellular Proteomics* 12(12), 3824-3833 (2013).

Durban J, Pérez A, Sanz L *et al*. Integrated “omics” profiling indicates that miRNAs are modulators of the ontogenetic venom composition shift in the Central American rattlesnake, *Crotalus simus simus*. *BMC Genomics* 14,234 (2013).

Dutertre S, Jin A, Kaas Q *et al*. Deep venomics reveals the mechanism for expanded peptide diversity in cone snail venom. *Mol & Cell Proteomics 1*2(2): 312-29. doi: 10.1074/mcp.M112.021469. Epub 2012 Nov 14. (2013).

Sunagar K, Undheim EAD, Chan AHC *et al*. Evolution stings: the origin and diversification of scorpion toxin peptides scaffolds. *Toxins (Basel)* 5(12): 2456-87. doi: 10.3390/toxins5122456 (2013).

He Q, Duan Z, Yu Y *et al*. The venom gland transcriptome of *Latrodectus tredecimguttatus* revealed by deep sequencing and cDNA library analysis. *PLoS One* 8(11): e81357 (2013).

Wong ESW, Nicol S, Warren WC *et al*. Echidna venom gland trasncriptome provides insights into the evolution of monotreme venom. *PLoS One* 8(11): e79092 (2013).

Aird SD, Watanabe Y, Villar-Briones A *et al*. Quantitative high-throughput profiling of snake venom gland transcriptomes and proteomes (*Ovophis okinavensis* and *Protobothrops flavoviridis*). *BMC Genomics* 14: 790 (2013).

Margres MJ, Aronow K, Loyacano J *et al*. The venom-gland transcriptome of the eastern coral snake (*Micrurus fulvius*) reveals high venom complexity in the intragenomic evolution of venoms. *BMC Genomics* 14: 531 (2013).

Zhao YJ, Zeng Y, Chen L *et al*. Analysis of transcriptomes of three orb-web spider species reveals gene profiles involved in silk and toxin. *Insect Sci* doi: 10.1111/1747917.12068 (2013).

Wong ESW, Hardy MC, Wood D *et al*. SVM-based prediction of propeptide cleavage sites in the spider toxins identities toxin innovation in an Australian tarantula. *PLoS One* 8(7): e66279 (2013).

Torres AF, Huang C, Chong CM *et al*. Transcriptome analysis in venom gland of the predatory giant ant *Dinoponera quadriceps*: insights into the polypeptide toxin arsenal of Hymenopterans. *PLoS One* 9(1): e87556 (2014).

Robinson SD, Safavi-Hemami H, McIntosh LD *et al*. Diversity of conotoxin gene superfamilies in the venomous snail, *Conus victoria*. *PLoS One* 9(2): e87648 (2014).

Fry BG, Roelants K, Champagne DE *et al*. The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. *Annu Rev Genomics Hum Genet* 10, 483-511 (2009).

Mans BJ, Louw AI, Neitz AW. Biochemical perspectives on paralysis and other forms of toxicoses causes by ticks. *Parasitology* 129 Suppl, S95-111 (2004).

Steen NA, Barker SC, Alewood PF. Proteins in the saliva of the Ixodida (ticks): pharmacological features and biological significance. *Toxicon* 47, 1-20 (2006).

Schwarz A, von Reumont BM, Erhart J *et al*. *De novo Ixodes ricinus* salivary gland transcriptome analysis using two next-generation sequencing methodologies. *The FASEB j.* 27, 4745-4756 (2013).

Dobzhansky T. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher* 35, 125-129 (1973).

Abu-Asab M, Chaouchi M, Amri H. Evolutionary medicine: A meaningful connection between omics, disease, and treatment. *Proteomics Clin. Appl.* 2, 122-134 (2008).

Segovia M, Carrasco HJ, Martinez CE *et al.* Molecular epidemiologic source tracking of orally transmitted Chagas disease, Venezuela. *Emerg. Infect. Dis.* 19, 1098-1101 (2013).

Holmes EC. Molecular epidemiology and evolution of emerging infectious diseases. *Br.*

*Med. Bull.* 54, 533-543 (1998).

Ojosnegros S, Beerenwinkel N. Models of RNA virus evolution and their roles in vaccine design. *Immunome Res.* 6, S5 (2010).

Wang J, Wong ES, Whitley JC *et al.* Ancient antimicrobial peptides kill antibiotic-resistant pathogens: Australian mammals provide new options. *PLoS ONE* 6**,** e24030 (2011).

Komatsu K, Zhu S, Fushimi H *et al.* Phylogenetic analysis based on 18S rRNA gene and matK gene sequences of Panax vietnamensis and five related species. *Planta Med.* 67, 461-465 (2001).

Patterson C. *Molecules and Morphology in Evolution: Conflict Or Compromise?* Cambridge University Press, (1987).

Crowe TM. Molecules vs morphology in phylogenetics: a non-controversy. *Trans. R. Soc. South Afr.* 46, 317-334 (1988).

Brawand D, Soumillon M, Necsulea A *et al.* The evolution of gene expression levels in mammalian organs. *Nature* 478, 343-348 (2011).

Goetz F, Rosauer D, Sitar S *et al.* A genetic basis for the phenotypic differentiation between siscowet and lean lake trout (*Salvelinus namaycush*). *Mol. Ecol.* 19, 176-196 (2010).

Hittinger CT, Johnston M, Tossberg JT *et al*. Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. *Proc. Natl. Acad. Sci.* USA 200910449 (2010). doi: 10.1073/pnas.0910449107

Smith SA, Wilson NG, Goetz FE *et al.* Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* 480, 364-367 (2011).

Chiari Y, Cahais V, Galtier N *et al.* Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biol.* 10, 65 (2012).

Wen J, Xiong Z, Nie ZL *et al.* Transcriptome sequences resolve deep relationships of the grape family. *PLoS ONE* 8, e74394 (2013).

Bapteste E, Brinkmann H, Lee JA *et al.* The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium, Entamoeba*, and Mastigamoeba. *Proc. Natl. Acad. Sci.* USA 99, 1414-1419 (2002).

Hughes J, Longhorn SJ, Papadopoulou A *et al.* Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (Beetles). *Mol. Biol. Evol.* 23, 268-278 (2006).

Bourlat SJ, Juliusdottir T, Lowe CJ *et al.* Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* 444, 85-88 (2006).

Roeding F, Hagner-Holler S, Ruhberg H *et al.* EST sequencing of Onychophora and phylogenomic analysis of Metazoa. *Mol. Phylogenet. Evol.* 45, 942-951 (2007).

Sukkapan P, Jia Y, Nuchprayoon I *et al*. Phylogenetic analysis of serine proteases from Russell's viper (*Daboia russelli siamensis*) and *Agkistrodon piscivorus leucostoma* venom. *Toxicon Off. J. Int. Soc. Toxinology* 58, 168-178 (2011).

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57-63 (2009).

Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361-375 (2005).

Chan CX, Ragan MA. Next-generation phylogenomics. *Biol. Direct* 8: 3, doi:10. 1186/ 1745-6150 (2013).

*** An argumentation of how next-generation sequencing data requires new strategies of sequence alignments leading to next-generation phylogenomics.*

Lin GH, Wang K, Deng XG *et al.* Transcriptome sequencing and phylogenomic resolution within Spalacidae (Rodentia). *BMC Genomics* 15 : 32 doi: 10.1186/1471-2164 (2014).

McCormack JE, Maley JM, Hird SM *et al.* Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Mol. Phylogenet. Evol.* 62, 397-406 (2012).

Rokyta, D. R., Wray, K. P., Lemmon, A. R., Lemmon, E. M. & Caudle, S. B. A high-throughput venom-gland transcriptome for the eastern diamondback rattlesnake (*Crotalus adamanteus*) and evidence for pervasive positive selection across toxin classes. *Toxicon Off. J. Int. Soc. Toxinology* 57, 657-671 (2011).

Dutertre S, Jin AH, Vetter I *et al*. Evolution of separate predation- and defence-evoked venoms in carnivorous cone snails. *Nature Communications* 5:3521, doi: 10.1038/ncomms4521 (2014).

* *A molecular demonstration of how cone snails are able to rapidly switch between distinct venoms cocktails in response to predatory or defensive stimuli.*

Schwarz A, Cabezas-Cruz A, Kopecky J, *et al*. Understanding the evolutionary structural variability and target specificity of tick salivary Kunitz peptides using next generation transcriptome data. *BMC Evol. Biol.* 14:4, doi: 10.1186/1471-2148 (2014).

Li P, Deng W, Li T *et al*. Illumina-based *de novo* transcriptome sequencing and analysis of *Amanita exitialis* basidiocarps. *Gene* 532, 63-71 (2013).

Hartmann S, Vision TJ. Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol. Biol.* 8, 95 (2008).

Wiens JJ. Missing data and the design of phylogenetic analyses. *J. Biomed. Inform.* 39, 34-42 (2006).

Nichols R. Gene trees and species trees are not the same. *Trends Ecol. Evol.* 16, 358-364 (2001).

Caravas J. Phylogenetic utility of mitochondrial and nuclear genes: a case study in the Diptera (true flies). *Wayne State Univ. Diss.* (2012). at <http://digitalcommons.wayne.edu/oa_dissertations/4.

## Figures and captions

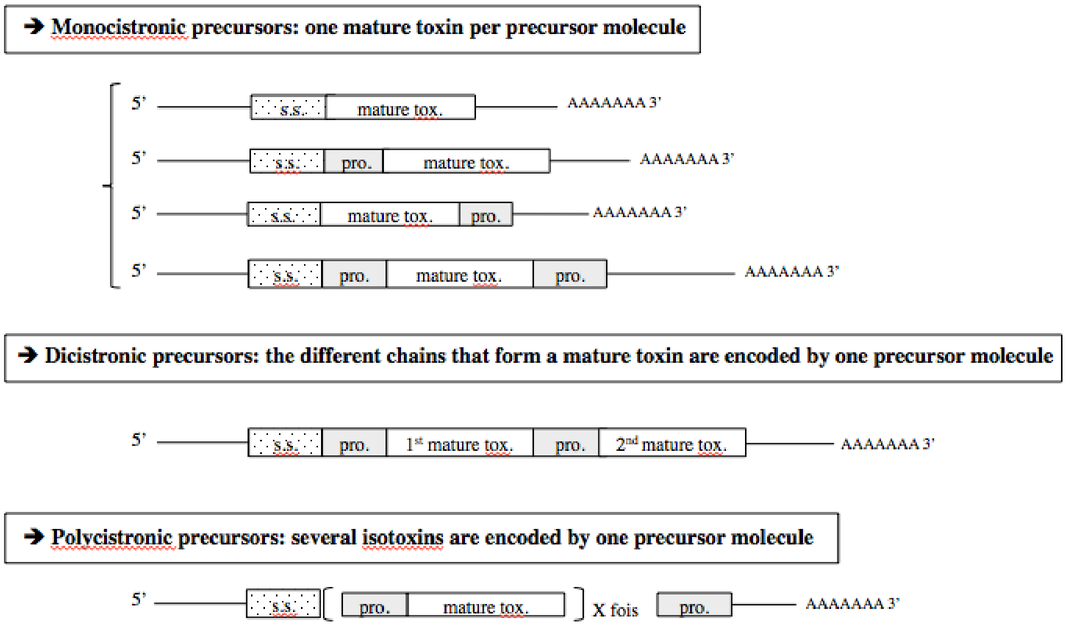
➔ Monocistronic precursors: one mature toxin per precursor molecule
5'
S.S.
mature tox.
AAAAAAA 3'
pro.
➔ Dicistronic precursors: the different chains that form a mature toxin are encoded by one precursor molecule
1st mature tox.
2nd mature tox.
➔ Polycistronic precursors: several isotoxins are encoded by one precursor molecule
X fois

**Figure 2**. *Animal toxin precursors.* Overall structures and organisations of the different mRNA precursor molecules encoding animal toxins. The signal peptide sequence at the 5' end of the ORF is indicated by "S.S". The propeptide sequences if present, are indicated by "pro" and are shown in light grey. The 5' and 3' untranslated sequences are represented as thin lines. The polyadenylated tail at the 3' end of the precursors is indicated by "AAAAAAA".

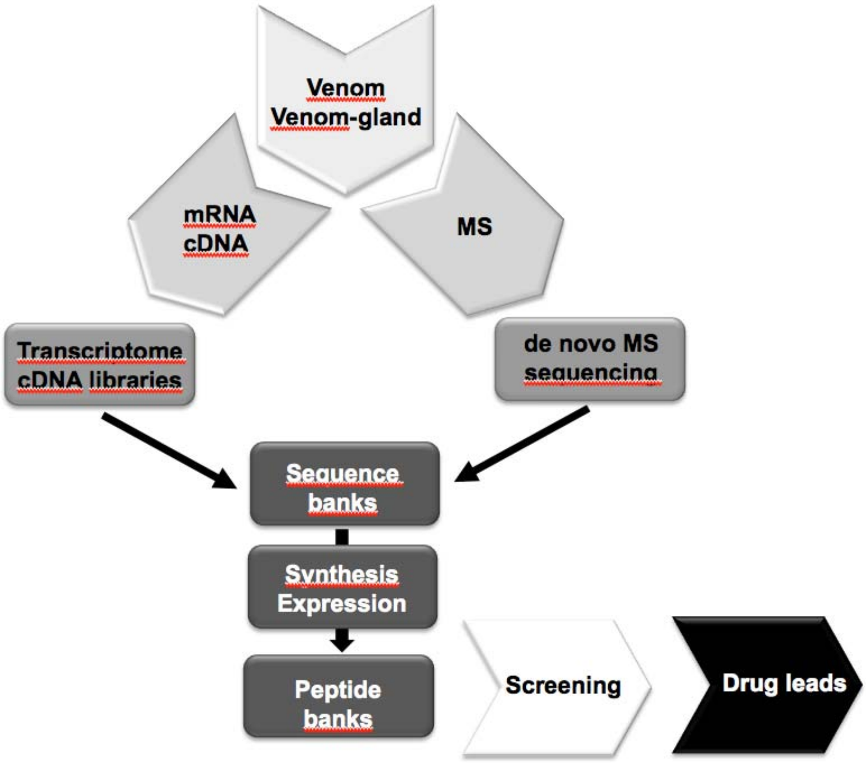

**Figure 3**. General overview of the "Omics-based high-throughput lead generation" process developed in the European VENOMICS project.

**Table 1**
Venomous animal species for which venom-gland NGS-transcriptomic analyses have been reported

| Animal species | 454 | 454 GS FLX | Illumina | Ion Torrent | Reference |
|---|---|---|---|---|---|
| *Snakes* | | | | | |
| *Crotalus adamanteus* | | X | | | Rokyta *et al.*, 2011 |
| | | X | X | | Rokyta *et al.*, 2012 |
| *Ovophis okinavensis*<br>*Protobothrops flavoviridis* | | | X | | Air *et al.*, 2013 |
| 8 Costa Rican snakes | | X | | | Durban J *et al.*, 2011 |
| *Atractaspis aterrima* | | X | | | Terrat Y *et al.*, 2013 |
| *Crotalus horidus* | | | X | | Rokyta *et al.*, 2013 |
| *Crotalus simus simus* | | X | | | Durban J *et al.*, 2013 |
| *Micrurus fluvius* | | | X | | Margres *et al.*, 2013 |
| *Cone snails* | | | | | |
| *Conus consors* | | X | | | Terrat Y *et al.*, 2012 |
| *Conus pulicarius* | X | | | | Liuisma AO *et al.*, 2012 |
| *Conus geographus* | | X | | | Hu H *et al.*, 2012 |
| *Conus miles* | | X | | | Jin AH *et al.*, 2013 |
| *Conus marmoreus* | | X | | | Dutertre S *et al.*, 2013 |
| *Conus victoriae* | X | | | | Robinson SD *et al.*, 2014 |
| *Scorpions* | | | | | |
| *Pandinus imperator* | X | | | | Roeding F *et al.*, 2009 |
| *Centruroides noxius* | | X | | | Rendón-Anaya M *et al.*, 2012 |
| 5 Australian scorpions | | X | | | Sunagar K *et al.*, 2013 |
| *Spiders* | | | | | |
| *Latrodectus tredecimguttatus* | | | X | | He Q *et al.*, 2013 |
| *Selenotypus plumipes* | | | X | | Wong ESW *et al.*, 2013 |
| *Gasteracantha arcuata*<br>*Gasteracantha hasselti*<br>*Nasoonaria sinensis* | | | X | | Zhao YJ *et al.*, 2013 |
| *Mammals* | | | | | |
| *Ornithorhynchus anatinus*<br>*Tachyglosus aculeatus* | | | X | | Wong ESW *et al.*, 2013 |
| *Crustaceans* | | | | | |
| *Speleonectes tulumensis* | | | X | | von Reumont BM *et al.*, 2013 |
| *Ants* | | | | | |
| *Dinaponera quadriceps* | | | | X | Torres AF *et al.*, 2014 |
| *Ticks* | | | | | |
| *Ixodes ricinus* | | X | X | | Schwarz A *et al.*, 2013 |