Tox|Note Manual

This document contains practical information on the use of **Tox/Note**, an online software package for analysis and annotation of spider venom-gland transcriptomes. **Tox|Note** was developed by Sandy S. Pineda, Pierre-Alain Chaumeil, Anne Kunert and Quentin Kass.

To use Tox|Blast, register by sending an email to: support@arachnoserver.org

If you loved or find Tox|Note useful, please send your comments, suggestions, bugs and reports to: support@arachnoserver.org... We look forward to hear from you.

Table of contents

About Tox|Note

Tox|Note — the spider toxin annotation and evaluation facility — is a bioinformatic pipeline that aims to fast-track the analysis of venom-gland transcriptomes generated by next-generation sequencing projects. The Tox|Note pipeline is designed with a user-friendly interface so that users can achieve the following tasks with minimal manual input:

• annotation of toxin transcripts **(Tox|Blast/Tox_Seek|)**

• prediction of signal peptide and propeptide cleavage sites in full-length spider-toxin precursors **(Spider|ProHMM**)

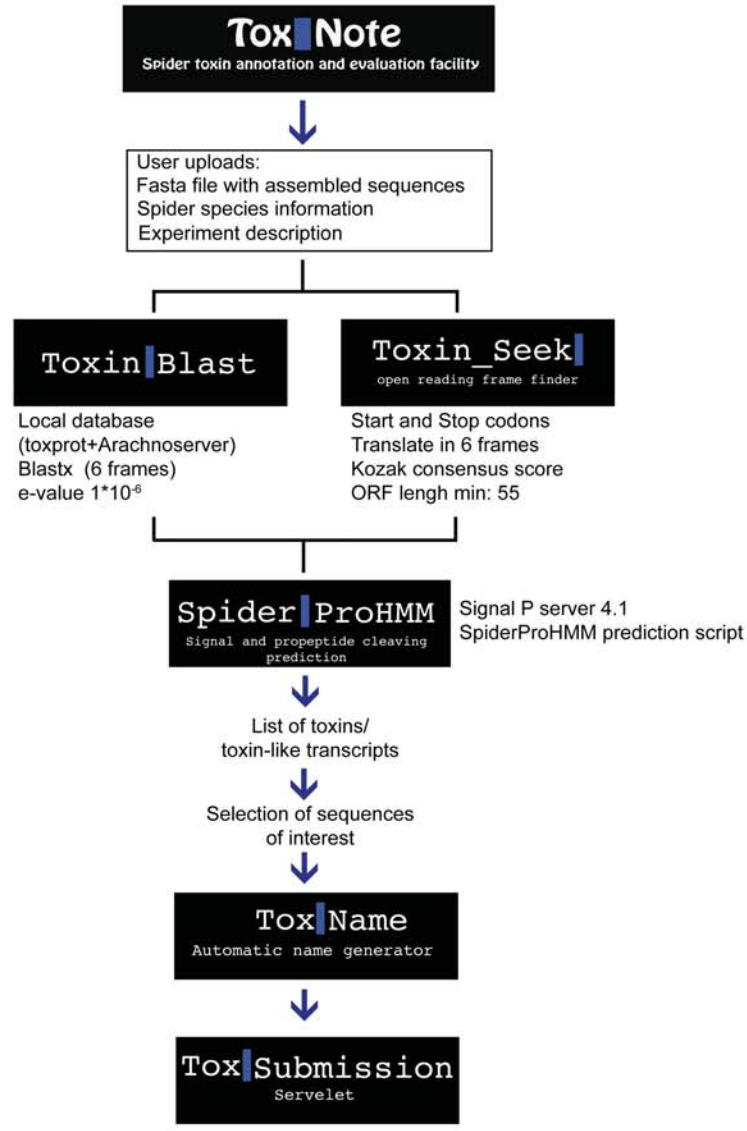
• automatic generation of rational toxin names **(Tox|Name**) based on published nomenclature rules (King et al., 2008).

Tox|Note also incorporates tools developed by the European Molecular Biology Laboratory (EMBL), the European Bioinformatics Institute (EBI), and the Australian Bioinformatics Resource (EMBL-ABR) (under the **Tox|Submission** tab). These resources make submission of sequencing project data to the European Nucleotide Archive (ENA) as simple as possible, offering the convenience of using just one website to achieve the isolation, annotation, and submission of sequences. Once submission is completed, ENA will review the submitted information and create unique accession numbers for projects and samples. These accession numbers will be sent directly to the user and subsequently fed back into the UniProt and ArachnoServer databases. These steps trigger an automated response from ArachnoServer that generates all of the required toxin cards, minimizing the amount of manual curation required for this database.

Tox|Note also provides other analysis capabilities, such as **Tox|Pred** and **Tox|Match**. In conjunction, these tools enable calculation of predicted molecular masses from mature peptide sequences isolated from the transcriptome and comparison of these masses with mass lists generated from proteomic and transcriptomic experiments. A complete diagram depicting the various tools and their inter- relationships is shown in Figure 1.

A

B



C

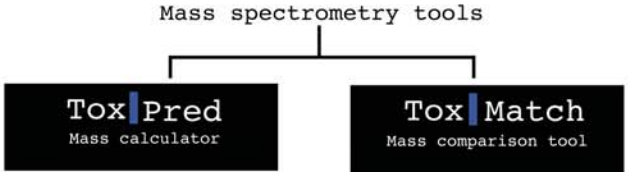


Figure 1: *(A) Screenshot from ArachnoServer 3.0 highlighting the Tox|Note tab (blue circle); (B) Schematic of the Tox|Note pipeline showing the integration of scripts and tools; (C) Overview of mass spectrometry tools.*

Shortcut to Tox|Note usage

Tox|Blast
This tool requires the following input files: a FASTA file containing the contig/singlet consensus sequences produced by any assembly software (e.g., Newbler, MIRA, VELVET/OASES, Trinity, etc). **Please note that the use of symbols in the headers and their length** are important features and may create errors while parsing the input file. We recommend avoidance of long descriptions and symbols in uploaded FASTA files, especially the use of pipes in headers. **If your header looks like the following example: >TR1|c0_g1_i1 len=509 path=[512:0-391 513:392-415 514:416-508] [-1, 512, 513, 514, -2], please make sure you cut the description and remove the pipe symbol, so the header would be: >TR1_c0_g1_i1**

Tox|Name
This tool requires the following input: IDs from the **Tox|Blast** csv file.

Tox|Submission
This section of the pipeline requires: (i) relevant information from the sequencing experiment (see **Tox|Submission** checklist); and (ii) relevant BAM and annotation files (**Tox|Name** file).

Tox|Pred
This tool calculates the theoretical mass of any mature peptide given in a FASTA file. Users can paste individual peptides in FASTA format or upload a FASTA file.

Tox|Match
This tool compares the experimental and theoretical masses from proteomic and transcriptomic experiments. Upload your text files of interest, set a match tolerance, and execute. Please note that **Tox|Match** assumes that the mass spectrometry masses do not require any adjustments (e.g., carboxyamidomethylation) as this is not supported at the current time.

Spider|ProHMM (standalone feature)
This tool predicts the signal peptide cleavage site using SignalP server 4.1 and HMMER 3.0 in conjunction with an algorithm developed in-house to discriminate putative propeptide cleavage sites. **Spider|ProHMM** requires sequences in FASTA format; these can either be pasted into the box provided or a FASTA file can be uploaded. Please note that the same rules of header length and symbols apply to **Spider|ProHMM**; we recommend avoidance of long descriptions and symbols in uploaded FASTA files.

Tox|Note in detail

Tox|Blast

To begin using **Tox|Note**, go to the main page of the ArachnoServer database (www.arachnoserver.org) and click on the **Tox|Note** tab. Alternatively, you can bookmark the **Tox/Note** page (http://arachnoserver.org/toxNoteMainMenu.html) for direct access (see screenshot in Fig. 2). If you have not used **Tox/Note** previously then you will first need to register to obtain a username and password, sending an email to support@arachnoserver.org.

First you need to upload a FASTA file (fa and/or fna also accepted)[1] from an **ASSEMBLED**[2] data set produced using any assembly software available (e.g., MIRA, VELVET/OASES, TRINITY etc.). It is important at this stage to **double check the headers in your FASTA file**, as very long headers with symbols can make **Tox|Note** crash. **If your header looks like the following example: >TR1|c0_g1_i1 len=509 path=[512:0-391 513:392-415 514:416-508] [-1, 512, 513, 514, -2], MAKE sure you cut the description and remove the pipe symbol, so the header would be: >TR1_c0_g1_i1**

To upload your FASTA file, click on the **Choose File** button and navigate to the location of the file of the file on your computer. Then do the following information: (i) add a name for the data set; (ii) add a description of the experiment; (iii) select the spider species name from the dropdown menu (if the species name is not in the dropdown menu then please send us an email message so we can add it to the list). Once all the boxes are filled in, press the upload button below the species name to upload your FASTA file.

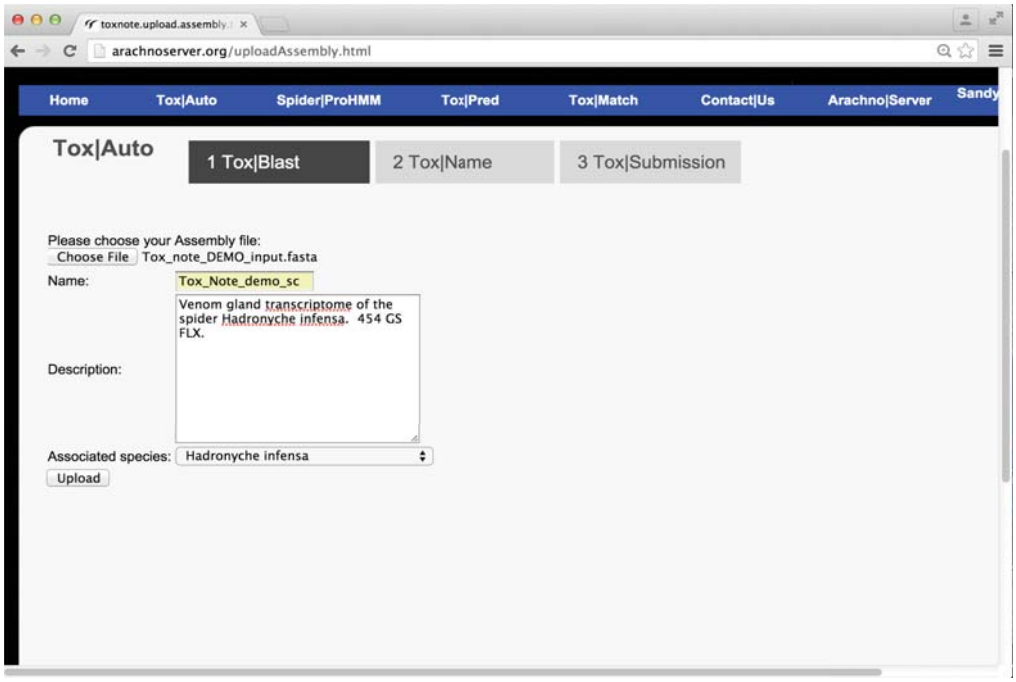


***Figure 2**: Screenshot showing fields shown in the **Tox|Blast** tab. The figure highlights how to upload a new FASTA file to the **Tox|Note** pipeline.*

[1]Note that the FastQ file format is not supported.

[2]The maximum number of contigs/singlets is 500,000, which should cover most projects. Please contact us if you want to analyse more than 500,000 sequences. If 454 or Sanger sequences are being used, it is possible to perform the analysis using unassembled sequences; just be aware of the maximum number of sequences that can be analysed.

After you hit the upload button, be patient and allow the hard-working spider on your screen to do all the work for you. After the analysis is concluded, an email will be sent to your account to alert you that the analysis is finished. As a guide, **Tox|Note** takes approximately 3–4 hours to analyse test sets of ~35,000 contigs/singlets.

Tox|Note performs a BLAST search of all sequences recovered from the transcriptome data against a consolidated databank containing all entries in the ArachnoServer and UniProt VenomZone (http://venomzone.expasy.org) databases. Sequences are translated in six frames and BLAST searches performed using Blastx (Blast+). Sequences returning a hit with a minimum Expect Value (E) of 1×10^{-6} and full open reading frame are recovered and reported by **Tox|Note** as toxins with hits to the database.

Blast+ note: In some instances, the local alignment between query and subject produced by blastx will not extend to both 5' and 3' ends of the ORF. Take the following example with nucleotide sequence ex1_nuc:

```
>ex1_nuc
ACACTTTGGGGTTCATTTTTTATTGCTATATGGAAATAAATTCTTCTCCAGTACAGACTGATATTTTACAGA
AACATTTCTTGTGCTTCAGTCATTCAAATTTTGAAAATCCTTGGACTCAAGTTTTATTTTTCAGTTAGCCCC
AACACTGGAACGCACTGAAACCATTTACTACGACATTTATTCCTTTTCCGGCTGTGTTTAGAGCTCCGTTGT
CACTAGTTTTATCGCTGCATCTACTGTACGGCTTATCACCTACGTAGACAGTATCGCAGGCAACAGTTGCAC
CGCAGCATTCGCAATGATAATCGCATTGCTCCCCTAATTTCTTGCTGCATTGCTCAGCTCTTTTTCGAAGAG
GCGCTTCTTCCTGATATTGCACCAGATCTTCAACTAACTGATCTCTTTCGGCAACAATTTCTTCCTTGGCCT
TTGAAGGCAAGGCTACAGCAACGGTGACCAAAACTGTGATCACGACGAAGGAATGCAGCAGCTTCATTATGA
AAACAATTACCGCAGTCAGGGGAGTACTCGGTTGGTATGCTAGTTCTGGTGCGTCCTGTGGAGTAGT
```

This nucleotide sequence produces the following ORF in Frame –1: ex1_prot, where M is the start codon and the peptide ends with a G residue.

```
>ex1_prot
MKLLHSFVVITVLVTVAVALPSKAKEEIVAERDQLVEDLVQYQEEAPLRKRAEQCSKKLGEQCDYHCECCGA
TVACDTVYVGDKPYSRCSDKTSDNGALNTAGKGINVVVNGFSAFQCWG
```

A blastx search using the >ex1_nuc sequence, using default settings, returns the following alignment against a sequence from the spider *Pelinobius muticus*:

putative mature sequence toxin-like ACSKQ [Pelinobius muticus]
Sequence ID: gb|ADF28499.1|**Length:** 115 **Number of Matches:** 1
Range 1: 26 to 115

Score	Expect	Method	Identities	Positives	Gaps	Frame
95.1 bits(235)	4e-21	Compositional matrix adjust.	42/91(46%)	60/91(65%)	1/91(1%)	-1

```
Query  412  AERDQLVEDLVQYQEEAPLRKRAEQCSKKLGEQCDYHCECCGATVACDTVYVGDKPYSRC  233
            +E  +L+E L   +E  P ++ A  CSK++GE+C++ C+CCGATV C T+YVG     +C
Sbjct  26   SETSKLLEKLGVSREAIP-QEMARACSKQIGEKCEHDCQCCGATVVCGTIYVGGNAVEQC  84

Query  232  SDKTSDNGALNTAGKGINVVVNGFSAFQCWG  140
              KTS+N  LNT G G+N V N F++  CWG
Sbjct  85   MSKTSNNAVLNTMGHGMNAVQNAFTSVMCWG  115
```

In this particular example, the alignment of query and subject does not extend to the 5' end of the transcript, missing the first 29 residues of the putative ORF. Normally, the BLAST information contained in this example would be sufficient to call this a significant hit. However, due to **Tox|Note**'s downstream processing, this incomplete extension violates the rules set in place to avoid retrieving information from incomplete transcripts.

Due to the nature of local alignments, and after consultations with user services at NCBI, we follow their advice in setting the composition-based stats to 0. Please note that this amendment may or may not fix the issue of reaching the end of the N- and C-termini of the peptide. Moreover, in cases where this composition-based stats setting fails to extend to either side of the transcript, the developers have chosen to recover any relevant information from BLAST (i.e., accession number of the hit) and append it to the last column of the **Tox|Blast** csv output file as an interim measure while we continue to test other scripts and software.

Not surprisingly, broken alignments also affect the retrieval of ORF coordinates. In this case, to avoid using shorter than expected coordinates, the **Tox|Blast** script was adjusted to minimise information loss. However, there will be certain instances (i.e., where BLAST hits are the only ones retrieved) that the user should be aware of that cause the coordinates to be shorter than expected.

In parallel to **Tox|Blast,** the entire data set is automatically loaded to the **Tox_Seek**| open reading frame finder. This is a complementary step to the BLAST search that aims to predict open reading frames (ORFs) *ab initio* from contigs and singlets.

The **Tox_Seek|** tool searches the entire dataset for complete ORFs (i.e., transcripts with start and stop codons and a minimum length of 55 residues). **Toxin_Seek|** also performs a Kozak consensus sequence analysis (Kozak, 1987) to allow computation of a Kozak score for each putative ORF. This enables **Tox_Seek|** to rank the starting ATG sites and pick the most likely ORF from the list of candidates.

The **Tox_Seek|** algorithm attempts to predict the majority of ORF with no hits to the database, including potentially novel "toxin-like" transcripts, and ensures that these ORFs are recorded and appended automatically to the **Tox|Blast** output. Once **Tox|Blast** and **Tox_Seek|** complete their iterations, the preliminary output comprises a list of candidate sequences that are then submitted to **Spider|ProHMM** for prediction of signal peptide and propeptide cleavage sites. This tool submits all sequences in the **Tox|Blast file** to the SignalP sever (SignalP 4.1[3] is used in this version of **Spider|ProHMM**) (Petersen et al., 2011) to predict the signal peptide cleavage site. After the signal sequence information is acquired, the signal sequence is trimmed from the toxin-precursor, and the remainder of the sequence is uploaded to **Spider|ProHMM** for prediction of propeptide cleavage sites.

Spider|ProHMM uses HMMER 3.0 (Eddy, 2011) to discriminate putative propeptide cleavage sites. **Spider|ProHMM** is a variation of the script described in 2013 (Wong et al., 2013) with the following adjustments: (1) in addition to identification of cysteine-rich regions, **Spider|ProHMM** now also considers the length of the inter-cysteine segments; (2) **Spider|ProHMM** selects the cleavage site furthest towards the C-terminus instead of the cleavage site with the lowest e-value. The new **Spider|ProHMM** script requires only a fraction of a second to analyse each sequence, making it amenable to analysis of large transcriptomic datasets and reducing the time it takes for **Tox|Note** to run. Once **Spider|ProHMM** completes its analysis it outputs the predicted signal peptide, propeptide, and mature toxin sequences for each transcript.

[3]Please note that the predictions may vary between the different versions of SignalP. Our team has tested all available versions of SignalP and we have the results available upon request. Because **Spider|ProHMM** using prediction algorithms, we cannot assure 100% certainty with respect to the predicted signal peptide and propeptide sequences.

These three steps in conjunction (**Tox|Blast**, **Tox_seek|** and **Spider|ProHMM**) lead to output of a .csv file that contains all contigs/singlets with a hit to known toxins along with toxins and putative toxin transcripts predicted by **Tox_Seek|**[4].

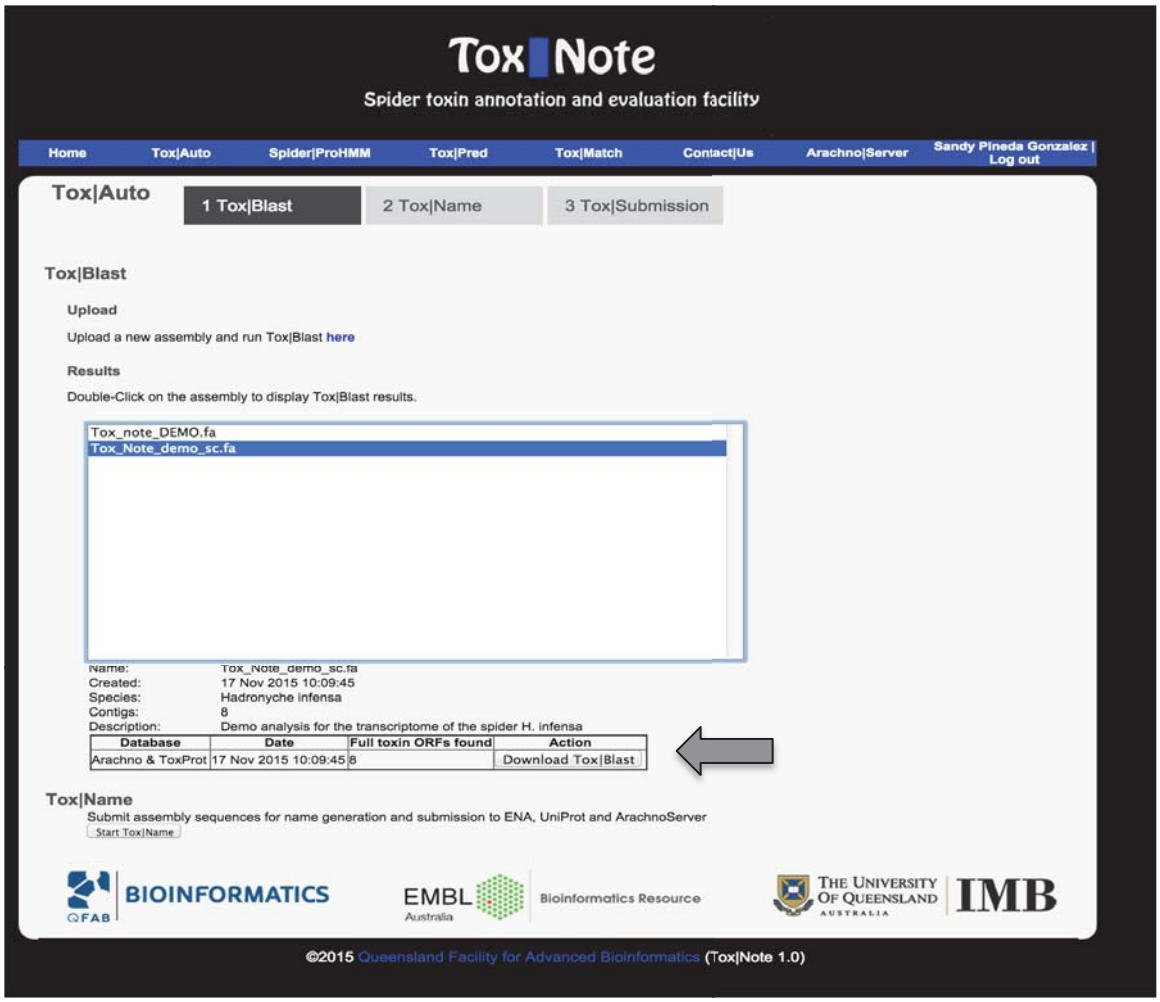


***Figure 3**: **Tox|Note** summary showing the results and the action buttons where users can download csv files.*

The **Tox|Blast** output only displays complete ORFs[5]. This file will also contain relevant information about the BLAST subject ID, BLAST subject length, E-value, identity, mismatches, ORF coordinates, translation frame, and predicted cleavage sites. If the transcripts were only found to be ORFs by **Tox_Seek|**, the BLAST information will appear blank, while the opposite will be the case if **Tox_Seek|** has no information but the transcript has a BLAST hit; the latter case will be recorded anyway, but these cases show that some transcripts are incomplete and do not have a stop codon.

Note that **Tox_Seek|** can make false predictions in the following cases:

[4]Please note that if the user requires BLAST information for the entire dataset, the results displayed by **Tox|Not**e will **ONLY** include **"Toxins/toxin-like"** as the aim of **Tox|Note** is to speed up the isolation of toxin-encoding sequences exclusively.

[5]This includes transcripts with a start and stop codon.

1) When two possible reading frames are detected, but one of the two has no obvious start site or lacks a stop codon due to an incomplete sequence. Take the following example:

```
>TR48672_c0_g1_i1
AAAGGACCAATTATTGGATAATGTGCCGTCTTCCTGCCTAACGCGTGACTGCCTCAGTGACAGGCGATCCGA
CCTTGGTTACATCAGCAGACATTGTGGTCCGTTCCTAACTCACTAGATTTGACCTCACTATTTTCCTGTGGG
GGGAACAGTCGAGTCTTGTGGTCACTTGCCGGAGTTTACTAACAGAGAAACGAACGAGTGCAAAGTTAGGGT
TTATTTAAGTTTGGTCGAAAGAATTCGCAGTTTTAATGATGATAAGGTTTTAAGGAGATAATTTGCATTAGT
AAAAGGTGTCGGGAGAAGTGAGCGGTGTTTTGCAAAGATCAAAGTGGTTGAAAGACTTCCGCGCACCGGTTG
GATCATCAGTGGCATTTTGAAAACAAGTGGCAAGTTTGACAACGAATGGAAAATTGCTTTACCGAAGAGAAA
ATGGATAGTAAAAAAGGCTGCAAGCAATGGCTGAAGCGGAGGACGCATCTCGAGAAGACCCTGATGTCTGTC
TGCGTCGTCCTCTTCCTCACCTGCACTCTCGTCATCTTGTTCGGAATCACTTTCGATGGTAAAAGTTCAACG
TCAAAGGTGGAAAGGACCGTCTGTTCTTCAGATGCCTGCATTCAAATAGCCTCGGTGATGCTGAAGAAGATG
GACCCAGCAACAGATCCTTGTGAAGACTTCTACGAGTTCTCTTGCGGAAGATACCTGAGGATGCACGAGATT
CCCGATGATTTCCACGAGCGATCAGTGGAGCAGACCTCTGTAGACGAAATTCGCTTGCAGGTCAAAAATTGG
TTGGAAACGGAAACTAAAGACGAAAATATCGCTGCCTTTAGCAAAGCTAAAATCTTGTACAGCACATGTATG
AATTTTAGTTTACCGGAAGACGATTACCGCCCTTTTCTGGAGAAGGTGTACGTGCAACAAATGAATGACACG
TGGCCTGTTTTGGACAGCCATTGGGAGGAAAAAGGATTTGAGAAGACATTCGCAGCTCTTACACTGCTCGAT
ATTCCTGCTGCTTTTCATCTAGAAATTGTTCCTGATGCTCGTGATTCATCAAAATACATCGCCAGATTGTCG
CCAGGAGAGCCATTGCTTGATCAGGAATTCTTTCAAGGGAAGAGAGAAAAATCCCTTATTACGTCTTACACT
TCGATGGTCGTATCCGCATTCATGATGCTTGGTTTGGATAGGAACAGAGCAGTCAGAGACTTCGAAGAAATC
CTCTCCATAGAAAGGGAATTGGTGGGATTCAGCAAGATGGCAAAGGAGGAATGTGGCCAACAGTCCGAAGAG
AATCCAAGCGGAAGTGGTTTGCATACAGAGAGATTTCCGCTCTCTCAATTAAATCAGAGGATGCCTTCGGGG
TTGAATTGGAAGGAAACGATGAAGCTTATCTTTACCGGAGCGAACATCACAGATAACATGGAAATCGAAATC
TATTGCGGGAAGCATATTTGGAATTACGCAGAATATCTTCTGTCAGGAACATTCAGTGCAAAATCTACAATG
GTCTATTTGGGATGGAGATTTCTGTTTCCCTTCATCCAGTACCTCGGCCAACCTTTTCACAGACTGCATCAA
GATTACAGTGAGCAAGTTACGGGAAGATTTTCTCAAAGAATCCACTCCAGATGGAAGGAATGTGTCCTTCTG
GTGGAACAAAGAATGCTTCCTGTTGTAGCGGCTGTCTACGGCGAACAGGAAGTGACGAAAGCCGTCCACGAT
TCGGTTGAGAAGATGTTGAAGAGTGTGAAGGCTAGTTTCGGGCATTTCCTGACTTCTGCGAGCTTCCTCAGT
GAGAGTGAAAGAAACAGGAGTAAGCAAAAGTTATCAAGACTGGTCTTCGAAATAGCGCTGATGAATTACTCT
AGGGACTTGGAGAAAGTGGATCGGATTTTTTCTCAGCTGAGTCTGGCAGACGATCACCTGTTATCGAACATC
GTTCGCTTGCAGCGCTGCAAGGTAGATAATCGGCTGCAGAAGATTCTCTCTCCTCACAAAGAAGGC
```

This transcript has one potential long ORF in frame 1, along with smaller frames including one in frame –1:

```
5'3' Frame 1
KGPIIG-CAVFLPNA-LPQ-QAIRPWLHQQTLWSVPNSLDLTSLFSCGGNSRVLWSLAGV
Y-QRNERVQS-GLFKFGRKNSQF----GFKEIICISKRCREK-AVFCKDQSG-KTSAHRL
DHQWHFENKWQV-QRMENCFTEEKMDSKKGCKQWLKRRTHLEKTLMSVCVVLFLTCTLVI
LFGITFDGKSSTSKVERTVCSSDACIQIASVMLKKMDPATDPCEDFYEFSCGRYLRMHEI
PDDFHERSVEQTSVDEIRLQVKNWLETETKDENIAAFSKAKILYSTCMNFSLPEDDYRPF
LEKVYVQQMNDTWPVLDSHWEEKGFEKTFAALTLLDIPAAFHLEIVPDARDSSKYIARLS
PGEPLLDQEFFQGKREKSLITSYTSMVVSAFMMLGLDRNRAVRDFEEILSIERELVGFSK
MAKEECGQQSEENPSGSGLHTERFPLSQLNQRMPSGLNWKETMKLIFTGANITDNMEIEI
YCGKHIWNYAEYLLSGTFSAKSTMVYLGWRFLFPFIQYLGQPFHRLHQDYSEQVTGRFSQ
RIHSRWKECVLLVEQRMLPVVAAVYGEQEVTKAVHDSVEKMLKSVKASFGHFLTSASFLS
ESERNRSKQKLSRLVFEIALMNYSRDLEKVDRIFSQLSLADDHLLSNIVRLQRCKVDNRL
QKILSPHKEG

3'5' Frame 1
AFFVRRENLLQPIIYLAALQANDVR-QVIVCQTQLRKNPIHFLQVPRVIHQRYFEDQS--
LLLTPVSFTLTEEARRSQEMPETSLHTLQHLLNRIVDGFRHFLFAVDSRYNRKHSLFHQK
DTFLPSGVDSLRKSSRNLLTVILMQSVKRLAEVLDEGKQKSPSQIDHCRFCTECS-QKIF
CVIPNMLPAIDFDFHVICDVRSGKDKLHRFLPIQPRRHPLI-LRERKSLCMQTTSAWILF
GLLATFLLCHLAESHQFPFYGEDFFEVSDCSVPIQTKHHECGYDHRSVRRNKGFFSLPLK
EFLIKQWLSWRQSGDVF--ITSIRNNF-MKSSRNIEQCKSCECLLKSFFLPMAVQNRPRV
```

```
IHLLHVHLLQKRAVIVFR-TKIHTCAVQDFSFAKGSDIFVFSFRFQPIFDLQANFVYRGL
LH-SLVEIIGNLVHPQVSSARELVEVFTRICCWVHLLQHHRGYLNAGI-RTDGPFHL-R-
TFTIESDSEQDDESAGEEEDDADRHQGLLEMRPPLQPLLAAFFTIHFLFGKAIFHSLSNL
PLVFKMPLMIQPVRGSLSTTLIFAKHRSLLPTPFTNANYLLKTLSSLKLRILSTKLK-TL
TLHSFVSLLVNSGK-PQDSTVPPTGK--GQI--VRNGPQCLLM-PRSDRLSLRQSRVRQE
DGTLSNNWSF
```

In this particular example **Tox_Seek|** will incorrectly predict the ORF, as ORF frame +1 lacks a stop codon, and therefore it will choose frame –1 instead. Users need to be careful when sorting out the **Tox|Blast** output.

In other instances, we have noticed that **Tox_Seek|** might choose the wrong ORF even when the Kozak score of the selected ORF is high. Take the following example:

```
>RL_rep_c11533
ATCACTATGAAGTTTGCAGTTCTCTTTGGCGTTCTTTTAGTAACGCTTTTCAGCTACTCTTCAGCTGAAATA
CTTGATGATTTAGAGCAAGCGGACGACGCTGATGAGCTGTTATCTTTAATAGAAGAGCAAACCAGAGCCAAG
GAATGTACCCCAAGGTTTAGCGACTGTACTAATGATCGCCACAGTTGCTGCCGAGGCGAATTGTTCAAAGAT
GTCTGCACATGCTTTACGCAGAAAACGGAGGAAACGAGTTCTGTACATGCCAACAACCCAAACATTACAAGT
ATATTGAAAAGGCACAGACAAGCTTAAGAAATTCGGCAGCAAGATTAAGAAATGGTTCGGTTAATGAGACAA
TATCGTTCGTAATGGATATGCTTAATAAATCCAAATATTTCT
```

Compare frame +1 and frame –2, highlighted in red:

```
5'3' Frame 1
ITMKFAVLFGVLLVTLFSYSSAEILDDLEQADDADELLSLIEEQTRAKECTPRFSDCTND
RHSCCRGELFKDVCTCFTQKTEETSSVHANNPNITSILKRHRQA-EIRQQD-EMVRLMRQ
YRS-WICLINPNIS
5'3' Frame 2
SL-SLQFSLAFF--RFSATLQLKYLMI-SKRTTLMSCYL--KSKPEPRNVPQGLATVLMI
ATVAAEANCSKMSAHALRRKRRKRVLYMPTTQTLQVY-KGTDKLKKFGSKIKKWFG--DN
IVRNGYA--IQIF
5'3' Frame 3
HYEVCSSLWRSFSNAFQLLFS-NT--FRASGRR--AVIFNRRANQSQGMYPKV-RLY--S
PQLLPRRIVQRCLHMLYAENGGNEFCTCQQPKHYKYIEKAQTSLRNSAARLRNGSVNETI
SFVMDMLNKSKYF
3'5' Frame 1
RNIWIY-AYPLRTILSH-PNHFLILLPNFLSLSVPFQYTCNVWVVGMYRTRFLRFLRKAC
ADIFEQFASAATVAIISTVAKPWGTFLGSGLLFY-R-QLISVVRLL-IIKYFS-RVAEKR
Y-KNAKENCKLHSD
3'5' Frame 2
EIFGFIKHIHYERYCLINRTIS-SCCRIS-ACLCLFNILVMFGLLACTELVSSVFCVKHV
QTSLNNSPRQQLWRSLVQSLNLGVHSLALVCSSIKDNSSSASSACSKSSSISAEE-LKSV
TKRTPKRTANFIV
3'5' Frame 3
KYLDLLSISITNDIVSLTEPFLNLAAEFLKLVCAFSIYL-CLGCWHVQNSFPPFSA-SMC
RHL-TIRLGSNCGDH-YSR-TLGYIPWLWFALLLKITAHQRRPLALNHQVFQLKSS-KAL
LKERQRELQTS--
```

In this particular case, BLAST analysis of both ORFs clearly showed that the toxin transcript is the one in frame +1. However, the computation and comparison of the Kozak score[6] of both frames, i.e., Frame +1= 0 and Frame –2= 5.6500006 shows, that in cases where the ORF has a score of 0, **Tox_Seek|** will choose any other ORF with a higher

[6]To calculate the Kozak score, the script that **Tox_Seek|** uses requires at least 10 nucleotides before the start codon. If 10 nucleotides are not available before the start of the sequence, the score will be 0.

score and this will be the sequence displayed in the **Tox|Blast** output file. If the computed value is 0 but no other ORFs are detected, the sequence will be recorded anyway and needs to be verified by the user.

In all of the above situations, if the user doubts the result, we always recommend checking all the ORFs before using the prediction. **Because Tox|Note is completely automated, ORFs detected by the users as wrong or doubtful SHOULD NOT be used for further analysis or downstream applications including the database submission**. We hope these issues will be minimal, but users are urged to be cautious.

Tox|Name: automatic toxin name generator

The **Tox|Blast** output is the first step towards generating information about toxins and toxin-like transcripts in a venom-gland transcriptome. Once completed, the user is directed to **Tox|Name tab**, the automatic toxin name generator. This step is crucial, as it enables the user to automatically apply rational nomenclature rules for naming the toxin (King et al., 2008) and use this information to proceed to the European Nucleotide Archive (ENA) submission process.

To generate names for a group of toxins the user simply needs to copy and paste the IDs of all contig/singlets from the **Tox|Blast** output file that require a name into the designated box (Figure 4). **Tox|Name** then automatically generates rational names for each toxin (Figure 5). Users have the option of downloading this file as a csv, xml or a PDF file.

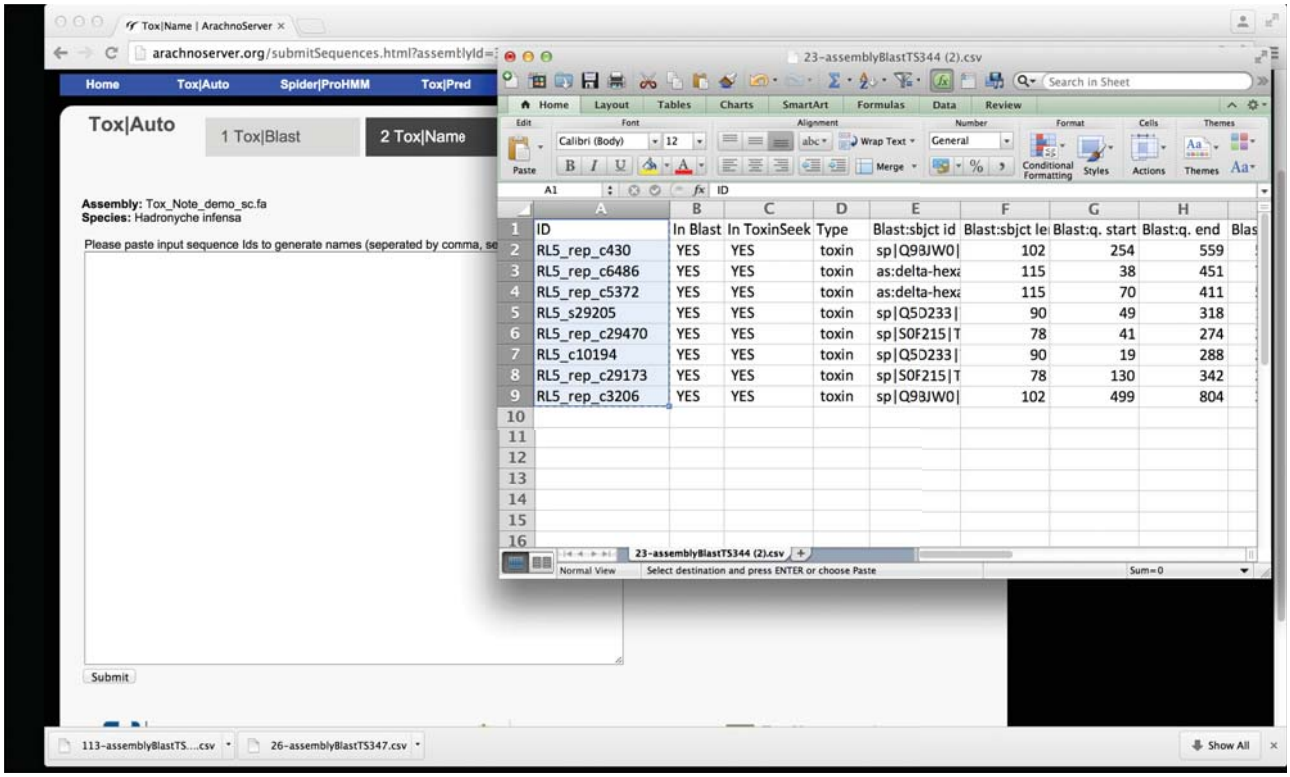


***Figure 4**: Example of a csv file with the IDs needed to start the automatic name generator. (**Tox|Name** tab). Simply copy and paste the IDs from the **Tox|Blast** output into the provided box and press the **Submit** button.*

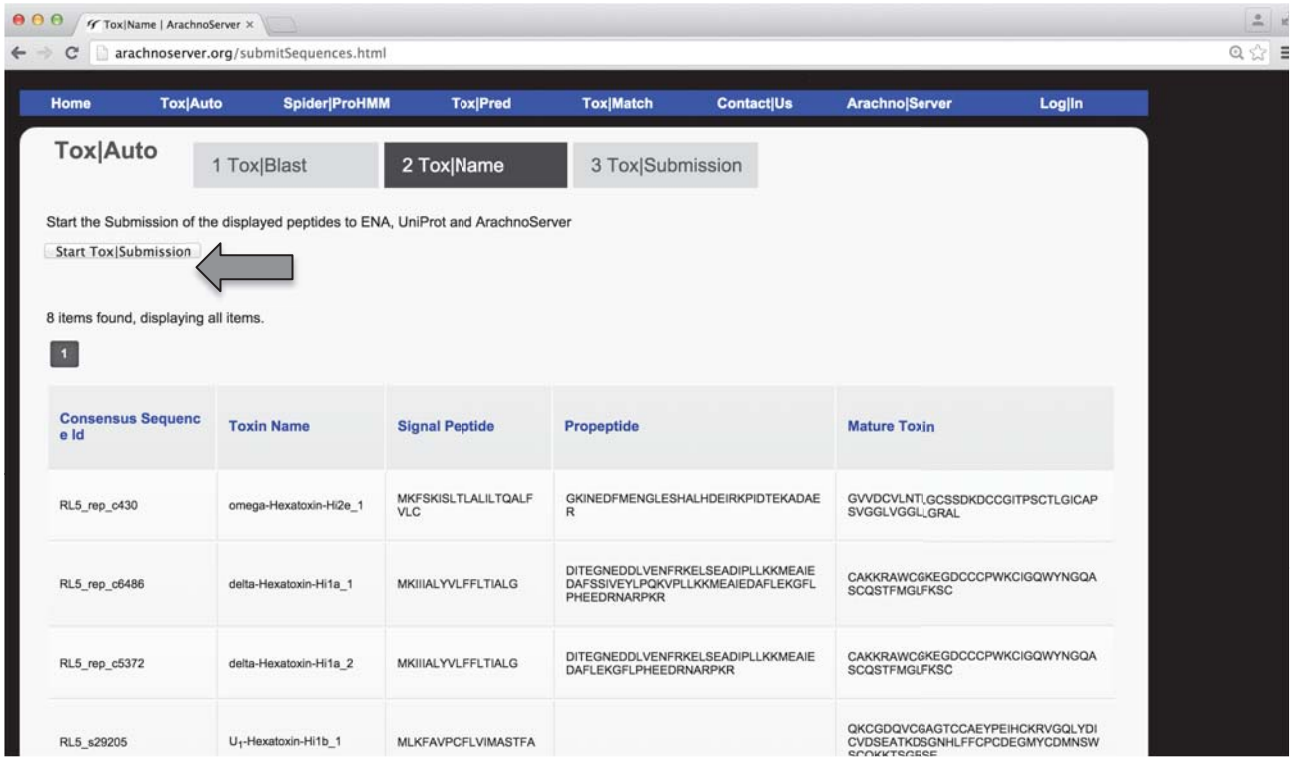


***Figure 5**: Example of a **Tox|Name** output, showing the result of automatic name generation based on the rational nomenclature (King et al., 2008) used in ArachnoServer and VenomZone. From this step you can access the tab for **Tox|Submission** (grey arrow), described in detail below.*

Currently Tox|Name works with all species recorded on ArachnoServer, except with peptides isolated from the Sicariidae family, which do not follow the standard nomenclature described in King 2008.

Tox|Submission: automated submission to ENA and ArachnoServer

No sequence annotation process is complete without submission of all available sequence information to the relevant databases. **Tox|Note** endeavours to achieve this automatically without the need to connect to any external websites; the required information is automatically submitted to ENA and subsequently to UniProt and ArachnoServer.

Sequences can be released immediately, or held privately by ENA for up to two years in accordance with their policies. The user will receive the relevant project and accession numbers after all sequences have been verified by ENA. Subsequently, these accession numbers will be forwarded to UniProt and ArachnoServer. This event will trigger the creation of a toxin card on ArachnoServer. Please note that for the first release of **Tox|Note**, users will have to finish their submission in one session. Saving and returning later to the submission form is currently **NOT supported**, but we hope it will be in the next release of **Tox|Note**.

NOTE: users must provide all relevant information from their respective experiment(s) before they can submit the project. The form below is an example of the standard information that users need to provide to ENA when submitting sequences, and the same information will be used by **Tox|Note** to generate submissions to ENA.

There are a total of six sections to be completed with experimental information. **Tox|Submission** also requires the experimental BAM file, the FASTA file from the assembly (supplied at the beginning of the **Tox|Blast**), and the SFF file (if dealing with 454 experiments) or Fastq files from Illumina. Please follow this checklist before starting the submission; this will save time and will give you hints about information required for the submission process. Fields with asterisks are mandatory (shown in the text list in red).

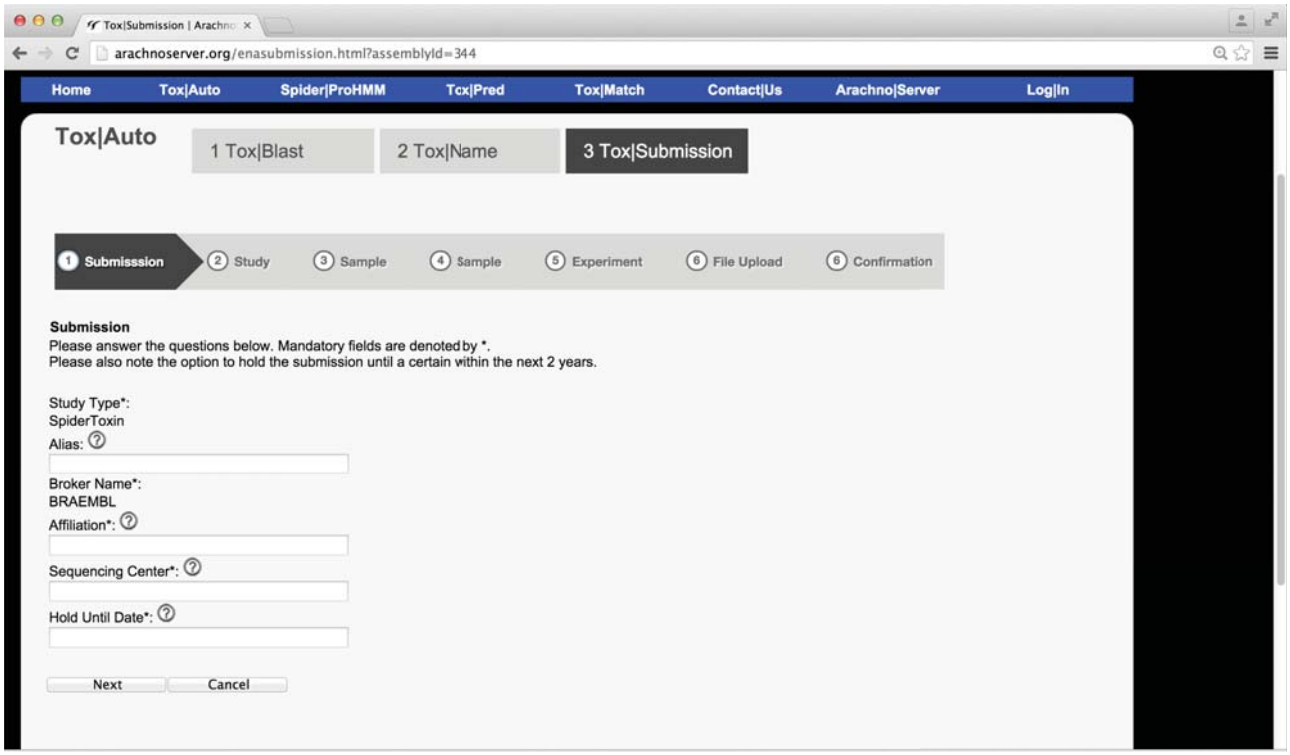


***Figure 6**: **Tox|Submission** tab at a glance, showing the six main steps that need to be completed in order submit an annotation package to the European Nucleotide Archive. Please check the submission checklist (below) before you start your submission.*

Tox|Submission check list

	Field	Description	Example/ Expected value	Your study
	Section 1 - Submission			
1	Alias*		Use the name you assigned to your project	
2	Broker name*	EMBL-ABR	This is a default text	This is a default text
3	Affiliation*	Name of the institution in which the research is being conducted.	Institute for Molecular Bioscience	
4	Sequencing centre*	Sequencing centre were samples were sequenced.	Australian Genome Research Facility	
5	Hold date*	Date when the user wants the sequences to be released.	15/03/16	
	Section 2 - Study			
6	Title*	Name of the project within which the sequencing was organized.	Probing the chemical diversity of venom from the Australian funnel-web spider *Hadronyche infensa.*	
7	Abstract*	Brief summary of the	Spiders produce an extraordinarily complex	

		research that was undertaken.	venom for defense, prey capture, and competitor deterrence. However, very little is known about the genetic and transcriptional mechanisms used by spiders to generate such complex chemical cocktails. A combined proteomic and transcriptomic approach was used to examine the range of peptides and proteins expressed in the venom of Australian funnel-web spiders. This analysis revealed that the venom comprises more than 3000 peptides and proteins divided across 32 superfamilies.	
	Section 3 -Sample			
8	Taxon ID*	Taxon ID from the NCBI taxon identification website. It will be selected from species at the assembly upload.	153481	
9	Scientific Name*	Scientific name selected from the species at assembly upload.	*Hadronyche infensa*	
10	Collection date*	The time of sampling, either as an instance (single point in time) or interval. If an exact time is not available, the date/time can be right truncated i.e. all of these are valid times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008; Except:2008-01; 2008 all are ISO8601 compliant	2007	
11	Source material identifiers	For cultures of microorganisms: identifiers for two culture collections; for specimens (e.g., organelles and Eukarya): voucher condition and location.	NA	
12	Sample collection device or method	The method or device employed for collecting the sample; i.e., PMID,DOI or URL, biopsy, niskin bottle, push core, venom-gland dissection.	Venom-gland dissection.	
13	Size of sample collected	Amount or size of sample (volume, mass or area) that was collected	6 venom glands.	
14	Sample material processing	Any processing applied to the sample during or after retrieving the sample from environment. This field accepts OBI, for a browser of OBI (v 2013-10-25) terms please see http://purl.bioontology.org/ontology/OBI	All spiders used for complementary DNA (cDNA) library construction were first anesthetized using CO2, frozen at –80°C for 10 min, and then dissected at 4°C. To access the venom gland, the chelicerae were first removed from the base of the structure (towards the carapace). Once separated, each chelicera was individually cut from the ventral side up to the base of the fang in order to approach the venom gland from underneath. The cheliceral muscle that surrounds the venom gland was detached to expose the isolated venom gland. Dissected venom glands were placed immediately in TRIzol® reagent (Invitrogen) to be processed.	
15	Isolation growth condition	Publication reference in the form of PubMed ID (PMID), digital object identifier (doi) or url for isolation	Individual spiders were housed at 23–25˚C in plastic containers (approximately 155 × 155 × 140 mm) in dark cabinets. The substrate in each container comprised 1/3 moist peat moss and 2/3 washed sand. Spiders were fed fortnightly	

		and growth condition specifications of the organism/material. Mandatory for MIGS and MIMARKS Specimen.	with 2/3 of a 2-4-day-old mouse, which was removed 24 h later if not eaten.	
16	Propagation	This field is specific to different taxa. For phages: lytic/lysogenic, for plasmids: incompatibility group (Note: there is a strong view that phage propagation should be named obligately lytic or temperate, therefore we also give this choice. Mandatory for MIGS of eukaryotes, plasmids and viruses.	NA	
17	Geographic location (latitude and longitude)*	The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system. The geographical origin of the sample as defined by the country or sea. Country or sea names	Orchid Beach, Fraser Island, Queensland Australia. 153.3223452, -24.9582736	
	Section 4 -			
18	Investigation type*	Nucleic Acid Sequence Report is the root element of all MIGS/MIMS compliant reports as standardized by Genomic Standards Consortium. This field is either eukaryote, bacteria, virus, plasmid, organelle, metagenome, mimarks-survey or mimarks-specimen	Nucleic acid sequences and peptide sequences from venom-gland preparations from the Australian funnel-web spider Hadronyche infensa. Eukaryote/Artropoda/Chelicerata/Arachnida/Araneae/Mygalomorphae/Hexathelidae/Hadronyche	
19	Project name*	Name of the project within which the sequencing was organized.	Can be the same as title	
20	Sequencing method*	Sanger dideoxysequencing, pyrosequencing, ABI-solid, etc.	Pyrosequencing	
21	Experimental factor (EFO)	Experimental factor are essentially the variable aspects of an experiment design which can be used to describe an experiment, or set of experiments, in an increasingly detailed manner. This field accepts ontology terms from Experimental Factor Ontology (EFO) and/or Ontology for Biomedical Investigations (OBI). For a browser of EFO (v 2.43) terms, please see http://purl.bioontology.org/ontology/EFO; for a browser of OBI (v 2013-10-25) terms, please see http://purl.bioontology.org/ontology/OBI	Parent EFO: http://www.ebi.ac.uk/efo/EFO_0001032. cDNA EFO: http://www.ebi.ac.uk/efo/EFO_0004187	

22	Library construction method*	Library construction method used for clone libraries	One hundred nanograms of mRNA preparation was used to construct a cDNA library using the methods described in the cDNA rapid library preparation method manual and emPCR method manual (Rev Jan 2010 Brandford, CT, USA). Once the library was constructed and amplified it was re-analysed using the Bioanalyzer 2100 (Agilent Technologies) and then sequenced using a ROCHE GS-FLX	
23	Library screening strategy	Enriched, screened, normalized.	Other-TSA	
24	Library reads sequenced	Number of reads sequenced/Total number of clones sequenced from the library.	300770	
25	Observed Biotic Relationship	Is it free-living or in a host and if the latter what type of relationship is observed.	Free-living	
26	Sub specific genetic lineage	This should provide further information about the genetic distinctness of this lineage by recording additional information i.e biovar, serovar, serotype, biovar, or any relevant genetic typing schemes like Group I plasmid. It can also contain alternative taxonomic information.	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Chelicerata; Arachnida;Araneae; Mygalomorphae; Hexathelidae; *Hadronyche*	
27	Relationship to oxygen	Is this organism an aerobe, anaerobe? Please note that aerobic and anaerobic are valid descriptors for microbial environments	Aerobe	
	Section 5 -Experiment			
28	Platform*	Select from drop down menu.	LS454	
29	Instrument model*	Select from drop down menu.	454 GS FLX	
30	Library source*	Select from drop down menu.	Transcriptomic	
31	Library selection*	Select from drop down menu.	cDNA	
32	Library strategy*	Select from drop down menu.	Other-TSA	
33	Library layout*	Select from drop down menu		
34	Library name	Name assigned to the library.	RL5-GKSPGFW	
	Section 6 –File upload			
35	Read File*	FASTA file containing assembled sequences		
36	Bam file upload*	BAM file describing how transcripts were constructed from the reads. Transcripts have to be annotated 5' or 3' (forward strand). *Be readable by SAM Tools.	Converted MAF to SAM and SAM to BAM using SAM tools.	
37	Assembly method*	Assembly software used to generate contigs	Mira 3.2	

Notes:

1. Please note that it is mandatory to complete the fields marked with an asterisk (*) and red text in order to complete the submission form.
2. Avoid excessive use of symbols in the names of sequences, contigs, singlets, and others, especially the forward slash symbol (\). These symbols can create issues while parsing all the scripts in the pipeline.

Tox|Pred and Tox|Match

Tox|Pred and **Tox|Match** can be used in conjunction with **Tox|Note** or as standalone features. **Tox|Pred** and **Tox|Match** were designed with the idea that venom-gland transcriptomes are often sequenced in parallel with proteomic experiments on venom.

Any FASTA file containing only mature peptide sequences from a transcriptome can be used to calculate the theoretical mass for each peptide. **Tox|Pred** allows users to upload files or copy/paste text into the provided box. Currently **Tox|Pred** does not support calculation of the mass of toxins with post-translational modifications (except C-terminal amidation) or chemical modifications such as cysteine alkylation. The user should not use this feature if such modifications were made.

The theoretical toxin masses calculated by **Tox|Pred** can be compared with experimental masses from mass spectrometry experiments to generate a list of matching masses using the **Tox|Match** application on the following tab. The user can adjust the error tolerance from 0 to1 Da, including decimal values.

A maximum of ~300 sequences can be computed at the time using the online prediction tab or ~1000 using the upload tab.

A version of **Spider|ProHMM** has been made available as a standalone tool so users can decide if they want to use SpiderP (the previously published SVM approach available on the main ArachnoServer site) or the new HMM approach. A maximum of ~300 sequences can be used on the online prediction tab while ~1000 can be done through the upload tab.

References

EDDY, S. R. 2011. Accelerated Profile HMM Searches. *PLoS Computational Biology* 7**,** e1002195.

KING, G. F., GENTZ, M. C., ESCOUBAS, P. & NICHOLSON, G. M. 2008. A rational nomenclature for naming peptide toxins from spiders and other venomous animals. *Toxicon* 52**,** 264–276.

KOZAK, M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research* 15**,** 8125–8148.

PETERSEN, T. N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* 8**,** 785–786.

WONG, E. S., HARDY, M. C., WOOD, D., BAILEY, T. & KING, G. F. 2013. SVM-based prediction of propeptide cleavage sites in spider toxins identifies toxin innovation in an Australian tarantula. *PLoS ONE* 8**,** e66279.