FEBS Letters 588 (2014) 3891–3897

journal homepage: www.FEBSLetters.org

Gene structure, regulatory control, and evolution of black widow venom latrotoxins

Kanaka Varun Bhere [a,1], Robert A. Haney [a,1], Nadia A. Ayoub [b], Jessica E. Garb [a,*]

[a] Department of Biological Sciences, University of Massachusetts Lowell, MA, USA
[b] Department of Biology, Washington and Lee University, Lexington, VA, USA

ARTICLE INFO

Article history:
Received 22 July 2014
Revised 28 August 2014
Accepted 29 August 2014
Available online 12 September 2014

Edited by Takashi Gojobori

Keywords:
α-Latrotoxin
Venom
Neurosecretion
Genomics
Molecular evolution
Latrodectus

ABSTRACT

Black widow venom contains α-latrotoxin, infamous for causing intense pain. Combining 33 kb of *Latrodectus hesperus* genomic DNA with RNA-Seq, we characterized the α-latrotoxin gene and discovered a paralog, 4.5 kb downstream. Both paralogs exhibit venom gland specific transcription, and may be regulated post-transcriptionally via musashi-like proteins. A 4 kb intron interrupts the α-latrotoxin coding sequence, while a 10 kb intron in the 3′ UTR of the paralog may cause non-sense-mediated decay. Phylogenetic analysis confirms these divergent latrotoxins diversified through recent tandem gene duplications. Thus, latrotoxin genes have more complex structures, regulatory controls, and sequence diversity than previously proposed.

1. Introduction

Venom proteins can evolve rapidly through gene duplication and adaptation in response to selection imposed by diverse and co-evolving prey [1–4]. Venom toxins also have important biomedical applications, including receptor characterization and drug discovery [5,6]. However, few studies have examined the structure and regulation of genes encoding venom proteins [7]. For spiders, one of the largest venomous clades, nearly all sequences come from venom gland cDNAs [8,9]. Thus, the roles of gene duplication, alternative splicing, and regulatory controls in generating venom molecular complexity are poorly understood.

The spider genus *Latrodectus* includes black widows (multiple species) and the Australian red-back spider (*Latrodectus hasselti*), the venoms of which have potent neurotoxic effects on vertebrates. *Latrodectus* venom contains latrotoxins, a family of neurotoxic proteins that share a unique N-terminal domain flanked by 11–20 ankyrin motif repeats [6,8,10]. Four latrotoxins have been functionally characterized from *Latrodectus tredecimguttatus*, and while three (α-latroinsectotoxin, α-latrocrustotoxin and δ-latroinsectotoxin) elicit neurotransmitter release in arthropods [11], α-latrotoxin forms calcium channels in vertebrate presynaptic neuronal membranes, thereby triggering massive neurotransmitter exocytosis [12–14]. α-Latrotoxin is responsible for the extreme pain resulting from black widow bites [6], and is important for studying neurosecretion, and hence has received considerable scientific attention.

While *Latrodectus* venom contains multiple latrotoxins, they have only been identified in three genera of the family Theridiidae, suggesting latrotoxins represent a recently expanded protein family [8,15,16]. Evidence from RNA-Seq data indicates ⩾20 divergent latrotoxins are expressed in venom glands of the Western black widow (*Latrodectus hesperus*) and most are phylogenetically distinct from the functionally characterized latrotoxins [16]. We hypothesize these transcripts are encoded by distinct loci that are spatially clustered in the genome and that their transcription and translation is tightly controlled to ensure strong venom gland-specific expression.

We sequenced 33 kb from the *L. hesperus* genome encompassing the α-latrotoxin gene, which we integrated with venom gland

Abbreviations: FPKM, fragments per kilobase per million library reads; TE, transposable element; LINE, long interspersed nuclear element; MaSp, major ampullate silk protein; TSS, transcription start site; TF, transcription factor; MBE, musashi binding element; LTR, long terminal repeat; NMD, non-sense mediated decay; FDR, false discovery rate

* Corresponding author at: Department of Biological Sciences, 198 Riverside Street, Olsen Hall 520, University of Massachusetts Lowell, Lowell, MA 01854, USA. Fax: +1 978 934 3044.

E-mail address: Jessica_Garb@uml.edu (J.E. Garb).

[1] These authors contributed equally to this paper.

http://dx.doi.org/10.1016/j.febslet.2014.08.034

expression data (RNA-Seq and Expressed Sequence Tags (ESTs)) to reveal α-latrotoxin's gene structure, quantify its expression, and investigate its regulation. This revealed a divergent, highly expressed latrotoxin paralog 4.5 kb downstream of α-latrotoxin, and long introns and putative regulatory elements in both paralogs, providing novel insights into latrotoxin evolution and production.

2. Materials and methods

2.1. Genomic sequencing

A genomic library was constructed from eight *L. hesperus* females to cover the estimated 1261 Mb genome [17]. Library construction details are in Supplementary Material and Ayoub et al. [18]. α-Latrotoxin primers were used to PCR-screen the genomic library, revealing three positive clones. The clone with the smallest insert (estimated from a *Bam*HI digest) was used to make a shotgun library from three complete, separate digests (*Eco*RI, *Pst*I and *Eco*RV). Resulting fragments were ligated into pZErO™-2 plasmids (Invitrogen) and electroporated into TOP10 *Escherichia coli*. For each digest, a library of 192 clones was screened for insert size. 1–2 kb inserts were sequenced using Sp6 and T7 primers. Sequences were edited and assembled in SEQUENCHER 4.2 (Gene Codes Corp.). Primer walking was employed to complete insert sequencing. The complete sequence was deposited at NCBI (Accession KM382064).

2.2. Genomic annotation

Open Reading Frames (ORFs) were predicted from the assembled genomic insert sequence using getORF, retaining ORFs encoding $\geqslant$30 amino acids. Predicted proteins were subjected to BLASTp searches against NCBI's non-redundant (*nr*) protein database. Gene structure and expression were explored using female *L. hesperus* RNA-Seq data [19] from three tissues: (1) venom gland, (2) total silk gland tissues, and (3) cephalothorax minus venom glands [16,19]. This included 133 million high quality 75–100 bp paired-end sequence reads collectively. Trinity [20] was used to assemble tissue-specific reads, and overlapping sequences across libraries were merged using CAP3 [19]. Tophat 2.0.8b [21] was used to align RNA-Seq reads from tissues to the genomic sequence to map introns, which were visualized with Integrative Genomics Viewer (IGV) 2.3 [22]. Cufflinks 2.0.2 [23] was used to assemble transcripts from Tophat mappings, filtering out isoforms representing <10% of major isoform abundance. Cuffmerge was used to merge assemblies from tissues, and Cuffdiff was used to produce expression estimates for transcripts in each tissue. OrfPredictor [24] was used to translate proteins from transcripts in the frame of the top *nr* BLASTx hit. We also mapped *L. hesperus* venom gland ESTs [25] (NCBI dbEST Accessions JZ577614–JZ578096) to the genomic sequence.

The insert sequence was annotated with MAKER [26]. Gene prediction was performed using Augustus under *Drosophila melanogaster* training parameters and mapping: (1) *de novo* Trinity assembled transcripts [19] (2) Cufflinks transcripts predicted from the insert, (3) venom gland ESTs [25] and (4) latrotoxins translated from Trinity transcripts. RepeatMasker [27] was used to predict repetitive elements and low complexity regions with the *D. melanogaster* database. The programs NNPP 2.2, Match 1.0, UTRscan, and PipMaker [28–31] were used to predict regulatory elements (Supplementary Materials).

2.3. Evolutionary analyses

We obtained latrotoxin proteins from NCBI's *nr* and extracted *L. tredecimguttatus* latrotoxins [8] from the transcriptome shotgun assembly archive using tBLASTn (e-value cutoff 1e-20), with the first 320 amino acids from *L. hesperus* latrotoxins [16] as queries. *L. tredecimguttatus* translations were obtained with OrfPredictor [24]. We aligned these sequences with *L. hesperus* latrotoxins predicted by Cufflinks from the genomic insert and *de novo* assembled from RNA-Seq data [16,19] with COBALT [32]. Phylogenetic analyses were performed with Mr. Bayes 3.2.2 [33], with a "mixed" amino acid model and gamma distribution, running 5×10^6 generations. Clade posterior probabilities were from a 50% majority-rule tree, excluding the first 25% of trees. We produced a nucleotide alignment with PAL2NAL [34] from the protein alignment (Supplementary Materials). PAML 4 [35] was used to test for positive selection among sites by comparing Model M7 (several *dN*/*dS* categories between 0 and 1), and model M8 (adds category with unconstrained *dN*/*dS*), and among sites along specific branches by comparing models A and $A_{\omega 2=1}$. MEME [36] was also employed to find sites under diversifying selection across the latrotoxin phylogeny with an false discovery rate (FDR) < 0.05.

3. Results

3.1. Venom expressed tandem latrotoxins

We sequenced 33342 bp of *L. hesperus* DNA from a genomic library clone containing the α-latrotoxin gene. Sanger sequencing and shotgun assembly of 236 subclones resulted in six contigs 968–8347 bp long, and the assembly was finished through primer walking. getORF predicted 129 proteins in both directions, 15 with significant BLASTp hits (Table 1). Two proteins (1401 and 1393 amino acids) had top BLASTp hits to α-latrotoxin. The upstream predicted protein had a top hit to *L. hesperus* α-latrotoxin with 99% identity and represents the α-latrotoxin locus, while the downstream latrotoxin had a best match to *Steatoda grossa* α-latrotoxin at 44% identity, and represents an adjacent paralog (Table 1). The translated tandem *L. hesperus* latrotoxins share 43% protein sequence identity.

Mapping of *L. hesperus* RNA-Seq data [19] with Tophat to the insert indicated that both latrotoxin genes are highly transcribed in venom glands (Fig. 1), with Cufflinks venom gland fragments per kilobase per million library reads (FPKM) of 80157 for α-latrotoxin and 81563 for the downstream paralog. FPKM in cephalothorax and silk glands was far lower for both latrotoxins (α-latrotoxin: cephalothorax = 1318, silk glands = 1442; downstream paralog: cephalothorax = 76, silk glands = 0).

Venom gland RNA-Seq and EST evidence revealed one 4241 bp phase 1 intron in α-latrotoxin's 25th codon from the stop codon and a 10005 bp intron in the 3′ UTR of the adjacent latrotoxin (Figs. 1 and 2, Table 2). The Cufflinks α-latrotoxin transcript had a 4257 bp coding region, a 316 bp 5′ UTR and a 2805 bp 3′ UTR (Table 2). The Trinity α-latrotoxin transcripts consisted of two overlapping (by 25 bp) fragments, one with a 150 bp 5′ UTR and the other with a 4828 bp 3′ UTR. The downstream paralog's Cufflinks transcript contained a 4182 bp coding region, which was collinear with the genomic sequence, indicating no intron interrupting the coding region. The Cufflinks transcript for this locus had a 675 bp 5′ UTR and a 341 bp 3′ UTR flanking the 10005 bp intron located 57 bp into the 3′ UTR. The Trinity transcripts comprised two overlapping (by 25 bp) fragments, which lacked the first 10 bp of coding sequence, and with the downstream transcript terminating 6 bp before the Cufflinks transcript, but agreeing in intron position. A single bp insertion in the Trinity transcript produced a frameshift and premature truncation, but mapped reads suggested its incorrect assembly. The end of α-latrotoxin's 3′ UTR and the putative transcription start site (TSS) of the adjacent latrotoxin were separated by 4545 or 2522 bp, depending