High-throughput peptidomics and transcriptomics of animal venoms for discovery of novel therapeutic peptides and innovative drug development

Reporting

Project Information

VENOMICS
Grant agreement ID: 278346

Funded under
Specific Programme "Cooperation": Health

Total cost
€ 9 360 275,25

EU contribution
€ 5 998 786,65

Coordinated by
COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES
France

Start date
1 November 2011

End date
31 October 2015

Final Report Summary - VENOMICS (High-throughput peptidomics and transcriptomics of animal venoms for discovery of novel therapeutic peptides and innovative drug development)

Executive Summary:
Context:
Animal venoms are complex chemical cocktails, comprising a wide range of biologically active reticulated
peptides that target with high selectivity and efficacy a variety of membrane receptors such as ion channels and G-Protein Coupled Receptors. Venoms can therefore be seen as large natural libraries of biologically active molecules that are continuously selected and highly refined by the evolution process, up to the point where every molecule is endowed with pharmacological properties that are highly valuable in the context of human use and drug development. Nevertheless the use of venoms for drug discovery is a rapidly emerging but still mostly unrealized prospective, due to several major difficulties including the availability of material, the sample size (most venomous animals are small to very small) and the complexity of venoms.

The aim of VENOMICS:
The vision associated with the VENOMICS project is to investigate in depth the enormous structural and pharmacological diversity of venom peptides through the development, integration and implementation of a novel research paradigm combining cutting-edge “omics” technologies in a high-throughput workflow. VENOMICS aims at replicating in vitro the diversity of venoms to generate original peptide banks to be used in drug discovery programs.

Results:
The project started with the sourcing of 323 animal samples from insects, arthropods, reptiles, fish, conus and cnidarians, representing 203 different species. 200 venom glands and 173 venoms were analyzed by transcriptomic and proteomics, respectively, in order to generate a toxin bank rich of 25,000 sequences. 3616 sequences, selected to represent the most diversity as possible in term of animals, size and structure have been produced and organized in a HTS compatible way. The Screening of these toxins against therapeutic targets related with allergies, diabetes, auto-immune diseases and inflammation allowed identifying active drug candidates.

Conclusion:
VENOMICS is a successful project which built a validated strategy to explore and exploit animal venoms in a scale never reached before. The active toxins identified at the end of the project demonstrate that the idea imagined four years ago has been transformed in a validated strategy where each bottleneck has been successfully solved. VENOMICS will have a major impact on public health issues by offering both innovative receptor-targeted drugs as well as novel therapeutic avenues for a number of current unmet medical needs. It will thus have a strong impact for European competitiveness, at both an academic and economic level.

Project Context and Objectives:
Project context
Venomous animals have developed an arsenal of small reticulated toxins used in defense and predation. Based on various disulfide-linked scaffolds, they represent an enormous structural and pharmacological diversity, from 10 to 120 residues linked by 1 to 9 disulfide bridges. Mass spectrometry and transcriptomics studies have shown the presence of up to 1,000 peptides in the venom of single species of cone snails and spiders. Therefore the global animal venom resource can be seen as a collection of more than 42,000,000 peptides and proteins of which only ~5000 are known (see figure 1 Venomous animals).

The pharmaceutical industry is in dire need of innovation since intense progress in the identification of “druggable” targets at the molecular level has not necessarily resulted in the identification of suitable
ligands. Successes in bringing to market therapeutic antibodies, proteins and peptides have increased the attractiveness of “biologics” as novel drugs, with several hundred antibodies and peptides under clinical development. Reticulated peptides from venoms are a truly novel class of biologics. They are highly target-specific but have much lower immunogenicity than antibodies and show much higher resistance to degradation than linear peptides. Peptide drugs from venoms are few but significant developments are occurring in the fields of pain, infection, cancer, hypertension and blood hemostasis control.

Considerable efforts have been developed by academic labs to study venoms. However, it has become apparent that pharmaceutical industries hesitate in using venom peptides as candidate molecules to discover novel therapeutics. This is mainly due to the large number of technical issues that represent venom manipulations. Identifying a drug lead from the hundreds of peptides found in single venom is a needle-in-a-haystack problem. The classical bioassay-guided isolation of bioactive peptides approach is time consuming, risky and not applicable to HTS.

Project objectives
VENOMICS proposes a totally new paradigm that completely bypasses the classical approach to identify novel therapeutics within venoms. Instead of isolating the bioactive peptide after several rounds of bioassay, then identify its sequence and finally develop a production protocol, VENOMICS will firstly solve the issues of sequence identification and production before screening. The VENOMICS definition of toxin sequence is a peptide of less than 120 residues and reticulated by at least one disulfide bridge. The project starts by the sourcing of venomous animals coming from various phyla, regions, sizes... Thanks to the analysis of the transcriptome of the venom glands and the proteome of venoms, sequences of all the toxins present in theirs animals will be identified (SG, Spain and ULG, Belgium). From that unique data sequence bank, the ever first synthetic toxins library will be produced by recombinant expression (NZytech, Portugal and UMED, France) and chemical synthesis (CEA, France), depending on their size and the presence of post-traductional modifications of the toxin. Finally, this toxin bank, organized in an HTS compatible way, will be screened on therapeutic targets related to unmet therapeutic needs (CEA, France and ZP, Denmark). The project is under the management of Absiskey, France (figure 2: Venomics workflow).

SOURCING: The sourcing of venoms will allow to obtain the largest venom and venom gland collection in the world. Commercial supply and field collection of venomous animals from Europe and French tropical areas will permit the collection of 200 different venomous species, most of them never explored. This biobank will cover the most possible biodiversity, including species from different sizes, ranging from ultra-small to large animals. The collection of biological material complied with ethical issues and European directives concerning protected species and conservation of biodiversity.

TRANSCRIPTOMIC ANALYSIS: The transcriptomics analysis of 200 venom glands, mainly from unsequenced species represents a huge effort. The challenge of analyzing >200 cDNA/EST libraries in a reliable manner lie in the optimized use of massive sequencing technologies. The traditional microarray high throughput approach fails in the discovery of new transcripts in an unbiased manner. VENOMICS will go further by testing and using two different second generation sequencing platforms 454 and Illumina to select the more suitable tools to generate high quality data. In the absence of genomic information, high-throughput transcriptomics will be achieved through massive sequencing of cDNA/EST libraries and
together with new bioinformatics tools allow exhaustive and accurate annotation of the generated libraries. Methods adapted for ultra-small samples should be also developed. The major issue of interpreting and annotating the huge amount of data provided by assembly could be possible thanks to the development of dedicated new-generation assembly and annotation software tools and an optimized bioinformatics analysis pipeline including appropriate quality controls.

PROTEOMIC ANALYSIS: The challenge for proteomics analysis of venoms is that it should obtain the sequences or at least part of the sequences of the toxins without any previous knowledge of reference venoms. The development of an automatic full de novo sequencing of venom peptides using mass spectrometry is a previously unsolved challenge. This can be achieved only thanks to the development and improvement of novel mass fragmentation methods, high-accuracy measurements and hyphenation to nanochromatography using the most cutting-edge mass spectrometry techniques available. These techniques need also to be applicable for the first time to ultra-small animals in a “nano-VENOMICS” approach. Development of specific interpretation software will be necessary for high-throughput and reliable automated data processing, as validated by built-in control protocols.

SEQUENCES BANK: One of the major issues to be addressed by VENOMICS will be the exploitation of these massive datasets via specialized bioinformatics and in particular the development of reliable assembly methods that will permit mature peptide sequence reconstruction from sequencing data. Automation of both proteomics and transcriptomics workflows, interpretation and integration of sequence data in a structured manner are major technological challenges that we will address to generate a bank 50,000 sequence.

PRODUCTION: Large scale parallel reticulated peptide production is a challenge never attempted since it involves both a production challenge and the additional problem of refolding disulfide-linked peptides into active forms. Selected toxins will have less than 120 residues and reticulated by 1 to 9 disulfide bridges. Several folds, structure families, will have to be produced with maybe a specific protocol for each of them. In addition, significant part of toxins will contain post-translational modifications which can be also a challenge to reproduce. The selection criteria for the production strategy will be based on toxin length. The toxins smaller than 35 residues will be produced by chemical synthesis while the longer ones will be purified from bacterial expression. After the chemical synthesis, the refolding step will be addressed with the implementation of a highly parallel refolding platform. Finally, quality control strategies will be built-in at all stages of the process, from sequencing efforts through peptide production and validation.

SCREENING: The ultimate objective is the validation of the VENOMICS strategy by functional assays aiming at identifying drug leads from the peptide library. Drug discovery efforts will focus on diseases related to inflammation, diabetes, auto-immune diseases, obesity and allergies. The choice of the targets is of prime importance and will be done just before the starting time of this task.

CONCLUSION: VENOMICS is an ambitious project that has to face technical bottlenecks never attempted before. It is clearly the biggest project on animal toxins either built in the world. VENOMICS proposes a seminal paradigm change with, for the first time, the possibility to perform high-throughput exploration and exploitation of venoms, for human benefits.
Project Results:
SOURCING
See Figure 3: Sourcing

The SOURCING team was in charge of providing VENOMICS consortium with the biological tissues (venom glands, venom ducts, salivary glands, telson...) and fluids (venom and salivary) necessary to ensure the success of the project. The CEA was in charge of that task.

Venomous animal diversity
The first prerequisite was to have access to the largest diversity of venomous animal species ever explored! Indeed, when other European or international projects focused on few different phyla of venomous animals, we aimed in VENOMICS, to provide the scientific community with the largest and more diverse libraries of venom/salivary compounds ever constituted.

To reach that goal, we combined collect expeditions together with the purchase from a private collector and a private company. Together these three different sources allowed us to gather 203 different specimens, including terrestrials, marines and flying species, distributed as illustrated in table 1 “Sourcing results”.

In the beginning, the animals providing the venom were found through expeditions in different French territories: French Guiana, Mayotte Island & Polynesia, but in the last year a company specialized in breeding poisonous animals (AlphaBiotoxine, Belgium), including some rare species, was contracted. Thus, all the 40 conus species that are part of VENOMICS collection are from Mayotte (Indian ocean) and French Polynesia (Pacific ocean) expeditions. On the other hands, half of the spiders and scorpions are from a private collector from France. Finally, all the snake specimens together with rare and peculiar species were bought from the very professional Belgium Company “AlphaBiotoxin”.

Together, and in addition to “classical” snakes, scorpions, spiders, cone-snails and cnidarians (jelly-fishes, sea-anemones), animals under analysis included, scolopender, fishes, ants, a poisonous species of lizard that had never been studied before, terebra, octopuses, and different insects including bees, wasps, bumble-bees.

Respect of natural resources
Whatever is the source of the 203 different species of venomous animals that were gathered along the VENOMICS project, we paid an important caution to respect the natural resources to which we had access. Thus, we limited the number of collected specimens to the strict number necessary to ensure the amounts and volumes of biological samples compatible with the experiments to be carried out within VENOMICS’ consortium. For instance, during Mayotte and Polynesia expeditions, supernumerary cone snails collected during the day, were systematically dropped back to water.

Another illustration of such and ethical behavior, is illustrated in the case of the very and unique specimen of venomous lizard (Helodermat horridum exasperatum), to which we had access via AlphabioToxin Company. Thus, we experimented a strategy based on the extraction by surgery under anesthesia of only one of the venomous salivary of the lizard. This was done with the help of a veterinary surgeon. The animal fully recovered and we established that the scientific results obtained with the extracted salivary gland
were of an excellent quality. This demonstrates the interest of such a strategy of tissue extraction that could, certainly be applied with success to rare snakes too!

Venomous species size limitation
One of the other challenges of VENOMICS project was also, to explore some very small venomous animals, which remain the most abundant on the Earth. Indeed, the huge majority of the spiders are few millimeters large, rendering the extraction of their venoms and venomous tissue difficult! To address that question, we included in VENOMICS some very small spiders. These animals, not only have a small size, but they are also very rare and only one or two specimens could be collected! Aware also about the quality of the material extracted, to ensure the next coming experimental phases of the project, we proceed to a speed and partial dissection of the tissues of interest. Despite that “contamination” with body tissues, we noticed that the deepness of the Omics technologies that we employed in VENOMICS (ie. transcriptomic of the venoms gland cells performed using the most recent DNA sequencing strategies) fully compensated the dilution of the generated data sequences with body elements! Combined with an appropriate strategy of annotation of the generated libraries, we have been able to identify several hundreds of toxins, among which many are totally news. Together, these recent data demonstrate that it is now possible to explore very small venomous animals, opening huge and novel perspectives in the field of drug-candidate discovery!

PROTEOMIC ANALYSIS
Proteomics aims at characterizing the proteome that is the peptide/protein content of a biological medium, such as a cell culture, a tissue extract, a plasma or... a venom. In VENOMICS, the strategies applied to the characterization of venom proteomes have been driven by the following properties of these particular samples.

A) Venom compositions vary from one species to another. Some venoms are mainly composed of small (1-5 kDa) and highly modified peptide toxins, such as cone snails or hymenopterans, whereas snakes, scorpions or spiders are known to produce larger toxins, generally spread from 5 to 10 kDa. In each case, toxins are folded by disulphide bridges that have to be reduced (removed) and alkylated (blocked) to unfold the peptides and to increase the efficacy of the later structural characterization (sequencing). Venoms containing small toxins were directly analyzed after these treatments. However, because larger toxins are harder to characterize, an additional step of sample preparation was included for venoms containing large toxins (M>5 kDa). After the reduction and the alkylation of the disulphides, such toxins were cut into small pieces through an enzymatic digestion performed under sweat conditions. The digestion was set up to generate peptides having with masses between 1 and 5 kDa, allowing these venoms to be analyzed in the same way than the previous ones.

B) Venoms are complex mixtures. Venoms are complex mixtures, made of a high number of different compounds. This number usually fluctuates from few hundreds (ex: snakes) to few thousands (ex: cone snails), depending of the considered species. Since the goal was to study the maximum of individual peptide, technical reasons prevent any direct characterization from crude venoms. The use of a powerful technique of purification is then mandatory. The technique, called ultra-performance liquid chromatography, was chosen to provide an effective separation of the toxins. Moreover, due to the poor amounts of venoms provided by the smallest VENOMICS species (bees, wasps, cone snails...), the nano-scale of this technique was selected (nano-UPLC), permitting the analysis for each venom from 500 ng,
that is two millionth of a gram.

C- Each toxin must be individually characterized. The main purpose of proteomics is to characterize the protein/peptide content of a biological sample. Characterization of toxins in VENOMICS was performed by mass spectrometry, allowing the measure of exact masses (MS) and the generation of informative data on each sequence (MS/MS experiments). The ideal mass spectrometer for this task had then to provide not only a high accuracy and a high resolution (for exact mass measurements), but also a high efficiency in fragmentation (for toxin sequencing). In few words, after nano-UPLC separation, toxins were eluted into the Q-exactive. This instrument was able to measure their exact masses and fragment the peptide with a high efficiency. Toxin sequencing was assisted by dedicated software called Peaks 6 (BSI) (see figure 4: Mass spectrometer).

This approach allows obtaining full or partial sequences for many toxins which have to be integrated with transcriptomic data.

TRANSCRIPTOMIC ANALYSIS

Transcriptomics is the study of gene expression in a cell, tissue or organ, and it’s based in RNA sequencing. Specifically, de novo transcriptomics is focused in previously unknown organisms, targeting natural huge biodiversity. This technology opens the door to the study of nature without previous knowledge of it and with a very high resolution, allowing the identification and quantification of transcripts in a broad spectrum, even with those of a low expression. The isolation of venom gland RNA, from the sourcing part, followed by Next Generation Sequencing yields hundreds sequences from each single gland. The most economical next-generation sequencing technologies are those that generate short sequence reads, typically in the range of 30–100 bp, and are the method of choice for “resequencing” model organisms (e.g. the Illumina technology). In this case, the analysis is performed by mapping the short-reads onto the reference genome or transcriptome. This approach has recently been used for transcriptome profiling in a method called RNA-seq that is expected to allow major breakthroughs in transcriptome analysis.

However, the assembly of all the sequences obtained to construct the whole transcriptome is a real challenge when there is no genome reference since de novo assemblies of sequences without a known reference using short reads have been considered difficult and researchers working on non-model organisms have often turned to the more expensive longer sequence reads (250–450 bp) obtained by the 454 Life Sciences (Roche) technology. However, the applicability of short-reads methods as an appropriate choice for de novo transcriptome assembly has recently received attention. By reassembling the transcriptome of a species with a known genome using a de novo assembler, we have shown that short-reads can be of considerable utility for assembling transcriptomes of non-model organisms.

Despite the fast development of assemblers able to efficiently handle more and more reads, transcriptome assembly is still difficult. For instance, elongation of contigs is not only impeded by repeats or allelic variations but also by alternatively spliced transcripts. Moreover, while genomic sequencing coverage is generally uniform across the genome, transcriptome coverage is highly variable, depending on gene expression level, excluding the use of coverage information to resolve repeated motifs. Therefore, the quality of a de novo transcriptome assembly is highly dependent on the user-defined sequence overlap.
length between two reads required to consider them as contiguous (referred as k-mer length). The best k-mer value for a given assembly depends on the sequencing depth, the read error rate, and the complexity of the genome/transcriptome to be assembled. For transcriptome assembly, in which coverage is not uniform, using higher k-mer length will theoretically result in a more contiguous assembly of highly expressed transcripts. On the contrary, poorly expressed transcripts will be better assembled if lower k-mer lengths are used. These theoretical expectations have been experimentally supported in a controlled de novo transcriptome assembly of a model organism. The choice of the k-mer length is then a subjective decision of whether to emphasize on transcript diversity by using a short k-mer length (that will lead to the assembly of numerous and highly fragmented transcript fragments), or to emphasize on contiguity by using a longer k-mer length (that will allow the recovery of longer transcript fragments but at the cost of a lower transcript diversity). Hence, in most cases, an intermediate k-mer length is chosen to reach a compromise between these two extremes. Therefore, an approach for de novo transcriptome assembly that takes advantage of the assembly performances of various k-mer lengths is highly desirable.

Sequencing
Since no much information was available about the reliability of different NGS platforms regarding de novo transcriptome reconstruction, Sistemas Genómicos performed different pilot studies in order to identify the most suitable NGS platform. Pilot samples were sequenced with both the 454-Roche and Illumina platform. Data analysis consisted in raw data quality control, quality filtering, read preprocessing, assembly and validation. From pilot tests analysis, SG concluded that total amount of raw data for Illumina was 200 times higher than in 454. No more than 2Mb were assembled with 454 data, many singlets (i.e. much information was probably missing) and higher number of putative toxins transcripts were found with Illumina technology rather than 454.

After platform study, Illumina pair-end NGS platform to assure sequence and data quality was selected.

Biodiversity
Small size animals represent the majority of venomous species. Venom gland extracts from animals of less than 5mm can contain more than 300 compounds. Including this huge biodiversity is one of the main challenges of the project. Low input protocols have been developed to adjust the sequencing process for very small amounts of RNA. We have worked with 50 to 200 nanograms of RNA, reducing up to 5 to 20 fold the required material previously needed.

Bioinformatics Assembly
Bioinformatic analysis was a crucial step of the process to identify peptide sequences. This analysis started with NGS raw data quality control, followed by primers, adapters, indexes and low quality bases exhaustive cleaning. Next step consisted in joining the pair-end reads and dereplicate all reads before entering in the assembly step itself. To assure the quality of the assembly, different algorithms (as De Bruijn Graph) were used, developing de novo robust and comprehensive assembly pipeline. See figure 5: Transcriptomic workflow.

Annotation
Once assembled, the different venom-gland or salivary-gland transcriptomes were annotated using different bio-informatic and specific de novo-developed tools and scripts. The objective being to constitute the largest library of toxin-like sequences, among which many will be then selected, synthesized and
To reach that goal, the totality of the assembled sequences was first annotated against public “general databases”. Then a specific “toxin annotation” was performed using sequence homology within Toxin & Venom public and homemade databases. In average, this process resulted in about 30-40 % of positive annotation, corresponding to “true toxins”: sequences that display strong and unambiguous sequence identities an/or homologies with previously known compounds issued from venom glands. Sequence alignments of these true toxins were carried out to identify the more original isoforms to be synthesized. Aside these “true toxins”, 30 to 50% of the assembled transcripts failed to match to any sequences of the actual databases! They were categorized as “unknown compounds”. No doubt that among them several precursors encoding peptides or proteins, enzymes are part of the secreted venoms or salivary! At this level, the specialists in toxins of VENOMICS consortium in close relation with our bioinformaticist, elaborated original scripts to explore that “black box” of sequences. They used the main features characterizing peptide venom-toxins:
- Presence of a signal peptide for secretion
- Presence or not of pro-peptides, that are found in many toxin precursors
- Identification and analyze of the putative mature sequences, susceptible to correspond to the peptide as found in the venom!
- Research of a high content in cysteine residues (susceptible to form stabilizing and typical disulfide-bridges)

The sequences of the toxin-candidates thus identified, were then cross-validated with proteomics data of the corresponding venoms or salivary, providing the consortium with strong indications that we are facing true de novo toxins! Some of these newly identified toxin-candidates, were selected, synthesized, and tested. Many of them display original sequences, suggesting plausible new spatial organizations, targets and mode of action. These new compounds may lead to original intellectual properties as patents.

Venomicco and the SEQUENCES BANK
Venomicco is a modular framework to manage big data complete life cycle coming from venoms analyses. It’s a complex component providing API services to develop a Web Platform that enables user to interact with the data in a transparent way using friendly interfaces. It is used for: analysis results storing, results browsing, results exploiting and results publishing. Venomicco includes a persistence manager layer to construct the data base to store data in a transparent and automatic way and is able to make management easier. It also includes a Data broker layer that handles data communication and data interoperability between components. This database contains all the results of the consortium including:
- Transcriptome assembling
- Proteomic profiles
- Peptides selected from transcriptomic and proteomic integration
- Peptides produced by chemical synthesis or recombinant expression

Venomicco also include bioinformatics tools for big data exploitation such as BLAST, MUSCLE and visualization.
VENOMICS constitutes the biggest and most completed Database worldwide including 203 venomous species, 218 whole transcript assemblies, 170 proteomic profiles to generate a bank of 25,000 sequences
of animal toxins.

PRODUCTION
Proteomics and transcriptomics analysis of almost 200 venomous species provided a fabulous bank of toxin sequences. However, primary peptide sequences provide no information on the biological activities of the venomic molecules. Between the High-throughput sequencing technologies developed in this project and the High-throughput screening strategies that already exist in pharmaceutical industries, we had to develop High-throughput production protocols to provide enough quantities of purified and active synthetic toxins to be screened. Two strategies were selected according to the size of the toxins. Peptides shorter than 35 residues have been taken in charge by the French Alternative Energies and Atomic Energy Commission to be produced by chemical synthesis. Toxins longer than 35 residues have been taken in charge by NZYTech and by Marseille University to be produced by recombinant expression. The specific complexity of toxins compare to classical proteins is that they toxins contained from 1 to 9 disulfide bridges. These bridges, that should link the right cysteine couples, are very difficult to obtain.

Chemical synthesis
The High throughput chemical production of small toxins containing or not PTMs (post-translational modifications) was undertaken in CEA. During the first stage of the project, different strategies and protocols have been successfully implemented. The Solid phase peptide synthesis (SPPS) of the toxins was optimized on an automated Prelude synthesizer (See figure 6: Prelude Synthesizer) using a fast Fmoc strategy combining large excess of protected amino acids and coupling reagents associated to short coupling times.

Two protocols of synthesis, one for very small toxins (<25 AA) and one for small toxins (25 to 50 AA) were validated allowing to obtain 36 linear toxins (<25AA) in one week with excellent purities (>75%).

After synthesis, toxins need a maturation step consisting in disulfide bridges formations. Rather than using the longer two step strategy (in red in the figure 7), we developed a one-step strategy based on the random oxidation of the crude toxins and compatible with the HT production.

Thus, a universal folding protocol was designed and validated for toxins containing 2 and 3 disulfide bridges. This was obtained after screening 60 different refolding conditions by tuning the different components (pH, buffers, additives and red-ox couples) present in the oxidation solution.

Production:
Following these optimized protocols (synthesis and refolding) the production phase of VENOMICS was initiated. 931 toxins were selected in the database resulting from transcriptomic and proteomic dataset and synthesized in order to cover the largest structural diversity. Finally, 880 toxins presented the required analytical criteria for testing, only 45 were failures and 6 were excluded due to high similarity in term of sequence.

Analysis of the population:
These toxins came from marine cones (56%), spiders (29%), scorpions (8%), then from centipedes,
snakes and others species. This chemically-synthesized bank is constituted by 44% of toxins with a size between 25-35AA, 35% between 15-25AA and 16% between 10-15AA. Most of these toxins have 3 disulfide (50%), 2 disulfide (35%) and 1 disulfide (14%) and finally only 2% of the population contain 4 disulfide.

The presence of PTMs is mainly found in two species (cones and spiders) and 33% of our bank contain PTMs with a repartition of PTMs as represented in Figure 8.

Finally, a detail analysis of 880 chemically synthetize toxins shows a huge diversity of PTMs combination in our population of toxins. If toxins with a single PTM (amidation or hydroxyproline) are predominant, our bank contains a large panel of toxins containing 2, 3, 4 and even five PTMs (see details in Figure 9).

In conclusion, we have synthesized 931 toxins and 880 toxins were put in 11 plates for testing.

Recombinant expression
1) HTP cloning
The high-throughput gene synthesis of ~ 5000 genes encoding venomic peptides was performed at NZYTech. An HTP gene synthesis platform was developed to produce synthetic genes encoding venom peptides. This pipeline includes 7 steps that allow the successful synthesis of several plates of 96 genes per week. The first step corresponds to gene design and codon optimization for maximizing expression in Escherichia coli; using NZYTech codon optimization software (ATGenium) multiple DNA sequences are designed simultaneously from peptide sequences provided by CEA. In steps 2, 3 and 4 oligonucleotides required for gene assembly are designed, synthesised and assembled through PCR using optimal conditions. Synthetic genes are cloned using NZYTech LIC protocol into the E. coli expression vector pHTP4. Bacterial transformation and DNA preparations are accomplished using high throughput protocols. DNA sequences are checked for the presence of sequence errors using the Sequencing Analysis tool. All steps are automated using a liquid handling robot (see figure 10: Synthetic production workflow).

A total of 4,992 genes (96 x 52 plates) were designed from 4,992 venom peptide sequences. The DNA sequences were optimized for expression in E. coli and the important properties of synthetic genes are shown in the table 2.

We developed 6 new expression vectors and preliminary testing with 200 genes revealed that pHTP4 expression vector is the most indicate for the production of venomic peptides and was used in the HTP cloning platform. This vector contains a fusion tag – DsbC – that promotes the disulfide bond formation in periplasm of bacterium cells. See the expression clones construction in Figure 11.

After gene synthesis and cloning into expression vector, gene integrity was verified by Sanger sequencing to detect the presence of errors in DNA sequences. From the initial sample, 3,912 genes were shown to be positive when only 1 clone was screened, for 809 genes 2 clones were required to obtain a non-error one and for 242 genes 3 clones were screened (see Figure 12).

The most common type of sequence errors identified in synthetic genes were deletions and insertions as illustrated in Figure 13.
In conclusion, NZYTech HTP gene synthesis platform was used to produce 4,992 genes encoding venomic peptides with a success rate of 1.3, meaning that 1.3 clones need to be screened to obtain a positive one.

2) High throughput Recombinant expression of oxidized toxins

A new high throughput production strategy and protocol for the production of correct reticulated toxins expressed E. coli have been successfully implemented during the VENOMICS project. Starting from clones generated by NZYtech, the production of the full library was started on month 36. The length of the toxins expressed in E. coli is ranging from 35 to 120 amino acid long. These toxins, without PTM contains between 1 to 9 disulfide bridges. These criteria made the recombinant and solid solid-phase synthesis protocols very complementary to obtain the most diverse bank of toxin possible.

To produce the vast majority of the toxins of VENOMICS, the team developed a completely new protocol. This procedure allowed the processing of 2x96 toxins every week during six months in a row. Around months 43, out of 4003 toxins checked for expression, 2360 toxins had successfully passed all the quality controls (mass spectrometry). This gave a 59% overall success for these 4003 toxins. See the proportion of toxins ranked by final concentration (4003 toxins) in Figure 14.

From the 59% success rate in toxin production, 30% of the toxins were above 20 µM concentration while only 0.67% of the toxins could not be produced due to technical problems (clone, expression or purification issues) demonstrating that this pipeline is very robust. These 4003 toxins were selected from seven organisms with two in big proportions (Spiders 45%, Scorpions 40%) and the others (cone snails 8%, snakes 2%, centipedes 1%, ants 1%, terebres 1%) in less extend. Note that if we compare the sourcing of the recombinant pipeline with the SPPS, there is a difference in the proportion of the sampling of cone snails and scorpions between the two pipelines (cone snails venoms contains mainly short toxins and scorpions a majority of long peptides). If we look closely to the sequence of the 4003 toxin selected, this batch span over 80 different “cysteine patterns” (number of disulfide bridges and cysteine position/pairing). The pipeline performed well for almost all categories. The production results are presented in Figure 15.

Toward the end of the project, the AFMB produced, following the same protocol, a second set of 960 toxins (to reach a total of 4963 toxins altogether). This batch was designed to have a chance to get more diversity of toxins from the sourcing of animals that were analyzed late.

In total, we have successfully purified 2736 out of 4963 toxins altogether (55% success rate). These toxins were aliquoted and frozen by AFMB into 96 well plates and sent to the screening partners (CEA and Zealand Pharma).

Conclusions

The teams of the NZytech/AFMB had in charge the toxin production of the VENOMICS project by recombinant methods. Both teams set up very high-throughput and efficient pipelines that allowed generating a bank of 2736 oxidized toxins that could for the vast majority be screened for pharmacological activities by the CEA and ZP.

The combination of both production methods (synthetic chemistry and recombinant expression) turned out to be as planned of paramount importance to mimic closely the diversity of the toxins found in the natural
environment. 25% of the bank was made by solid phase synthesis while 75% was produced by recombinant expression. The solid phase synthesis pipeline gave excellent success rate (95%) to produce short, 2-3 disulfide bridges with or without PTM but was limited in term of throughput to around 1000 toxins during the project and needed a refolding step after synthesis to get the toxins in oxidized form. The recombinant expression pipeline gave good success rate (55%) to produce long, 1-9 disulfide bridges without PTM directly without refolding strategies, it was not limited in term of throughput.

These protocols currently represent the quickest animal toxins production protocol in the world.

To conclude, the library of 3616 toxins produced by the VENOMICS consortium represents the biggest and most diverse collection of toxins in the world. Interestingly, this bank represents a large diversity in term of species, size, disulfide contents and PTMs.

SCREENING

This unique bank of 3616 toxins is ready now to be screened against therapeutic targets. They have been selected for their link with diseases as diabetes, obesity, inflammation, allergies. Zealand Pharma focused on functional assays, either cell-based phenotypic screening or cell-based assays on targets such as ion channels or GPCRs. French Alternative Energies and Atomic Energy Commission focused on GPCR sensitive to peptides. Screening provides primary Hits that should be confirmed by subsequent experiments. When the hit is confirmed, many other pharmacological tests have to be done to establish its affinity, selectivity and mode of action. Some time, engineering can be done to ameliorate it, this is called the hit to lead optimization. If these first results are convincing, then the lead compound is challenged in diseases models in order to be validated. Then, clinical phases can be stated.

Zealand Pharma assays

One of the phenotypic assays was the Mixed Lymphocyte Reaction (MLR), which identifies peptides that can suppress the immune response between human peripheral blood lymphocytes from two donors. Screening hits from this assay are candidates for development into drugs that can be used to treat autoimmune and inflammatory diseases.

Five screening assays were set up and optimized, then used to screen the VENOMICS toxin library. A number of hits were identified. Following conclusion of the VENOMICS project, Zealand Pharma will continue working with these hits in order to develop them into drug candidates. This will involve hit confirmation, then modification of the toxin peptides to fine tune their characteristics for treatment of the relevant diseases.

Production of the toxin library was delayed due to difficulties encountered in that part of the project plan. This meant that there was limited time to carry out the screens and to do the follow up work on the hits. However, the high quality of the VENOMICS library meant that the screening went very smoothly.

French Alternative Energies and Atomic Energy Commission (CEA) assays:

The screening strategy developed for four GPCR sensitive to peptides are competitions of radioactive ligands. That technology is very robust, quick and reliable.

This figure 16 illustrates how a screening by competitive tests is performed and where are the hits. This figure 17 shows how hits are confirmed or not. 10,400 screening have been performed which allow us to identified 318 hits. Out of these 318 hits, 88%
could be confirmed. These two results are incredible high. Classically, when a synthetic bank of molecule, design randomly by chemists is screened on GPCR, the hit rate is less than 0.04% with a success rate close to 50%. Here, the screening done on natural library, design by the nature since millions of years provide us a hit rate of 3% with a success rate of 88%.

The results are summarized in Table 3.

These results demonstrate that the VENOMICS peptide bank is 100 times more efficient than the synthetic bank screen by the industry since tens of years.

Work is pursuing to terminate the screening and the validation part, then the best candidates will be selected to continue the drug development process.

CONCLUSION

VENOMICS is the biggest projects in toxinology dedicated to the identification of drug candidates with more than 300 toxins identified as drug candidates. VENOMICS succeed to develop a unique vision of how to explore and exploit animal venoms for the benefit of human health. Snakes, scorpions, spiders, bee and so on are not only dangerous animals. They are also, for each of them, providers of toxins that, when correctly used can become the next innovative drugs.

Nature is still the first source of therapeutic innovation and VENOMICS is currently the best strategy adapted for venoms, especially because most of the venomous animals are small, rare and difficult to study. We demonstrated for the first time that a natural bank of toxins can be 100 times more efficient than the traditional synthetic ones used by the pharmacological industries to identified new molecules.

The VENOMICS project will be continued through three different ways. The first axis will evaluate the biological properties of each toxin active on the therapeutic targets screened during the project. Affinities, selectivity and mode of action will be firstly determined. For the toxins having the best pharmacological properties, experiment on models of the targeted diseases will confirmed their effects. Then for the best candidate, pre-clinical and clinical studies will be initiated.

The second axis will interest the commercial exploitation of VENOMICS. VENOMICS possesses a sequence bank rich of 25,000 sequences and a toxin bank, rich of 3616 toxins. These two banks are now available for private companies to perform their own screening. Some companies, like Sanofi, already expressed a certain interest to use the VENOMICS technology.

The third and probably not the last axis is purely scientific. VENOMICS accumulated a considerable amount of data from the transcriptomic and proteomic analysis of almost 200 venomous animals, most of them never studied before... More than 1,000 sequences have been identified by animals and this unique database represents an opportunity to study unknown proteins like enzymes or to study the fabulous but mostly mysterious evolutionary process of venomous animals.

Potential Impact:
VENOMICS has successfully achieved its objectives which are summarized below:

- 203 animal species and their venoms collected: spider, scorpion, insect such as bees, bumble bees, ant, wasp, cone shell, snakes, lizard, octopus ...
- Constitution of a database of 25,000 sequences
Constitution of a 3,616 venom toxins bank displayed in 48x96-wells plates
- 8 therapeutic targets representing unmet-therapeutic targets tested;
- More than 300 “primary hits” observed and 88% further confirmed meaning that the toxins had an effective desired action on the therapeutic targets. The hits are potential starting points of further drug development.

These factual tangible project results are the first steps towards wider scientific and socio-economic impact which are presented in this section.

The first immediate “Scientific” impact is the validation of the VENOMICS concept:

“VENOMICS has proven that a large and diverse venom toxins database and peptide bank is feasible and has a great potential for drug lead compound identifications.”
Nicolas Gilles, CEA, Project coordinator.

VENOMICS has a great potential in the context of drug discovery. The development of a new drug is a long and expensive process usually taking more than 12 years for a global cost of more than 1b€ per new drug launch on the market. Drug discovery is the first process towards new drug delivery whereby a drug candidate or lead compound is identified and partially validated for the treatment of a specific disease. The drug discovery phases last usually 2 to 5 years with the following steps:
- Target identification: This is the therapeutic target identification derived from the disease mechanism identification, performed upstream of VENOMICS.
- Lead identification: This is where VENOMICS offer a solution.
- Lead optimization: This is where nature provides already optimized leads.

Following the Drug discovery, the Drug development process can be carried on, with pre-clinical and clinical development phases, up to regulatory approval and marketing sales. The Drug development last usually 6 to 15 years.

VENOMICS is a solution for the Hits identification phase of the Drug discovery.

The high potential for drug lead identification is demonstrated by the results obtained with a small number of animal toxins used during the therapeutic targets screening performed during the project. Only 3,616 toxins were used and more than 300 hits were identified on 75% of the therapeutic targets screened.

The VENOMICS toxins bank is therefore a powerful tool for drug identification which can be directly used by stakeholders.

The drug development global market has been historically dominated by big pharmaceutical industries mainly based in the US (Pfizer, Abbott, Merck...) and EU (Sanofi, Roche, Novartis, Astra Zeneca...) and Biotech companies (smaller companies focused on R&D). The whole Drug development R&D is a huge sector gathering more than 100 000 employees in the EU (+52% between 1990 and 2014 according to EFPIA). This sector is facing a major issue: the cost of R&D per new approved medicine is increasing exponentially since 1960 and this trend is not sustainable. The number of drugs invented per billion dollars of R&D invested has been cut in half every nine years for half a century (from 50 drugs/b$ in the 1950’s to
less than 1 drug/b$ in the 2010’s). This trend is due to several factors among them:
- Increasing costs;
- Reduced clinical success
- Long cycle duration;
- Cautious regulatory environment;
- New health challenges related to older people (chronic diseases vs acute disease).

This development cost increase is correlated with a loss of revenue for big pharmaceutical industry due to the Patent expiration of blockbuster drugs known as the ‘Patent cliff’. This patent expiration started in 2008 led to $b loss of revenue (- $ 33b in US in 2012).

In this context drug development stakeholders are eagerly looking for new lead compound with a proven pharmacological activity against therapeutic targets. Toxins present in venoms are ideal candidates. Toxins are molecules of intermediate size (Peptides) which have interesting properties for a drug:
- They are naturally occurring biologics: safer than synthetic drugs and have a greater efficacy, selectivity and specificity.
- Peptides possess bioactivities that are of major interest for drug discovery as peptides and their derivatives control and coordinate most physiological processes. In contrast to synthetic substances, peptides are degraded into their component amino acids without leading to toxic metabolites.
- They are much more stable than peptide thanks to their high contain of disulfide bridges.
- Bigger than small molecules produced by chemical engineering which suffer from reduced target selectivity that often ultimately manifests in human side-effects.
- Smaller than Protein therapeutics which are specific for their targets due to many more interactions with them, but this comes at a cost of low bioavailability, poor membrane permeability, metabolic instability and most importantly, immunogenicity.

Big pharmaceutical companies are now more and more looking for outsourced solutions in order to reduce the internal costs and use innovative tools and companies for drug development. This is also a good opportunity for VENOMICS results exploitation. VENOMICS database can be interesting for both actors: Big pharma or outsourced biotech companies. VENOMICS already received expression of interest of these actors.

VENOMICS database and peptide bank exploitation will have an impact on the EU pharmaceutical sector. The EU pharmaceutical economy (700 000 employees) has still a world leadership position but is facing emerging competitors such as China. EU is still at the origin of more than 30% of the new drugs launched on the world market each year but this proportion is decreasing with the rise of “pharmerging” economies (- 10% since 1990).

VENOMICS has the potential to strengthen the EU pharmaceutical sector competitiveness by offering new drug development solution:
- By facilitating the drug discovery and providing at will cost effective toxins database;
- By providing totally new toxins, never tested before, with a high pharmacological activity.

Health
Peptide drugs have proven to be efficient drugs. Today, there are more than 50 peptide drugs that have been approved for clinical use and the increasing number of peptides entering clinical trials now supports the notion that peptide drugs have a long and secure future. The targeted therapeutic areas of the present peptides include but are not limited to oncology, metabolic, cardiovascular and infectious diseases, all of which represent important markets.

For Zealand Pharma and the French Alternative Energies and Atomic Energy Commission the VENOMICS project generated a library of 3,616 novel venom peptides that both partners can use in drug discovery efforts. Screening of the library has already been carried out on several targets. For one of these, peptides were identified that inhibit the target with impressive potency. Zealand Pharma and the French Alternative Energies and Atomic Energy Commission are now analyzing these venom peptide hits to determine which of them can be used as a starting point for developing a new drug candidate to be taken into clinical trials. Sufficient quantities of the library remain available to carry out screening on approximately ten additional targets during the coming years, and will be very useful in projects that are in need of new functional peptides that can be optimized for their therapeutic properties.

Biodiversity

The impacts of VENOMICS on the biodiversity are related to the development of a drug discovery technology able to use very small amount of venom and thus very small animals. Before VENOMICS strategy, big amounts of venoms were required because the process followed iterative fractionation of venoms. Thus, only the large animals with large amount of venoms could be explored and exploited. This strategy didn’t target the largest part of venomous species:

- 90% of the animal biodiversity is made of small animals (< 1 cm);
- Only 5% of spiders are Tarantula (big ones > 10 cm), the vast majority of spider have a size below 3 cm.
- Hymenoptera (wasp, bee, ants, etc) are the most important order among insect with more than 230 000 species. They are the most frequent venomous animal among insect and their size range from 0.1 mm to 10 cm.

The 203 Venomous animal species used during VENOMICS have been collected on French territories or purchased to specialized companies. Most of them were small animal species. The expected impacts of VENOMICS were:

- To improve the capacity to use venoms from very small animal and thus explore new venoms from a wide range of small animal species.
- To improve the capacity to use very few number of animals, mostly one and thus minimize the sourcing and the impact on the environment.
- To synthetize and produce toxins in laboratory with no more needs of the natural product extracted from the animal.

All these impacts are reached at the end of the project with the following key figures:

- The smallest animal successfully used were two different ants providing enough material to process the VENOMICS technology and include the venom toxins in the database.
- Few microliters of venom were enough to perform its analysis and include its toxins in the database. Ants’ venoms were analysed with less than 5 microliters.
- 10 fold reduction of the amount of Venom required for analysis compare to pre- VENOMICS technologies.
- The number of Venom transcriptome (the complete analysis of the Venom proteins) produced during VENOMICS (203) is 10 time higher than the number of transcriptome produce before VENOMICS worldwide in 10 years (~20).
- Venom extraction of a rare Varan (monitor lizard) has been performed under general anesthesia in order to save the animal which recovered perfectly. This strategy could be used on rare animals.

Science/Technology and other impacts
All the results mentioned above have been achieved thanks to the scientific and technological innovation developed during VENOMICS. The natural origin of the toxins, their a priori unknown size and conformation were a scientific obstacle towards the development of high throughput technologies for their characterization (proteomic, transcriptomic, and sequencing), production and screening. Each project partner has improved its field of expertise in sourcing, venom extraction, “omic” technologies, production and screening. The addition of the individual competencies has been combined in a unique workflow with unmet performances:
• CEA improved 10-fold its toxin synthesis capacity and reduce 10-fold its screening timing. These technological progresses open new future scientific perspectives to deeply study venoms, to engineering toxins of interest and to make new collaborations around the world.
• Sistemas Genomicos has developed Venomicco, a specific software handling all the data produced during VENOMICS. SG has developed de novo transcriptomics technology within VENOMICS project. This new technology will allow the company to offer targeted services to pharma industry and opening market diversification. New strategic lines risen from this project for SG will be:
  o Identification of novel peptides, sRNAs and enzymes sequences with industrial application
  o Application of NGS in drug discovery processes reducing complexity, time and costs of the process for pharma companies.
  o Mechanism of action characterization for active compounds.
  o Modular bioinformatics platform to manage the complete life cycle of big data analyses.
• UIG developed a specific process for the handling of venomous samples avoiding multiple fractionation and able to analyze the toxins with just a few microliters of venom.
• NZytech adapted its molecular biology strategy by developing an HTP cloning and synthetic gene pipeline that is now used by all the other projects of the company.
• AMU-AFMB has developed a new HTP protein production pipeline that is hundred times quicker than any pipeline used to produce toxins before VENOMICS. Not only is this protocol quick to express toxins but indeed this procedure is for any kind of proteins the quickest recombinant protein production protocol ever described in the world. It is universal and was already successfully used in the team on hundreds of proteins outside of the VENOMICS project.
• On top of being very HTP, it offers the unique opportunity to obtain toxins oxidized from E. coli cultures, a big problem for most of the therapeutic protein expressed in E.coli.
• ABSISKEY was in charge of the overall project monitoring and reporting. VENOMICS was an opportunity to develop a network of high end scientist in biotech and pharma activities. The promising results of VENOMICS deserve to be proposed to the European Community for further developments through, for example, the Horizon 2020 funding program.

Dissemination
VENOMICS project has already produced a number of scientific and public communications that are listed
in the project website: www.VENOMICS-project.eu several more scientific publications are currently in preparation.

The project brochure is in the annexes of this report.

The dissemination key-figures are the following:
- 7 Scientific papers published in peer-reviewed journals;
- 8 Scientific reports at international congress
- 3 press event organized during the project in Lisboa, Paris and Valencia;
- 5 Press agencies dissemination activities;
- 11 publications in national newspapers (El Pais, Le Monde, Publico, Biofutur etc.);
- 11 TV and radio programs (the latest one published by Euronews is available online: http://www.euronews.com/2016/01/18/how-i-learned-to-stop-worrying-about-snakes-and-love-the-venom/)
- 124 publications in digital press (the most renowned net version of European newspapers published news about VENOMICS).

Since always, nature is seen as a prolific source of medicines. This is also the case for venoms, firstly use as a whole by ancient Egypt or by the traditional chines medicine. Later one, toxins were isolated, biologically characterized and use as drugs. Most of the toxins on the market have not been chemically modified as they were already very efficient. The claim raison is that toxins are highly refined by the evolution process, up to the point where every molecule is endowed with pharmacological properties. For the first time, VENOMICS project is able to compare efficiencies of a natural bank of toxins with the one of chemical libraries, screened by the pharmaceutical industries since tens of years. The screening tests performed at the CEA have demonstrated that the VENOMICS toxins bank provided tens of hits which have to be studied.

These numbers show that the efficacy of the VENOMICS strategy is hundred times more elevated than a randomized and synthetic chemical bank. By extension, the VENOMICS bank of toxins represents indeed a synthetic bank of 361,600 molecules but much more cheaply to screen.

Concluding SWOT for the VENOMICS results market exploitation

Strength: adapted to any sample size of the venomous animal. First strategy adapted to European venomous animals. Provide the first natural bank of animal toxins 100 times more efficient than classical organic banks.

Weakness: expensive process. Need a large panel of technics and apparatus

Threat: VENOMICS is a prototype, which deserves to be transformed in a commercial adventure. This professionalization will reduce dramatically the cost of this technology.

Opportunity: make toxins a real interest for the pharma by renders them compatible with pharma criteria’s. Promoting the vision that animal toxins are true innovative drugs.

Potential impact per partner are summarized in Table 4.