A Comprehensive Multi-Omic Approach Reveals a Relatively Simple Venom in a Diet Generalist, the Northern Short-Tailed Shrew, *Blarina brevicauda*

Zachery R Hanf[1] and Andreas S Chavez [1,2,*]

[1]Department of Evolution, Ecology, and Organismal Biology, The Ohio State University

[2]Translational Data Analytics Institute, The Ohio State University

*Corresponding author: E-mail: chavez.102@osu.edu.

Accepted: 5 June 2020

Data deposition: Raw RNA-sequencing data used for the transcriptome assembly of Blarina brevicauda's submaxillary gland from samples WH4 and WH5 are available from NCBI Sequence Read Archive (SRA) under SRA run accessions SRR11880369 and SRR11880370. Raw proteomic data for Blarina brevicauda's from samples of saliva from WH4 and WH5 are available from the Mass Spectrometry Interactive Virtual Environment (MassIVE, https://massive.ucsd.edu/, last accessed June 10, 2020) under accession no. MSV000085512). Sixteen new sequences used in tests for positive selection have been deposited in GenBank under accession numbers MT559764-MT559779.

Abstract

Animals that use venom to feed on a wide diversity of prey may evolve a complex mixture of toxins to target a variety of physiological processes and prey-defense mechanisms. *Blarina brevicauda*, the northern short-tailed shrew, is one of few venomous mammals, and is also known to eat evolutionarily divergent prey. Despite their complex diet, earlier proteomic and transcriptomic studies of this shrew's venom have only identified two venom proteins. Here, we investigated with comprehensive molecular approaches whether *B. brevicauda* venom is more complex than previously understood. We generated de novo assemblies of a *B. brevicauda* genome and submaxillary-gland transcriptome, as well as sequenced the salivary proteome. Our findings show that *B. brevicauda*'s venom composition is simple relative to their broad diet and is likely limited to seven proteins from six gene families. Additionally, we explored expression levels and rate of evolution of these venom genes and the origins of key duplications that led to toxin neofunctionalization. We also found three proteins that may be involved in endogenous self-defense. The possible synergism of the toxins suggests that vertebrate prey may be the main target of the venom. Further functional assays for all venom proteins on both vertebrate and invertebrate prey would provide further insight into the ecological relevance of venom in this species.

Key words: de novo genome assembly, long-read sequencing, transcriptomics, proteomics, shrews, venom.

Introduction

Predatory venoms are typically comprised a complex mixture of toxins to aid in prey capture (Casewell et al. 2013). Venom toxins often originate from duplications of genes that subsequently undergo positive selection and neofunctionalization to produce proteins and short peptides that disrupt key regulatory processes or bioactivities of specific prey (Escoubas and King 2009; Fry et al. 2012). In some venomous animals that feed on a multiple prey from a diversity of taxonomic groups, venom is comprised a very complex mixture of toxins that reflect their dietary complexity (Daltry et al. 1996; Barlow et al. 2009; Phuong et al. 2016; Pekár et al. 2018).

If dietary breadth is a driver of venom complexity, then one would predict that venomous animals in the Order Eulipotyphla (shrews, moles, hedgehogs, and solenodons) would have extreme complexity of venom toxins. Predatory venom is found in only a few extant eulipotyphlan species (Dufton 1992; Ligabue-Braun 2015; Rode-Margono and Nekaris 2015; Casewell 2019) and may be found in a few other species (Nussbaum and Maser 1969; Folinsbee 2013; Camargo and Álvarez-Castañeda 2019). The selective pressures leading to the evolution of venom in these shrew species is unclear because both venomous and nonvenomous species feed on a diversity of prey items from widely divergent animal

groups, including Arthropoda, Annelida, Mollusca, and Chordata (Hamilton 1930; Hamilton 1941; Eadie 1952). Some shrew species, including some of the venomous species, are known to cache prey for later consumption, which can be a useful strategy for supplying energy to support their extremely high metabolic rates (Churchfield 1990). Thus, there has been a strong debate as to which selective pressure has led to venom adaptations in eulipotyphlans, including whether the evolution of toxins has been driven by either the need to hunt larger prey (i.e., small vertebrates) or the need to extend the preservation of cached food items by paralyzing prey (Pearsson 1942; Tomasi 1978; Martin 1981). These hypotheses are not mutually exclusive and it is possible that both selective pressures have driven the evolution of venom in these few species. Alternatively in some animals, venom has evolved for defensive purposes, however, this behavior is not known to exist in shrews.

Blarina brevicauda (the northern short-tailed shrew) has the most potent venom among eulipotyphlans (Pearson 1942; Dufton 1992; Folinsbee 2013). Like most other shrews, *B. brevicauda* are diet generalists that consume a wide array of vertebrate and invertebrate prey (Hamilton 1930; Hamilton 1941; George et al. 1986). Early functional studies on *B. brevicauda*'s venom revealed that extracts from their submaxillary gland cause respiratory arrest and even death in vertebrates, as well as paralytic effects on invertebrates (Pearson 1942; Martin 1981). More recent studies have isolated the Soricidin peptide from *B. brevicauda*'s submaxillary glands and found this to have paralytic effects on invertebrates (US Patent No.: US8003754B2). Another study isolated the BLTX toxin, a paralog in the kallikrein-1 subfamily, that contributes to multiple symptoms in mice, including rapid and irregular breathing, low blood pressure, and loss of muscle movement (Kita et al. 2004). However, to explain the full manifestations of all of these symptoms in envenomated mice, other unknown toxin constituents in the saliva are thought to be present (Kita et al. 2004). Therefore, a more complete annotation of the saliva proteome, as well as the submaxillary-gland transcriptome may help reveal additional toxin proteins in the venom repertoire.

Despite the major advances that next-generation sequencing technologies has brought to the field of genomics for nonmodel systems, one persistent issue is generating accurate de novo assemblies of novel genes and their transcripts when using short-read (100–500 bp) sequence data alone (Rhoads and Au 2015; Magi et al. 2018). For example, ambiguous overlap in short reads between similar genes can lead to erroneous chimeric contigs, which can be difficult to distinguish from biologically real transcripts. This problem is particularly an issue in venom systems because many toxin genes arise from gene duplications that have high-sequence similarity with their paralogs (Hargreaves and Mulley 2015). One solution to this problem is including data from long-read sequencing platforms (Sunagar et al. 2016), such as Pacbio and Nanopore MinION, which can be used to bridge genomic scaffolds when generating reference genomes and to sequence entire transcripts instead of assembling fragments of transcripts together from short-read sequences. However, one of the main caveats to long-read sequencing are the higher rates of sequencing error than short-read sequencing. To resolve this, researchers are using approaches that combine data from both short-read and long-read sequencing platforms to correct errors in long-read data, as well as filtering out erroneous transcripts assembled from short-read data (Haas et al. 2013; Hackl et al. 2014; Tan et al. 2018).

The main goals of this study were to identify all known and candidate toxins in the venom repertoire of *B. brevicauda* by using an integrative multiomic approach (genomic, transcriptomic, and proteomic) and to investigate the evolutionary relationships and rates of these toxin genes. To do this, we first generated a de novo transcriptome assembly of the submaxillary gland using both short-read and long-range sequencing data. We then assembled a de novo reference genome to be used with the reference transcriptome to bioinformatically search for transcripts with homology to known venom components found in other venomous systems. From the list of putative venom genes, we then explored patterns of gene expression, evidence for positive selection, and 3D-protein structure to provide more insight into their potential role as a venom component. These transcriptomic results were then compared with proteomic profiles of saliva from *B. brevicauda* to determine whether toxin transcripts from the venom-producing submaxillary gland were present in the saliva and are potentially being used as venom.

Materials and Methods

Animal Capture and Tissue and Saliva Procurement

We captured two wild *B. brevicauda* animals (one female: WH4, and one male: WH5) in pitfall traps near woodpiles at The Ohio State University's Waterman Farm Headquarters in June 2017. Saliva was collected from the shrews by allowing them to bite on a piece of sterile medical tubing. The tubes were then placed in a sterile Eppendorf tube, put on ice, and were immediately stored at −80 °C. Shrews were then euthanized *via* an extended period of inhaled isoflurane and their submaxillary glands and heart were immediately placed in RNA Later (Invitrogen) at room temperature for 24 h then followed by freezing at −80 °C.

High-Molecular Weight gDNA Extraction

High-molecular weight genomic DNA (HMW gDNA) was isolated from the heart tissue of WH4 *B. brevicauda* animal using a Puregene Kit (Qiagen) following the manufacturer's protocol with slight modifications. These modifications included replacing all steps that required vortexing with gentle inversions to reduce damage to DNA. HMW gDNA was quantified

with the Qubit dsDNA HS Assay Kit (Invitrogen) and the quality of the molecular weight was assessed using both genomic screen tapes on a Tapestation 2200 (Agilent) and pulse-field gel electrophoresis on a Pippin Pulse (Sage Sciences). Fragments smaller than 600 bp were removed from the HMW gDNA sample using the Pippin HT (Sage Science).

Genomic DNA Library Construction, Sequencing, and De Novo Genome Assembly

A Chromium Controller Instrument (10× Genomics) at the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center was used for sample preparation of a 10× Genomics "linked-read" library to be used to generate a de novo genome assembly. Sample indexing and partition-barcoded libraries were prepared using the Chromium Genome Reagent Kit (10× Genomics) according to manufacturer's protocols described in the Chromium Genome User Guide Rev A (https://support.10xgenomics.com/permalink/2ofuH1pVbWyeCg2s6u6EwY, last accessed June 10, 2020). In summary, approximately 1 ng of HMW gDNA in Master Mix was combined with a library of Genome Gel Beads and partitioning oil to create Gel Bead-In-Emulsions (GEMs) within a microfluidic Genome Chip. HMW gDNA was partitioned across ~1 million GEMs where library construction took place. The library construction incorporated a unique 16-bp barcode, an Illumina R1 sequencing primer, and a 6-bp random primer sequence. GEM reactions were isothermally incubated (for 3 h at 30 °C; for 10 min at 65 °C; held at 4 °C), and barcoded fragments ranging from a few to several hundred base pairs were generated. After incubation, the GEMs were broken and the barcoded DNA was recovered. Solid-phase reversible immobilization beads were used to purify and size select fragments for library preparation.

Standard library prep was performed according to the protocol described in the Chromium Genome User Guide Rev A (https://support.10xgenomics.com/permalink/2ofuH1pVbWyeCg2s6u6EwY, last accessed June 10, 2020) to construct one sample-indexed library using 10× Genomics adapters. The final library contained the P5 and P7 primers used in Illumina bridge amplification and was quantified by qPCR. Genomic data were generated using 150-bp paired-end sequencing on an Illumina HiSeq X machine at the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center.

We assembled the "linked-read" HiSeq data using Supernova 2.1.0 assembler (Weisenfeld et al. 2017) using the default recommended settings. Genome-wide statistics were calculated on the total number of phase blocks and the N50 of individual phase block sizes in the pseudohap outputs produced in the Supernova assembly. Statistics about the genome assembly were also ascertained using the stats.py script that is part of the BBMAP suite (Bushnell 2014).

RNA Isolation and Sequencing

To generate RNA-seq data for the submaxillary-gland transcriptome, we extracted total RNA from the submaxillary glands of both shrews using the Qiagen RNeasy Plus Mini Kit following the manufacturer's protocol. RNA concentration and quality were assessed using RNA screen Tapes on a TapeStation 2200 (Agilent Technologies). Poly-A-selected RNA libraries were constructed from total RNA for both shrews using a Kapa mRNA Hyperprep Kit for Illumina platforms (Kapa Biosystems). Final library concentrations and fragment-size distributions were confirmed using a Qubit RNA HS Assay Kit (Invitrogen) and a TapeStation 2200 (Agilent Technologies), respectively. RNAseq data were generated from an average insert size of 330 bp and sequenced using 150-bp paired-end sequencing on an Illumina HiSeq X machine at the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center.

We also used the Oxford Nanopore's MinION Sequencing platform to generate long-range cDNA sequence data. Long-range sequencing of cDNA can be useful for resolving transcripts from highly paralogous venom genes that are difficult to assemble with only short-read sequencing data (Hargreaves and Mulley 2015). cDNA libraries were prepped with Nanopore's cDNA-PCR Sequencing Kit SQK-PCS108 (Oxford Nanopore Technologies, Oxford, UK) for only the WH5 shrew sample due limited amount of RNA. Libraries were prepared using Nanopore's cDNA protocol and suggested enzymes (https://nanopore.yilimart.com/static/images/media/cDNA-PCR%20Sequencing.pdf, last accessed June 10, 2020); LongAmp Taq 2X Master Mix: NEB; RNaseOUT: Thermofisher; SuperScript IV reverse transcriptase, 5× RT buffer, and 100 mM DTT: ThermoFisher Exonuclease 1: NEB. The final library was loaded into a FLO-MIN 106 R9 flowcell and data were collected until almost all of the flowcell's pores were inactive (~10 h).

Transcriptome Assembly and Annotation

We generated three de novo transcriptome assemblies of the submaxillary glands from RNA-seq data from both short-read Illumina data and long-read MinION data. These assemblies included: 1) a de novo assembly for WH4 using only short-read data (WH4-short read); 2) a de novo assembly for WH5 using only short-read data (WH5-short read); and 3) a de novo assembly for WH5 using both short-read data and long-read data from the MinION sequencing run (WH5-short-and-long reads). Long-read data are useful for transcriptome assembly because it can be used to help cluster short-read Illumina data and to resolve transcript paths in the de Bruijn graph during assembly. However, it is important to emphasize that MinION reads do not contribute actual sequence data to the final transcript assembly. Before generating assemblies, we preprocessed the RNA-seq reads using methods adapted from Singhal (2013) and Bi et al. (2012) with some modifications.

Briefly, Trimmomatic (Bolger et al. 2014) was used to trim adapter contamination and low-quality reads. Exact PCR and/or optical duplicate reads were removed using Super-Deduper (https://github.com/dstreett/Super-Deduper, last accessed June 10, 2020). Bowtie2 (Langmead and Salzberg 2012) was used to align the resulting reads against the *Escherichia coli* genome to remove potential bacterial contamination introduced during library preparation or sequencing that might be present in the raw data. Overlapping paired reads were then merged using Flash (Magoč and Salzberg 2011) and their final quality assessed using FastQC (Babraham Bioinformatics). Fast5 files were generated from our MinION sequencing run and were converted to fasta format using the Nanopore's Albacore 1.0.3 basecalling software. Bases whose quality scores were <10 were trimmed using Nanofilt (https://github.com/wdecoster/nanofilt, last accessed June 10, 2020).

All assemblies were generated using the default parameters in Trinity 2.8.5 software (Haas et al. 2013), except for the inclusion of the long-reads parameter (–long_reads) in the WH5-short-and-long reads assembly. We then used the TrinityStats.pl script to assess quality of the transcripts for each of the three assemblies. Next, we used TransDecoder (https://github.com/TransDecoder/TransDecoder/wiki, last accessed June 10, 2020) to identify candidate coding regions within each transcript and filtered by the longest open-reading frame. Assemblies were then annotated using Trinotate v3.2.0 (Bryant et al. 2017) to perform homology searches to known sequence data using BLAST+/SwissProt, to identify protein domains using HMMER v3.3 (http://hmmer.org, last accessed June 10, 2020) against PFAM (El-Gebali et al. 2019), and predicting protein signal peptides and transmembrane domains using SignalP v5.0 (Armenteros et al. 2019). Finally, expression levels for each transcriptome were estimated as transcripts per million (TPM) with Kallisto (Bray et al. 2016).

In an attempt to reduce potential concerns with missing or erroneous transcripts or isoforms from using only one transcriptome assembler, we also generated a merged assembly with de novo assemblies using both Oases 1.2.10 (Schulz et al. 2012) and trans-ABySS 2.0.1 (Robertson et al. 2010) at different k-mer settings and then merged these with our Trinity assembly using the EvidentialGene pipelines (Gilbert 2013; Holding et al. 2018). For the Oases assemblies, we used k-mer values of 31, 51, 71, and 81 and then merged all Oases assemblies with the merged-Assembly tool. For Trans-ABySS, we used k-mer sizes of 32, 52, 72, and 82 and then merged these using the transabyss-merge tool. For each individual data type (WH4-short read, WH5-short read, and WH5-short-and-long reads), we merged the Trinity assemblies with the merged Oases assemblies and the merged Trans-ABySS assemblies to generate a complete transcriptome using the tr2aacds.pl script from the Evidential gene (Evigene) assembly pipeline (Gilbert 2013) after normalizing the transcript names with the trformat.pl script. All assemblies were then assessed for completeness using TransRate (Smith-Unna et al. 2016).

To validate how well our Trinity assembly captured all the transcripts from the submaxillary gland, we assessed the percentage of MinION long-read sequences that were clustered with Trinity transcripts. Using MinION long-read RNA-seq data for this can be a powerful approach to test the completeness of the transcriptome assemblies because each long read is equivalent (or nearly so) to an entire transcript. To investigate the completeness of our Trinity transcriptome assemblies, we conducted a BLAST search of the ~ 500,000 filtered MinION reads against the Trinity-based WH5 (short-and-long reads) assembly. We then saved the best positive BLAST hit for each read with a sequence identity cut-off of 80%. As a consequence, any MinION read that did not have a positive BLAST hit against the Trinity assembly was considered to having been missed by the Trinity assembler. Since the observed error rates of MinION reads are generally between 10% and 15%, we then clustered these remaining MinION reads using CD-Hit-EST (Fu et al. 2012) at a sequence identity of 80% (-c 0.80). We then examined the number of unique clusters to determine whether any cluster was found in high abundance (>100 reads) and could potentially be relevant as an important toxin.

Bioinformatic Pipeline for Identifying Toxin Genes

To identify venom transcripts from our de novo transcriptome assemblies, we applied a pipeline described in Verdes et al. (2016) that filters out transcripts that do not possess signal peptides and do not match any toxins from existing toxin databases. We applied this method to each of the three Trinity de novo transcriptome assemblies. In detail, we filtered our annotated transcriptome to include only transcripts that possessed a signal peptide because this is needed for the transcript to be escorted out of the cell and potentially function as a toxin. Next, the remaining annotated transcripts with signal peptides were searched against a curated database of known toxin protein sequences from Tox-Prot (https://www.uniprot.org/program/Toxins, last accessed June 10, 2020), a subset of the UniProt database, using BlastP from the BLAST+ package (NCBI) with an e-value cutoff of 1×10^{-5}. Then, we filtered these remaining transcripts to only include those that were expressed at >1,000 TPM because these would be most biologically important as potential toxins. Lastly, we considered any of these filtered transcripts as candidate toxins if they were also found in the salivary proteome. We also looked for potential toxin-inhibitor genes by examining transcripts that matched with known venom-protein families, but were not present in the salivary proteome.

Phylogenetic Inference of the KLK1 Gene Subfamily

We investigated the evolutionary history of the kallikrein-1 (KLK1) gene subfamily to understand the origins of the BLTX toxin and other *B. brevicauda* KLK1 paralogs. To do this, we generated a phylogenetic tree containing KLK1 genes that we discovered for *B. brevicauda*, as well as those from other divergent mammal taxa representing Afrotherians, Euarchontoglires, and Laurasiatherians. Alignments were made from these sequences using the MUSCLE (Edgar 2004) plug-in for Geneious v. 11.0.4 (https://www.geneious.com, last accessed June 10, 2020), and Partitionfinder (Lanfear et al. 2012) was used to fit a nucleotide substitution model by partitioning the alignment by codons. We determined using jModeltest 2 (Guindon and Gascuel 2003; Darriba et al. 2012) that the general time-reversible model with invariable sites and a gamma-shaped distribution (GTR + I + G) (Tavaré 1986) was the best-fitting nucleotide-substitution model for our phylogenetic analyses. Maximum likelihood trees were inferred with rapid bootstrapping method (100 bootstraps) in RAxML (Stamatakis 2014), as well as with MrBayes (Ronquist and Huelsenbeck 2003). We estimated overall mean distance of KLK1 sequences in MEGAX (Kumar et al. 2018) to verify that our alignment was reliable for phylogenetic analyses (Thompson et al. 1999).

Selection Tests

To investigate whether toxin genes and toxin-related genes in *B. brevicauda* are rapidly evolving and if this differed from the type of selection on these genes across mammals, we conducted both site model and branch-site model tests (Yang and Swanson 2002; Zhang et al. 2005) using the CODEML program (Yang 2007) implemented in EasyCodeML (Gao et al. 2019). Selection tests were performed on known toxin genes, candidate-toxin genes, paralogous KLK1 genes, and candidate-inhibitor genes. We first downloaded one-to-one orthologs from across divergent mammal clades from Ensembl (Frankish et al. 2018) and generated alignments using MUSCLE (Edgar 2004) in Geneious v. 11.0.4 (https://www.geneious.com, last accessed June 10, 2020). Alignments were processed with Trimal with the -automated1 flag (Silla-Martínez et al. 2009) to remove spurious alignments and large uninformative gaps. Phylogenetic trees using maximum likelihood were then inferred using the GTR + I + G nucleotide model with 100 bootstraps in RAxML (Stamatakis 2014). We then ran the site-model test, which allowed the ω ratio (measure of natural selection acting on a protein) to vary among all sites across all mammalian lineages. For this test, we compared the M8 alternative model (beta and ω_S) for 11 different site classes against the null M8a model (beta and $\omega_S = 1$). Next, branch-site models were tested to identify episodic positive selection acting within the *B. brevicauda* lineage (foreground branch). This test compared the MA alternative model ($\omega_{FG}>1$ in the foreground branch) against the MA null model ($\omega_{FG} = 1$). Likelihood ratio tests were used to compare null and alternative models, approximating the statistic to a chi-squared distribution (P value = 0.05) with degrees of freedom equal to the difference in the number of parameters of the compared models. For genes statistically under positive selection, we considered sites under positive selection with a Bayes Empirical Bayes (BEB) posterior probability > 0.95.

Characterization of the Salivary Proteome

To prepare both WH4 and WH5 samples of shrew saliva for peptide sequencing, we soaked each salivary sample on the medical tubing in a 50 mM ammonium bicarbonate solution inside an Eppendorf tube and pipette washed this 20 times. The ammonium bicarbonate solution was then removed and placed in a separate Eppendorf tube. The wash procedure was repeated two times, after which the ammonium bicarbonate solution was pooled together and concentrated in a speed vacuum to a final volume of ~100 μl for digestion. We then measured 200 μl of 50 mM ammonium concentration of proteins using a Qubit Protein Assay Kit (Invitrogen). Each saliva sample was then digested by adding 5 μl of DTT (5 μg/μl in 50 mM ammonium bicarbonate) and incubated at 56 °C for 15 min. Next, 5 μl of iodoacetamide (15 mg/ml in 50 mM ammonium bicarbonate) was added to each sample and then incubated them in the dark at room temperature (23 °C) for 30 min. Following this, sequencing grade-modified trypsin (Promega, Madison WI) prepared in 50 mM ammonium bicarbonate was added to each sample with an estimation of 1:50 enzyme–substrate ratio and the reaction was carried out at 37 °C for overnight. After the digestion, acetic acid was added to the sample to quench the reaction. The samples were dried in a vacufuge and resuspended in 20 μl of 50 mM acetic acid. Peptide concentration was determined by nanodrop (A280nm).

Liquid chromatography-nanospray tandem mass spectrometry (LC/MS/MS) for protein identification was performed on a Thermo Scientific orbitrap Fusion mass spectrometer equipped with an EASY-Spray Source and operated in positive ion mode. Samples were separated on an easy spray nano column (Pepmap RSLC, C18 3 µ 100 A, 75 µm X250mm Thermo Scientific) using a 2D RSLC HPLC system from Thermo Scientific. Each sample was injected (2 μg) into the µ-Precolumn Cartridge (Thermo Scientific) and desalted with 0.1% Formic Acid in water for 5 min. Mobile phase A was 0.1% formic acid in water and acetonitrile (with 0.1% formic acid) was used as mobile phase B. Mobile phase B was increased from 2% to 20% in 40 min and then increased from 20% to 32% in 10 min and again from 32% to 95% in 6 min and then kept at 95% for another 2 min before being brought back quickly to 2% in 2 min. The column was equilibrated at 2% of mobile phase B (or 98% A) for 15 min before the next sample injection. MS/MS data were acquired with a spray voltage of 1.7 kV and a capillary temperature of 275 °C

is used. The scan sequence of the mass spectrometer was as follows: the analysis was programmed for a full scan recorded between m/z 375–1,700 and an MS/MS scan to generate product ion spectra to determine amino acid sequence in consecutive scans starting from the most abundant peaks in the spectrum in the next 3 s. To achieve high mass accuracy MS determination, the full scan was performed at FT mode and the resolution was set at 120,000. The AGC Target ion number for FT full scan was set at 4×105 ions, and maximum ion injection time was set at 50 ms. MSn was performed using ion trap mode to ensure the highest signal intensity of MSn spectra using CID (for 2+ to 7+ charges). The AGC Target ion number for ion trap MSn scan was set at 1× 104 ions, and maximum ion injection time was set at 30 ms. The CID fragmentation energy was set to 30%. Dynamic exclusion is enabled with a repeat count of 1 within 60 s and a low mass width and high mass width of 10 ppm.

Mgf files were searched using Mascot Daemon by Matrix Science version 2.3.2 (Boston, MA) and the database searched against a custom database comprised all translatable ORFs from both the WH4 and WH5 transcriptome assemblies, in order to have the highest probability of identifying a potential protein. The mass accuracy of the precursor ions was set to 10 ppm, and accidental inclusion of 1 13 C peaks was also included into the search. The fragment mass tolerance was set to 0.5 Da. Considered variable modifications were oxidation (Met), deamidation (N and Q), acetylation (K), and carbamidomethylation (Cys) was set as a fixed modification. Four missed cleavages for the enzyme were permitted. A decoy database was also searched to determine the false discovery rate (FDR) and peptides were filtered according to the FDR. Proteins with <1% FDR as well as a minimal of two significant peptides detected were considered as valid proteins. Proteins were annotated by performing homology searches to our WH4 and WH5 transcriptomes and to a human database using BlastP.

Protein Modeling and Docking of the KLK1 Gene Subfamily

Due to previous work showing that substitutions in regulatory loops surrounding the catalytic cleft of BLTX facilitates its toxicity (Aminetzach et al. 2009), we 3D modeled the newly discovered KLK1 paralogs to investigate their electrostatic similarity to the known toxin paralog. Input protein sequences were trimmed of their signal peptides and fed into the homology-based protein folding, online server, MUSTER (Wu and Zhang 2008). We visualized predicted protein models and examined the electrostatic potential of surface residues using the APBS electrostatics plug-in in PyMOL (Schrödinger, LLC 2015). We also modeled a Serine Protease Inhibitor that was found to be the most highly expressed gene in our transcriptome with the same approach as the KLK1's with the intention of modeling protein–protein interactions between this inhibitor and BLTX. We modeled the potential interaction of BLTX and the Serine Protease Inhibitor using ClusPro by treating BLTX as the receptor and the double-headed inhibitor as the ligand (Comeau et al. 2004). Results from this prediction were visualized using PyMOL.

Results

De Novo Reference Genome

We assembled a 1.66×Gb genome at 31× effective coverage from 3.6×10^8 reads for one individual shrew (WH4) using Supernova 2.1.0. Supernova estimated the total genome size of this sample to be 2.52 Gb. The assembly contained 166,552 scaffolds with a scaffold N50 of 0.34 Mb, of which 16,115 scaffolds were >10 kb in length. The 10× Genomic Library used for this assembly had a weighted-mean molecule size of 20.32 kb.

Transcriptome Assemblies of the Submaxillary Gland

We found relatively similar numbers of transcripts between the two Trinity transcriptome assemblies using only short-read Illumina data and the one Trinity assembly that combined short-read data with long-read MinION data. The two short-read-based transcriptomes contained 32,893 transcripts from 12.2×10^6 reads for the WH4-short read assembly and 30,587 transcripts from 12.7×10^6 reads for the WH5-short read assembly (supplementary table S1, Supplementary Material online). The combined short-read and long-read transcriptome (WH5-short-and-long reads) contained 30,719 transcripts from 12.7×10^6 short-reads and 5.1×10^5 postfiltered MinION reads.

Validation of Trinity Assemblies

The MinION cDNA long-read data had a couple of important effects on our final transcriptome of the submaxillary gland. First, the Trinity assembly using both Illumina short reads and MinION long reads (WH5-short-and-long reads) had substantially fewer transcripts ($n = 171$) than the Trinity assembly using only short-read data for the same WH5 individual ($n = 583$) and for the WH4 ($n = 642$) individual (fig. 1 and supplementary table S1, Supplementary Material online). A majority of the unique transcripts found in the WH5-short read assembly ($n = 227$) and the WH4-short read assembly ($n = 291$) were, on an average, lowly expressed (14.9 TPM ± 136.7 SD and 10.0 TPM ± 56.5 SD, respectively). Similarly, the shared transcripts between the WH5-short read assembly and the WH4-short read assembly ($n = 204$) were also, on an average, lowly expressed (35.3 TPM ± 322.6 SD). As a result, the addition of the MinION data in the Trinity assembly appeared to have removed many low copy transcripts.

The MinION long-read cDNA sequences represent full-length transcripts and were also used to detect additional

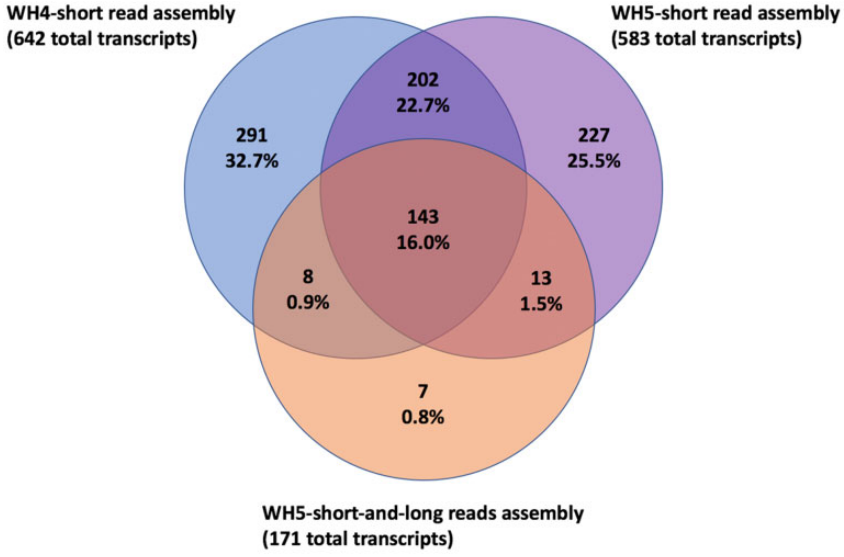

Fig. 1.—Venn diagram displaying the overlapped and unique transcripts with signal peptides among the three transcriptome assemblies.

transcripts that were not assembled by Trinity. After performing BLAST searches of our MinION-long reads (504,388) against the WH5-short-and-long reads transcriptome assembly, we found that 498,251 long-read transcripts (~98.8%) matched with an assembled transcript, leaving 6,137 MinION reads without a match. After clustering the remaining MinION reads to determine if any remaining long-read transcripts were at high abundance, we found 5,181 unique clusters. However, none of these unique clusters was higher in abundance than 51 reads (supplementary table S2, Supplementary Material online). Thus, we conclude that we have strong evidence that we did not miss any biologically important transcripts from our Trinity assembly.

Identification of Known and Candidate Venom Genes from Submaxillary-Gland Transcriptomes

Our bioinformatic search for toxin proteins from all three transcriptome assemblies resulted in a total of 13 highly expressed genes (>1,000 TPM) that had BlastP hits with known toxins from the Tox-Prot database (table 1 and supplementary table S4, Supplementary Material online). These 13 sequences consisted of *kallikrein-1* (*KLK1-BL1*), *Blarina toxin* (*BLTX*), *Blarinasin-1* (*Blarinasin*), *Blarinasin-2* (*Blarinasin*), a novel *kallikrein-1 paralog* (*KLK1-BL2*), *proenkephalin-A* (*PENK*: containing the Soricidin peptide), *phospholipase A2 group 1B* (*PA21B*), *cholecystokinin* (*CCKN*), *antileukoproteinase* (*SLPI*), *cystatin-M* (*CYTM*), *WAP four-disulfide core domain protein 2* (*WFDC2*), *endothelin-1* (*EDN1*), and a *double-headed protease inhibitor* (*IPSG*). All of these transcripts met the following criteria: contained a signal peptide, had homologous sequences with known toxins from the Tox-Prot database, and were expressed at >1,000 TPM in the submaxillary gland (see supplementary table S1, Supplementary Material online, for filtering of transcripts through each step of the pipeline). No additional candidate toxins were found from our merged assemblies using Evigene (supplementary table S4, Supplementary Material online).

Genomic Arrangement of KLK1 Paralogs

Using a combination of the new de novo genome assembly and the new de novo transcriptome assemblies, we identified three novel KLK1 serine-protease paralogs (*KLK1-BL1*, *KLK1-BL2*, and *KLK1-BL3*) that are in tandem array with two previously identified KLK1 paralogs (*BLTX* toxin and its nontoxic paralogs *Blarinasin-1* and *Blarinasin-2*) (fig. 2*A*). We followed the nomenclature recommendations for Kallikrein genes by Olsson et al. (2004) for naming these newly discovered paralogs. *KLK1-BL1*, *KLK1-BL2*, *BLTX*, and both *Blarinasins* were expressed at relatively high levels in the submaxillary gland (fig. 2*A*). *KLK1-BL3* was not expressed in the submaxillary gland. In addition, we were only able to find *Blarinasin-1* in our reference genome, but not *Blariansin-2*. *Blarinasin-1* and *Blariansin-2* have very high-sequence similarity (98.9%) and were expressed at relatively high levels in our transcriptome assemblies. We suspect that *Blarinasin-1* and *Blarinasin-2* might be different alleles of the same gene and that only one of these alleles is apparent in the reference genome. Alternatively, *Blarainsin-2* may also be found in another part of the genome that we were not able to assemble.

Phylogenetic Inference of KLK1

Our phylogenetic tree of the KLK1 gene subfamily revealed there were two ancient duplications of the KLK1 gene prior to the divergence between *B. brevicauda* and *Sorex araneus*,

Table 1
Gene Expression Profile (TPM) of Three Transcriptomes of the Submaxillary Gland

		TPM		
Gene	Protein	Transcriptome WH5-Short-and-Long Reads	Transcriptome WH5-Short Reads	Transcriptome WH4-Short Reads
IPSG	**Double-headed protease inhibitor**	**46,625**	**109,204**	**50,910**
FLP	Female-specific lacrimal-gland protein	40,179	N/A	43,352
BPIA2	BPI fold-containing family A member 2	26,562	23,284	11,843
MUC7	Mucin-7	21,608	16,780	19,360
PA21B	**Phospholipase A2 group 1B**	**15,099**	**15,955**	**10,101**
PENK	**Proenkephalin-A (contains Soricidin peptide)**	**10,477**	**16,591**	**8,985**
KLK1	**Blarinasin-2**	**9,605**	**9,396**	**2,742**
CCKN	**Cholecystokinin**	**9,454**	**7,223**	**3,044**
TKN1	Protachykinin-1	8,777	6,511	1,898
SPL2B	Short palate, lung, and nasal epithelium carcinoma-associated protein 2B	8,495	N/A	2,733
CAH6	Carbonic anhydrase 6	7,759	6,242	9,539
PIP	Prolactin-inducible protein homolog	7,417	5,904	13
SLPI	**Antileukoproteinase**	**6,897**	**3,906**	**N/A**
BLTX	**Blarina toxin**	**6,884**	**5,318**	**1,580**
KLK1-BL1	**Blarina Kallikrein-1**	**5,581**	**3,141**	**1,809**
KLK1-BL2	**Blarina Kallikrein-1**	**5,553**	**1,532**	**3,703**
EDN1	**Endothelin-1**	**5,243**	**2,002**	**5,891**
KLK1	**Blarinasin-1**	**4,813**	**N/A**	**824**
CYTM	**Cystatin-M**	**2,262**	**2.521**	**980**
OPRPN	Opiorphin prepropeptide	2,065	N/A	2,678
MUC19	Mucin-19	1,840	4,599	7,870
RNAS7	Ribonuclease 7	1,253	770	438
WFDC2	**WAP four-disulfide core domain protein 2**	**1,111**	**879**	**1,869**

Note.—Shown in bold are genes with BlastP hits to known toxins from the Tox-Prot database. List order is based on the 22 highest expressed genes from the WH5-short-and-long reads assembly.

common shrew (fig. 2*B*). This tree also showed that the KLK1 genes in other Eulipotyphlans (hedgehog, star-nosed mole, and Solenodon) duplicated in each of those lineages, but these duplications were independent of duplications that occurred in the ancestor of Blarina and Sorex. Low support values for nodes leading to divergence between *BLTX*, *Blarinasin-1*, *Blarinasin-2*, and the *Sor. araneus KLK1* XM_012935294.1 ortholog indicate that their order of divergence is unclear. The average amino acid identity of our alignment of the KLK1 gene subfamily was 0.60, which corresponds to a 40% identity and was above the 30% identity required for reliable alignments for phylogenetic analysis (supplementary table S5 and fig. S1, Supplementary Material online).

Positive Selection

Of the 16 transcripts that had high expression and BlastP hits to the Tox-Prot database, 11 were found to have undergone positive selection using the branch-site model tests in CODEML (table 2). All Kallikrein-1 paralogs, including the known toxin BLTX, were undergoing positive selection (fig. 2*C*). This included KLK1-BL3 paralog that we discovered in the reference genome, but was not expressed in the submaxillary-gland transcriptome or present in the salivary proteome. Of these 11 genes that have experienced positive selection, four of them (*Blarinasin-1*, *Blarinasin-2*, *PA21B*, and *SLPI*) only had weak evidence for positive selection at any particular site, whereas the other seven genes had statistical significance (>0.5 Bayesian posterior probability) of at least one site undergoing positive selection. Furthermore, of the 11 genes under positive selection using the branch-site model test, five were not undergoing positive selection among the mammal lineages (*KLK1-BL1*, *Blarinasin-1*, *Blarinasin-2*, *KLK1-BL2*, and *PENK*) according to the site-model test (table 2). We also did not detect positive selection in the tissue factor pathway inhibitor 2 (TFPI2) gene (table 2). This gene was not expressed at high levels in the submaxillary gland, but was one of the more abundant proteins in the saliva and has homology to known venom genes.

Abundance of Venom Proteins in the Salivary Proteome

Of all the transcripts that were highly expressed in the submaxillary gland and that had BlastP hits to the Tox-Prot database, only five were found to also be at high abundance

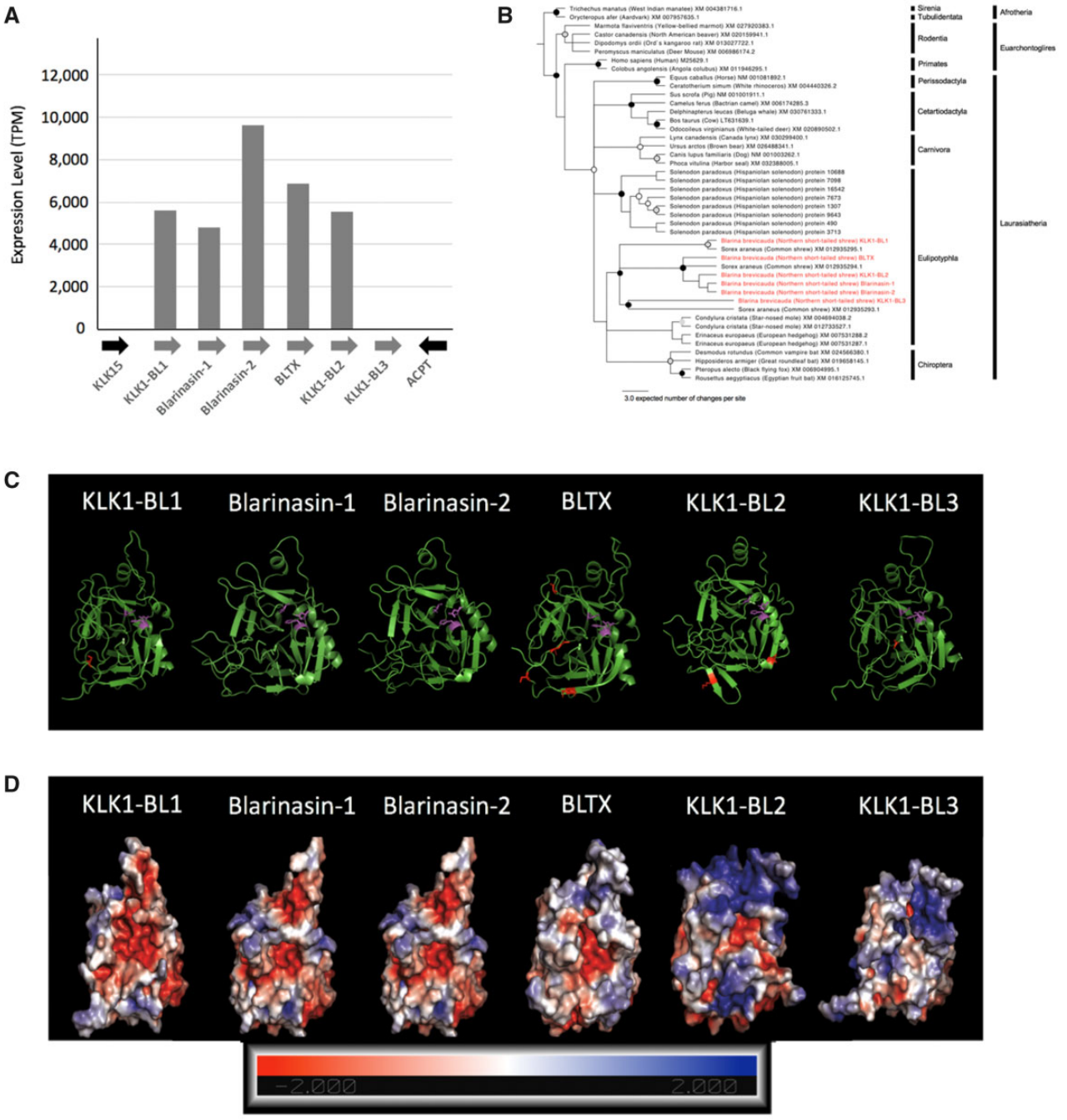

Fig. 2.—(*A*) Genomic orientation of KLK1-like tandem array in *Blarina brevicauda*. Flanking genes *(KLK15* and *ACPT)* are shown along with expression values in TPM for each *Blarina* KLK1 paralog in the WH5-short-and-long reads transcriptome assembly. (*B*) Phylogenetic reconstruction of KLK1 sequences from *B. brevicauda* and other mammalian taxa. Node labels indicate maximum-likelihood bootstrap support and Bayesian posterior probability (pp) support with black circles indicating 100% bootstrap support and 1.00 pp support, gray circles with black outline indicating 75–99% bootstrap support and 1.00 pp support, gray circles with no black outline indicating 75–99% bootstrap support and 0.95–0.99 pp support, and white circles with black outline indicating <75% bootstrap support and 0.95–0.99 pp support. Red labels indicate *B. brevicauda* KLK1 genes. (*C*) Ribbon diagrams for all five *B. brevicauda* KLK1 paralogs showing sites undergoing positive selection (red) and the catalytic triad Asp-His-Ser (purple). (*D*) Electrostatic potential of modeled surface residues for all five *B. brevicauda* KLK1 paralogs. Red indicates a more negative electric potential, whereas blue indicates a more positive electric potential.

Table 2

Results for Site Model and Branch-Site Model (*Blarina brevicauda* as Foreground Lineage) Tests Using CODEML

		Site Model (Mammals)		Branch-Site Model (Blarina Lineage)		
Gene (Protein Name)	**d*N*/d*S***	**M8 versus M8a 2Δln l (*P* Value, df)**	**M8 Selection Parameters**	**MA versus MA Null 2Δln l (*P* Value, df)**	**MA Selection Parameters**	**MA BEB No. of Positively Selected Sites (pp>0.95)**
KLK1-BL1 (Blarina kallikrein-1)	0.37	1.99 (0.16, 1)	$p_S = NA, \omega_S = NA$	**9.58 (<0.01, 1)**	**$p_S = 0.03, \omega_{FG} = 502.7$**	**1**
KLK1 (Blarinasin-1)	0.37	2.90 (0.09, 1)	$p_S = NA, \omega_S = NA$	**7.30 (<0.01, 1)**	**$p_S = 0.03, \omega_{FG} = 41.90$**	**0**
KLK1 (Blarinasin-2)	0.38	2.48 (0.12, 1)	$p_S = NA, \omega_S = NA$	**6.37 (0.01, 1)**	**$p_S = 0.03, \omega_{FG} = 38.72$**	**0**
BLTX (Blarina Toxin)	0.37	**5.21 (0.02, 1)**	**$p_S = 0.18, \omega_S = 1.51$**	**14.31 (<0.01, 1) ω_{FG} = 15.98**	**$p_S = 0.04, \omega_{FG} = 15.98$**	**4**
KLK1-BL2 (Blarina kallikrein-1)	0.37	0.55 (0.45, 1)	$p_S = NA, \omega_S = NA$	**21.83 (<0.01, 1)**	**$p_S = 0.05, \omega_{FG} = 331.4$**	**2**
KLK1-BL3 (Blarina kallikrein-1)	0.39	**4.92 (0.03, 1)**	**$p_S = 0.11, \omega_S = 1.69$**	**3.99 (0.05, 1)**	**$p_S = 0.05, \omega_{FG} = 10.38$**	**1**
PENK (Soricidin)	0.17	0.42 (0.51, 1)	$p_S = NA, \omega_S = NA$	**10.63 (<0.01, 1)**	**$p_S = 0.02, \omega_{FG} = 6.51$**	**2**
PA21B (Phospholipase A2 group 1B)	0.16	**3.92 (0.05, 1)**	**$p_S = 0.01, \omega_S = 4.24$**	**3.71 (0.05, 1)**	**$p_S = 0.03, \omega_{FG} = 3.35$**	**0**
SLPI (Antileukoproteinase)	0.59	**16.93 (< 0.01, 1)**	**$p_S = 0.34, \omega_S = 1.68$**	**5.63 (0.02, 1)**	**$p_S = 0.03, \omega_{FG} = 73.21$**	**0**
HYALP (Hyaluronidase PH-20)	0.58	**123.54 (< 0.01, 1)**	**$p_S = 0.37, \omega_S = 2.01$**	**4.15 (0.04, 1)**	**$p_S = 0.02, \omega_{FG} = 6.42$**	**1**
IPSG (double-headed serine protease inhibitor)	0.66	**48.67 (< 0.01, 1)**	**$p_S = 0.21, \omega_S = 2.36$**	**8.14 (<0.01, 1)**	**$p_S = 0.03, \omega_{FG} = 999.0$**	**1**
WFDC2 (WAP four-disulfide core domain protein 2)	0.21	**10.89 (< 0.01, 1)**	**$p_S = 0.06, \omega_S = 2.01$**	<0.01 (0.99, 1)	$p_S = NA, \omega_{FG} = NA$	NA
EDN1 (endothelin-1)	0.32	0.14 (0.71, 1)	$p_S = NA, \omega_S = NA$	0.00 (1.00, 1)	$p_S = NA, \omega_{FG} = NA$	NA
CYTM (cystatin-M)	0.19	0.03 (0.86, 1)	$p_S = NA, \omega_S = NA$	1.04 (0.31, 1)	$p_S = NA, \omega_{FG} = NA$	NA
CCKN (Cholecystokinin)	0.17	0.23 (0.63, 1)	$p_s = NA, \omega_S = NA$	0.00 (1.00, 1)	$p_S = NA, \omega_{FG} = NA$	NA
TFPI2 (Tissue factor pathway inhibitor 2)	0.30	2.06 (0.26,1)	$p_s = NA, \omega_S = NA$	0.00 (1.00, 1)	$p_S = NA, \omega_{FG} = NA$	NA

NOTE.—Shown in bold are tests with significant *P* values ≤0.05. d*N*/d*S*, ratio averaged across all sites and lineages; p_S, proportion of sites estimated to be under positive selection with $\omega_S > 1$ and $\omega_{FG} > 1$; NA, not applicable; BEB, Bayes Empirical Bayes analysis.

Table 3

Most Abundant Proteins (≥1.0% Relative Abundance) from the Proteomes of Both WH4 and WH5 Individuals

WH4 Saliva Proteome			WH5 Saliva Proteome		
Gene Name	**Protein Name**	**Relative Abundance (%)**	**Gene Name**	**Protein Name**	**Relative Abundance (%)**
SLPI	**Antileukoproteinase**	**20.0**	**SLPI**	**Antileukoproteinase**	**10.2**
BLTX	**Blarina toxin**	**11.3**	PLXNB3	Plexin-B3	8.8
KLK1-BL2	**Kallikrein-1**	**10.6**	**KLK1-BL2**	**Kallikrein-1**	**8.5**
PA21B	**Phospholipase A2 group 1B**	**9.4**	**BLTX**	**Blarina toxin**	**7.6**
KLK1	Blarinasin	8.5	KLK1	Blarinasin	7.6
EGF	Pro-epidermal growth factor	3.6	**PA21B**	**Phospholipase A2 group 1B**	**7.2**
TFPI2	**Tissue factor pathway inhibitor 2**	**3.2**	EGF	Pro-epidermal growth factor	4.6
OBP	Odorant-binding protein	2.3	BPIFA1	BPI fold-containing family A	2.9
ALB	Serum albumin	2.2	**TFPI2**	**Tissue factor pathway inhibitor 2**	**2.7**
PLXNB3	Plexin-B3	2.1	CSF2	Granulocyte-macrophage colony-stimulating factor	2.7
SORI	**Soricidin (peptide from the Proenkephalin-A gene)**	**1.8**	**SORI**	**Soricidin (peptide from Proenkephalin-A gene)**	**2.6**
ZG16	Zymogen granule protein 16	1.1	**HYALP**	**Hyaluronidase PH-20**	**2.0**
KRT10	Keratin, type I cytoskeletal 10	1.0	OBP	Odorant-binding protein	2.0
			ALB	Serum albumin	1.7
			KRT10	Keratin, type I cytoskeletal 10	1.6
			ZG16	Zymogen granule protein 16	1.4
			RNASE7	Ribonuclease 7 like	1.2
			SCGB2A2	Secretoglobin family 2A member 2-like	1.2
			CAH6	Carbonic anhydrase 6	1.0

NOTE.—Shown in bold are known or candidate-toxin genes.

Table 4

Most Abundant Salivary Proteins from *Blarina brevicauda* (≥1.0% Relative Abundance) and Functions of Their Homologs

Protein Name	Protein Function (Reference)
Antileukoproteinase	**Antimicrobial—inhibitor of serine proteinases (Amerongen and Veerman 2002)**
BPI fold-containing family A	Antimicrobial (Jönsson 2018)
Carbonic anhydrase 6	Taste perception (Amerongen and Veerman 2002)
Granulocyte-macrophage colony-stimulating factor	Antimicrobial (Brasil et al. 2012)
Hyaluronidase PH-20	**Spreading factors—promotes the diffusion of toxins (Tu and Hendon 1983; Bordon et al. 2015)**
Kallikrein-1 (BLTX)	**Vasodilation and cleavage of bradykinin—serine proteinase (Kita et al. 2004)[a]**
Kallikrein-1 (Blarinasin)	Cleavage of bradykinin—serine proteinase (Kita et al. 2005)[a]
Kallikrein-1 (KLK1-BL2)	**Cleavage of bradykinin—serine proteinase (Blanchard et al. 2015)**
Keratin, type I cytoskeletal 10	Epidermal barrier (Fischer et al. 2016)
Odorant-binding protein	Delivery and perception of odiferous molecules (Tegoni et al. 2000)
Phospholipase A2 group 1B	**Myotoxic, neurotoxic, anticoagulant (Fry et al. 2009)**
Plexin-B3	Neurogenesis (Artigiani et al. 2004)
Pro-epidermal growth factor	Tissue generation (Brasil et al. 2012)
Ribonuclease 7 like	Antimicrobial (Huang et al. 2007)
Secretoglobin family 2A member 2-like	Anti-inflammatory and mate selection (Jackson et al. 2011; Chung et al. 2017)
Serum albumin	Lubrication of oral tissues (Hatton et al. 1985)
Soricidin (peptide from the Proenkephalin-A gene)	**Paralysis—inhibits calcium channel activity (US Patent No.: US8003754B2)[a]**
Tissue factor pathway inhibitor 2	**Inhibitor of blood coagulation—kunitz-type serine protease inhibitor (Wood et al. 2014)**
Zymogen granule protein 16	Host defense—binds to *Staphylococcus aureus* (Heo et al. 2013)

NOTE.—Shown in bold are known or candidate toxins in *B. brevicauda* venom.

[a]Indicates previous studies that have extracted *B. brevicauda* proteins and performed functional tests for toxicity.

(>1.0% of total abundance) in the salivary proteome (SLPI, BLTX, KLK1-BL2, PA21B, and PENK) and were thus considered as known or candidate toxins (tables 3 and 4; supplementary table S6, Supplementary Material online). Our LC/MS analysis of the saliva yielded a total of 122 unique proteins with 109 proteins having at least two unique identifying spectra from the WH4 individual and 83 proteins from the WH5 individual. Of the 122 total proteins, 71 were shared between the two shrew individuals (supplementary table S6, Supplementary Material online). The most abundant protein in both shrew samples was Antileukoproteinase (SLPI). The previously known toxin BLTX (Kita et al. 2004) was the second most abundant protein (11.3% of the total protein abundance) in WH4's salivary proteome and the fourth most abundant (7.6%) in WH5's salivary proteome (table 3). The candidate toxin KLK1-BL2 comprised 10.6% of all salivary proteins in WH4 and 8.5% in WH5. Phospholipase A2 group 1B (PA21B), the newly identified candidate toxin, was found to be the fourth most abundant salivary protein (9.4%) in WH4 and the sixth most abundant protein (7.2%) in WH5. Proenkephalin (PENK), which contains the known toxin peptide Soricidin, was found to be 1.8% of all salivary proteins in WH4 and 2.6% in WH5.

We also identified two additional salivary proteins that were not expressed at high levels in the transcriptome, but were relatively abundant in the salivary proteome and are possible constituents of venom because they had BlastP matches with the Tox-Prot database. The Hyaluronidase PH-20 protein (HYALP) was relatively abundant (2.0%) in WH5 and less abundant in (0.6%) in WH4 (table 3 and supplementary table S6, Supplementary Material online). This protein is a nontoxin, but has been shown to be an important spreading factor for toxins in venomous lizards (table 4; Tu and Hendon 1983). The TFPI2 protein was present in moderate levels of abundance (3.2%) in WH4 and (2.7%) in WH5 (table 3) and is an important inhibitor of blood coagulation (tables 3 and 4; Wood et al. 2014).

The nontoxic Blarinasin (Kita et al. 2005) was also present in high abundance (8.5%) in WH4 and (7.6%) in WH5. However, we were not able to distinguish from the peptide sequences of Blarinasin whether they came from Blarinasin-1 or Blarinasin-2. In addition, despite being highly expressed in the transcriptome of the submaxillary gland, KLK1-BL1, Endothelin-1, and the double-headed serine protease inhibitor were either not found or found at very low levels in the salivary proteome (supplementary table S6, Supplementary Material online).

3D-Protein Modeling

We found evidence from the 3D protein structural modeling that of the three newly identified *B. brevicauda* KLK1 paralogs, KLK1-BL2 is possibly a toxin. The other two KLK1 paralogs (KLK1-BL1 and KLK1-BL3) are likely not toxic. We suspect KLK1-BL2 may be similar toxin as the known BLTX toxin because it has similar positively charged regulatory loops surrounding a negatively charged catalytic pocket. In contrast, the 3D protein modeling for KLK1-BL3 shows only positive residues in the regulatory loop, but not a negatively charged pocket (fig. 2*D*). In addition, the 3D structure of KLK1-BL1 is similar

to both of the nontoxic Blarinasins in having negatively charges in both the pocket and the surrounding regulatory loops (fig. 2*D*).

The presence of a highly expressed double-headed serine-protease inhibitor gene in the submaxillary gland and its potential to impact the function of BLTX (a serine protease) prompted us to investigate the potential for protein–protein interactions between these two using ClusPro. When treating BLTX as the receptor and the inhibitor as the ligand, ClusPro predicted docking of the protease inhibitor in the active site of BLTX (supplementary fig. S2, Supplementary Material online). Specifically, the protease-binding loop of the inhibitor bound to BLTX with the reactive P_1–P_1' site of the inhibitor directly positioned in the active site of BLTX. This configuration has been found as the general mode of inhibition for many serine protease inhibitors, suggesting the possibility that this inhibitor does interact with BLTX in an inhibitory manner inside only the submaxillary gland since it is found at a relatively low level within the saliva.

Discussion

Using a multiomic approach with analyses of molecular evolution, we show that *B. brevicauda* venom contains a relatively simple mixture of toxin-related genes. These include two toxins that have been previously identified with functional assays for toxic activity on either mice or invertebrates (BLTX and Soricidin) and five proteins that are newly described as candidate-toxin constituents (KLK1-BL2, PA21B, and SLPI, HYALP, TFPI2). We also discovered three additional KLK1 paralogs in tandem array in the genome with BLTX and Blarinasin-1. Interestingly, all five KLK1 paralogs are undergoing positive selection, even though three of them are unlikely to be toxic. Finally, we have identified three proteins that were highly expressed in the submaxillary gland, but not abundant in the saliva, and that may act as endogenous self-defense mechanisms to help ameliorate the toxic effects of the main venom component BLTX.

Known and Candidate Venom Components

Consistent with previous work (Kita et al. 2004; Aminetzach et al. 2009), the *BLTX* gene was found to be highly expressed in the submaxillary gland and undergoing positive selection. We also found it to be one of the most abundant toxin proteins in the saliva. A previous study has shown that this protein is lethal to mice and decreases blood pressure through the cleaving of kinins to bradykinin (Kita et al. 2004). Additional work has shown that this gene to be evolving under positive selection that is acting on lineage-specific insertions in and around the regulatory loops surrounding its catalytic pocket, which has led to increased substrate specificity for this protein relative to its KLK1 derivative (Aminetzach et al. 2009). Using a branch-site selection analyses, we also found evidence of positive selection that is occurring at four sites.

Using a combination of transcriptomic, proteomic, and a de novo reference genome, we have identified a candidate-toxin gene, *KLK1-BL2*, that belongs to the same kallikrein-1 gene subfamily as *BLTX*. *KLK1-BL2* was highly expressed in the submaxillary gland, was relatively abundant in the salivary proteome, and is undergoing rapid evolution. One of the strongest pieces of evidence supporting this paralog as a toxin is the similarity in electrostatic potential with BLTX. Further functional analysis or lethality assays are needed to assess the validity of this protein as a venom component.

We also found another previously identified toxin, Soricidin, to be highly expressed in the submaxillary gland, present in the saliva, and undergoing rapid evolution in *B. brevicauda*, but not across mammals. Soricidin is a small peptide from the proenkephalin gene, which was previously isolated from *B. brevicauda* and found to be highly effective at immobilizing mealworms (US Patent No.: US8003754B2). The effect on mealworms is consistent with the hypothesis that *B. brevicauda* venom is used to cache and immobilize invertebrate prey for long periods of time. Proenkaphalin is a precursor gene that undergoes post-translational cleavage resulting in multiple enkaphalins, which are short peptides involved in opioid receptor signaling (Henry et al. 2017). Proenkaphalins have been implicated as toxic components in both scorpion and fangblenny venom, where they exhibit hypotensive activity (Zhang et al. 2012; Casewell et al. 2017). Specifically, Soricidin has been shown to have high affinity for TRPV6 Calcium ion channels, and is capable of inhibiting the movement of calcium across the cellular membrane (Bowen et al. 2013).

We also discovered the candidate toxin, PA21B protein, in *B. brevicauda*'s venom. We suspect this may contribute to venom because it is highly expressed in the submaxillary gland, produced at high levels as a salivary protein, and is undergoing rapid evolution in *B. brevicauda*, as well as across mammals. Furthermore, phospholipase A2 enzymes "PLA2s" are common animal toxins that have been convergently recruited into venom arsenals across the animal kingdom (Fry and Wüster 2004) and depending on the specific venom system can have drastically different pharmacological effects including: neurotoxic, myotoxic, inflammatory, and hemolytic activities (Kordiš 2011). PLA2 has the general property of hydrolyzing phospholipids and is an important pancreatic enzyme, and thus plays an important role in lipid metabolism (Arni and Ward 1996). PLA2 paralogs and isoforms can have highly variable function and levels of toxicity even within the same venomous species/organism (Harris and Scott-Davey 2013). Interestingly, recent proteomic work from the saliva *Neomys fodiens*, an unrelated venomous-shrew species, revealed a peptide with homology to PLA2, which was speculated to contribute to the paralytic effects observed in *N. fodiens* venom (Kowalski et al. 2017). To our knowledge, the presence of PLA2 in the saliva of both *Neomys* and *Blarina*

may represent the first example of recruitment of these proteins into mammalian saliva, and potentially the first example of convergence of venom toxins in eulipotyphlans. Moreover, the extreme variability in PLA2 function across venom systems makes it difficult to suggest an exact function for PLA2 within the *B. brevicauda* venom system, but nonetheless it is a likely candidate venom gene. Further proteomic isolation and functional assays are needed to elucidate the functional role of this protein within this venom system.

The Hyaluronidase PH-20 (HYALP) protein found in *B. brevicauda*'s saliva has potentially an important function in aiding the diffusion of other toxin components in the prey. Hyalurondase PH-20 is most well known as a spreading and adhesion molecule that facilitates the penetration of sperm through the cumulus mass to the oocyte (Stern and Jedrzejas 2006). Homologous Hyalurondase proteins are often found with other toxins in venomous snakes and arthropods (e.g., spiders, scorpions, and hymenopteran insects: Černá et al. 2002). These "spreading factor" Hyalurondase proteins can facilitate the spread of toxins by degrading the extracellular matrix (Kreil 1995). Interestingly, Hyaluronidase proteins are also present in the Gila monster (*Heloderma suspectum suspectum*) venom (Tu and Hendon 1983; Sanggaard et al. 2015), which also contains Kallikrein toxins that have convergently evolved to have similar active sites as *B. brevicauda*'s BLTX (Kita et al. 2004; Aminetzach et al. 2009).

The TFPI2 protein also may have important function in aiding the diffusion of other *B. brevicauda* toxins in prey. This anticoagulant protein is a Kunitz-type serine protease inhibitor that inhibits tissue factors involved in thrombosis (Wood et al. 2014; Maroney and Mast 2015). This venom component has been found in ticks and some venomous snakes. The anticoagulating effects of TFPI2 may facilitate the spread of other *B. brevicuada* toxins in prey. Interestingly, this protein is not a typical component of mammalian saliva (Amerongen and Veerman 2002; Blanchard et al. 2015) and has possibly been recruited as a novel protein in *B. brevicauda*'s saliva.

Finally, *B. brevicauda*'s *Antileukoproteinase* gene has sequence similarity to the Waprin toxins from snake venom (Torres et al. 2003). We found this candidate toxin in *B. brevicauda* to be highly expressed in the submaxillary gland, the most abundant protein in the saliva, and also undergoing rapid evolution in *B. brevicauda*, as well as across mammals. In general, the antileukoproteinase protein acts as an inhibitor of serine-proteinases, aids in regulating innate immunity and wound healing (Zhu et al. 2002), and is an important antimicrobial agent in saliva (Tomee et al. 1998). However, this protein is a common component of mammalian saliva (Williams et al. 2006; Torres et al. 2018), and thus it is difficult to discern without further functional evidence whether this protein has a special toxic function in *B. brevicauda*.

Genomic Characterization and Evolution of the KLK1 Gene Subfamily

Genomic duplication events of genes involved in key physiological pathways have been shown to be a major mechanism for the recruitment of venom genes across many evolutionarily divergent animals (Gibbs and Rossiter 2008). The toxins associated with the KLK1 gene subfamily in eulipotyphlans follows a similar recruitment of venom genes *via* gene duplication. A recent study on another venomous eulipotyphlan, the Hispaniolan solenodon (*Solenodon paradoxus*), also KLK1-like proteins to be the main components of their venom (Casewell et al. 2019). Examination of the reference genome in *B. brevicauda* reveals a total of five KLK1 paralogous genes tandemly arrayed in the genome. The order of these genes include a likely nontoxin *KLK1-BL1*, a known nontoxin *Blarinasin-1* (Kita et al. 2005), a known toxin *BLTX* (Kita et al. 2004), a candidate toxin *KLK1-BL2*, and a likely nontoxin *KLK1-BL3*. Our phylogenetic results of the KLK1 gene subfamily show that the *KLK1-BL1* gene is the likely ancestral gene to the other *B. brevicauda* KLK1 paralogs because it diverged first among the KLK1 genes and in the ancestor of both *B. brevicauda* and *Sor. araneus* shrews. Moreover, the toxic *BLTX* paralog and the putatively toxic *KLK1-BL2* paralog are closely related in *B. brevicauda* and it is possible that the similarity in electrostatic potential between these paralogs reflects their shared history instead of convergence. In addition, our selection tests showed that all KLK1 paralogs in *B. brevicauda* are undergoing rapid evolution, including the *KLK1-BL3* gene even though it was not expressed in the submaxillary gland and was not present in the saliva proteome. The presence of KLK1 toxins in both *Sol. paradoxus* and *B. brevicauda* and the rapid evolution of both toxin and nontoxin KLK1 paralogs in *B. brevicauda* show that this is a dynamic gene family in eulipotyphlans and that there may be other selective pressures not related to venom function driving their rapid evolution.

Possible Endogenous Venom Defense

Many venomous snakes are known to have inhibitory proteins in their circulatory systems that can act as direct antagonists to their own venom components (Mackessey 2010; Santos-Filho and Santos 2017). It is possible that the double-headed protease inhibitor found in *B. brevicauda*'s submaxillary gland serves a similar function. This inhibitor was one of the most highly expressed transcripts in *B. brevicauda*'s submaxillary gland, but the actual protein was found at extremely low levels in the saliva in only one of the two shrew samples. Analysis of this gene shows it contains two Kazal-type serine protease domains in tandem. Kazal-type serine protease inhibitors can contain multiple Kazal domains with a variable amount of amino acids making up each domain. This variability is thought to confer different specificities for their target serine proteases (Rimphanitchayakit and Tassanakajon 2010). Kazal-type serine protease inhibitors have been found to have

a role in the venom of the eyelash and side-striped palm vipers (*Bothriechis schlegelii* and *Bot. lateralis*), however, they are relatively rare venom constituents (Durban et al. 2011). We have shown with models of protein–protein interaction that this inhibitor has the potential to bind with BLTX in a similar inhibitory manner as other previously characterized kazal-type protease inhibitors. Previous studies of kazal-type protease inhibitor domains have similarly shown protease-binding loops binding directly to active sites of the protease they are inhibiting (Krowarsch et al. 2003; Rimphanitchayakit and Tassanakajon 2010). Since the main component of *B. brevicauda* venom is a serine protease, BLTX, and we see little evidence for this inhibitor protein in the saliva, we surmise that this inhibitor may serve as a self-defense mechanism against BLTX within the submaxillary gland where it is highly expressed. If so, it is likely that this protein serves to combat the vasodilatory effects of BLTX by directly binding and inhibiting it within the submaxillary gland.

The Antileukoproteinase protein has been shown to be a major inhibitor of Kallikrein activity, including BLTX and Blarinasin (Kita et al. 2004, 2005). It is therefore peculiar that this protein is the most abundant individual protein among all salivary proteins when BLTX is purported to be a major venom component. Perhaps Antileukoproteinase has a role in self-defense for *B. brevicauda* because of the large amount of Kallikrein proteins combined (BLTX, KLK1-BL2, and Blarinasin) in the saliva. Moreover, Antileukoproteinase is also undergoing rapid evolution in *B. brevicauda* and it is possible that this protein is coevolving with the rapid evolutionary changes that are also occurring in Kallikreins proteins.

Another potential defense mechanism in *B. brevicauda* that is highly expressed in the submaxillary gland, but is not abundantly present as a salivary protein is Endothelin-1. Endothelin-1's are homologous and structurally similar to the snake venom protein Sarafotoxin, which is highly lethal to small vertebrates (Kochva et al. 1993) because it causes dramatic increases in blood pressure due to extreme vasoconstriction (Wollberg et al. 1989; Mackessey 2010). However, sarafotoxins have only been identified as venom constituents within the *Atractaspis* genus of snakes (Fry 2015). Therefore, we suspect the high expression of Endothelin in *B. brevicauda* may function as a vasoconstrictor as a means to ameliorate the vasodilatory effects caused by BLTX within the submaxillary gland. This would be consistent with previous work that has shown *B. brevicauda* to have a high tolerance to its own venom when injected with extracts from their own submaxillary gland (Pearson 1950).

Venom Simplicity in Relation to Diet Complexity

One of the lingering questions about the evolution of venom in *B. brevicauda*, and other venomous shrew species, is identifying the selective pressures that led to the production of venom. Diet complexity has been shown to correspond with greater complexity in venom composition in many organisms including cone snails, snakes, and spiders (Daltry et al. 1996; Phuong et al. 2016; Pekár et al. 2018). However, despite feeding on a wide-range of prey from different animal phyla, including Arthropoda, Annelida, Mollusca, and Chordata, *B. brevicauda* appears to have a simple mixture of toxins in their saliva (Hamilton 1941; George et al. 1986). Given that other nonvenomous shrew species also have broad diets (Hamilton 1930; Vander Wall 1990; Churchfield and Sheftel 1994), it is possible that envenomating prey is a specialized accessory for capturing prey in *B. brevicauda* rather than a representation of a major shift in feeding strategy. *Blarina brevicauda*, like other nonvenomous shrew species, rely heavily on an active foraging strategy, specialized dentition, and masticatory systems for capturing and consuming prey (Dufton 1992; Furió et al. 2010; Folinsbee 2013). *Blarina brevicauda*'s lack of specialized morphologies for delivering venom and active foraging strategy with high feeding frequency rather than a sit-and-wait foraging strategy highlights their uniqueness among venomous animals (Greene 1983; Nyffeler et al. 1994; Folinsbee 2013).

Other possible explanations for the simple toxin mixture in *B. brevicauda* are that they use venom for defensive purposes or that their venom has evolved relatively recently. Venomous animals, such as bees, wasps, ants, and some fishes, that deliver toxins for defensive purposes also tend to have simple venom mixtures (Casewell et al. 2013). However, these animals also typically have aggressive defensive behaviors or a special delivery apparatuses for injecting venom (Starr 1985; Church 2002; Casewell et al. 2013; Reed and Landolt 2019). These type of specialized defensive behaviors or morphological tooth features for delivering venom are not known in *B. brevicauda* (Folinsbee 2013). Alternatively, the simplicity of *B. brevicauda*'s venom may be due to its relatively recent origin. The *Blarina* lineage is $<$10 Ma (Folinsbee 2013), which represents an extremely young evolutionary lineage for a venomous animal (Sunagar and Moran 2015). The rate of evolution among different toxin families can vary dramatically and the evidence of positive selection in some of *Blarina*'s venom components is consistent with other evolutionary young taxonomic groups that are venomous (Sunagar and Moran 2015).

The potential toxin synergism between the paralytic effects by Soricidin, vasodilation effects by BLTX and possibly KLK1-BL2, the anticoagulant effects by TPFI2, and the spreading effect by HYALP points to vertebrates as a main prey target for *B. brevicuada* venom. This supports the "hunting big" hypothesis that proposes venom evolved in *B. brevicauda* to help facilitate the capture of greater numbers of small vertebrate prey. Certainly, the paralytic effects by Soricidin may still be important for the "hoarding small" hypothesis that proposes venom evolved to facilitate longer caching of living invertebrate prey, a behavior observed in *B. brevicauda*, as well as several other nonvenomous shrews (Hamilton 1930;

Hamilton 1941; Robinson and Brodie 1982; Martin 1984). To further understand the merit of both hypotheses, there needs to be functional studies for all toxin and their combinations on both vertebrate and invertebrate prey, particularly native prey. For instance, BLTX was only functionally experimented on mice and Soricidin was only functionally tested on mealworms. Furthermore, feeding ecology studies are needed to reveal how much vertebrate and invertebrate prey make up the total caloric amount in *B. brevicauda*'s diet, as well as in the diet of other venomous and nonvenomous shrews. A recent study comparing the feeding ecology of another venomous shrew species, *N. fodiens* with the sympatric nonvenomous shrew, *Sor. araneus*, found that *N. fodiens* caches fewer invertebrate prey than *Sor. araneus*, but is able to overpower and cache larger prey more quickly (Kowalski and Rychlik 2018), thus suggesting that venom evolution may have been driven by a dietary expansion toward larger prey.

The Value of Long-Read Sequencing for Venom Transcriptome Studies

Venom systems often arise through duplication of genes, which leads to bioinformatic challenges in poorly characterized venom systems when assembling short-read DNA sequence data into distinct paralogs. Sequence similarity of paralogous genes can sometimes lead to erroneously assembled chimeric transcripts using current short-read assembly methods (Grabherr et al. 2011). Our transcriptome assembly with both short-read and long-read sequences had significantly fewer transcripts with ORFs that contained signal peptides than the two other assemblies based solely on short-read data (171 vs. 583 and 642). Many of the extra transcripts in the assemblies with only short-read data were lowly expressed transcripts (often singletons), which perhaps represented erroneous assemblies.

Long-range sequencing platforms, such as Nanopore MinION, can be very useful for transcriptome studies because each long read represents, in theory, a full-length transcript that does not require assembly. Thus, these long-read sequences can serve to validate transcript assemblies from short-read sequences, as well as in detecting transcripts that were missed from a short-read assembly approach. There is also a reciprocal benefit of the using short-read data because its higher accuracy allows for better assistance with clustering of long-read data, which are prone to higher error rates in their sequences (Laver et al. 2015). Despite the usefulness of merging these approaches for discovering paralogous transcripts, further improvements in assembly methods for long-read MinION mRNA data are still needed to automate paralog detection more efficiently.

Conclusion

Our research represents one of the few integrative multiomic comprehensive characterizations of a venom system in a single study. The components of venom in *B. brevicauda* appear to be simple relative to their diet, but further functional assays of our newly identified putative toxins are needed to complete the characterization of this venom system. Venom has evolved in at least two other eulipotyphlan genera, and possibly several more given multiple anecdotal accounts of other shrew species paralyzing their prey (Dufton 1992; Folinsbee 2013). The molecular bases of venom in these other venomous shrews are just recently becoming characterized, but further work is needed to completely examine whether functional convergence has evolved from similar proteins. Herein, we provide a comprehensive investigation of the venom system in short-tailed shrews that furthers our understanding of how the evolution of toxicity for prey capture has arisen in mammals. These findings will be useful for future comparative studies with other venomousand nonvenomous eulipotyphlans.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We are grateful to Nicholas Casewell, H. Lisle Gibbs, and Marymegan Daly for fruitful discussions and/or review of this article. We are also grateful to three anonymous reviewers for comments on the article. The genomic sequencing was carried by the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation (Grant No. 1S10OD010786-01). The protein preparation and sequencing was carried out by the Proteomics Shared Resource at The Ohio State University, supported by NIH (Grant Nos. P30 CA016058 and S10 OD018056). This work was supported by the startup fund from the Ohio State University for A.S.C.

Literature Cited

Amerongen AN, Veerman EC. 2002. Saliva – the defender of the oral cavity. Oral Dis. 8(1):12–22.

Aminetzach YT, Srouji JR, Kong CY, Hoekstra HE. 2009. Convergent evolution of novel protein function in shrew and lizard venom. Curr Biol. 19(22):1925–1931.

Armenteros J, et al. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 37:420–423.

Arni RK, Ward RJ. 1996. Phospholipase A2—a structural review. Toxicon 34(8):827–841.

Artigiani S, et al. 2004. Plexin-B3 is a functional receptor for semaphorin 5A. EMBO Rep. 5(7):710–714.

Barlow A, Pook CE, Harrison RA, Wüster W. 2009. Coevolution of diet and prey-specific venom activity supports the role of selection in snake venom evolution. Proc R Soc B. 276(1666):2443–2449.
Bi K, et al. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13(1):403.
Blanchard AA, et al. 2015. Towards further defining the proteome of mouse saliva. Proteome Sci. 13(1):10.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.
Bordon KC, Wiezel GA, Amorim FG, Arantes EC. 2015. Arthropod venom Hyaluronidases: biochemical properties and potential applications in medicine and biotechnology. J Venom Anim Toxins Incl Trop Dis. 21:43.
Bowen CV, et al. 2013. *In vivo* detection of human TRPV6-rich tumors with anti-cancer peptides derived from soricidin. PLoS One 8(3):e58866.
Brasil C, Sepra M, de Franca T, de Castro J. 2012. Management of oral mucositis. Arch Oncol. 19(3-4):57–61.
Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 34(5):525–527.
Bryant DM, et al. 2017. A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. Cell Rep. 18(3):762–776.
Bushnell B. 2014. BBMap: a fast, accurate, splice-aware aligner. Berkeley (CA): Lawrence Berkeley National Laboratory (LBNL).
Camargo I, Álvarez-Castañeda ST. 2019. Analyses of predation behavior of the desert shrew *Notiosorex crawfordi*. Mammalia 83(3):276–280.
Casewell NR, et al. 2017. The evolution of fangs, venom, and mimicry systems in blenny fishes. Curr Biol. 27(8):1184–1191.
Casewell NR, et al. 2019. Solenodon genome reveals convergent evolution of venom in eulipotyphlan mammals. Proc Natl Acad Sci U S A. 116(51):25745–25755.
Casewell NR, Wüster W, Vonk FJ, Harrison RA, Fry BG. 2013. Complex cocktails: the evolutionary novelty of venoms. Trends Ecol Evol. 28(4):219–229.
Černá P, Mikeš L, Volf P. 2002. Salivary gland hyaluronidase in various species of phlebotomine sand flies (Diptera: Psychodidae). Insect Biochem Mol Biol. 32(12):1691–1697.
Chung AG, Belone PM, Bímová BV, Karn RC, Laukaitis CM. 2017. Studies of an androgen-binding protein knockout corroborate a role for salivary ABP in mouse communication. Genetics 205(4):1517–1527.
Churchfield S. 1990. The natural history of shrews. Ithaca (NY): Cornell University Press.
Churchfield S, Sheftel BI. 1994. Food niche overlap and ecological separation in a multi-species community of shrews in the Siberian taiga. J Zool. 234(1):105–124.
Comeau SR, Gatchell DW, Vajda S, Camacho CJ. 2004. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 20(1):45–50.
Daltry JC, Wüster W, Thorpe RS. 1996. Diet and snake venom evolution. Nature 379(6565):537–540.
Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 9(8):772.
Dufton MJ. 1992. Venomous mammals. Pharmacol Ther. 53(2):199–215.
Durban J, et al. 2011. Profiling the venom gland transcriptomes of Costa Rican snakes by 454 pyrosequencing. BMC Genomics 12(1):259.
Eadie WR. 1952. Shrew predation and vole populations on a localized area. J Mammal. 33(2):185–189.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.
El-Gebali S, et al. 2019. The Pfam protein families database in. Nucleic Acids Res. 47(D1):D427–D432.
Escoubas P, King GF. 2009. Venomics as a drug discovery platform. Expert Rev Proteomic. 6(3):221–224.
Fischer H, et al. 2016. Keratins K2 and K10 are essential for the epidermal integrity of plantar skin. J Dermatol Sci. 81(1):10–16.
Folinsbee KE. 2013. Evolution of venom across extant and extinct eulipotyphlans. Comptes Rendus Palevol. 12(7–8):531–542.
Frankish A, et al. 2018. Ensembl. Nucleic Acids Res. 46(D1):D754–D761.
Fry BG. 2015. Venomous reptiles and their toxins: evolution, pathophysiology and biodiscovery. New York: Oxford University Press.
Fry BG, et al. 2009. The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. Annu Rev Genomics Hum Genet. 10(1):483–511.
Fry BG, et al. 2012. The structural and functional diversification of the Toxicofera reptile venom system. Toxicon 60(4):434–448.
Fry BG, Wüster W. 2004. Assembling an Arsenal: origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences. Mol Biol Evol. 21(5):870–883.
Fu L, et al. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152.
Furió M, Agustí J, Mouskhelishvili A, Sanisidro Ó, Santos-Cubedo A. 2010. The paleobiology of the extinct venomous shrew *Beremendia* (Soricidae, Insectivora, Mammalia) in relation to the geology and paleoenvironment of Dmanisi (Early Pleistocene, Georgia). J Vertebr Paleontol. 30(3):928–942.
Gao F, et al. 2019. EasyCodeml: a visual tool for analysis of selection using CodeML. Ecol Evol. 9(7):3891–3898.
George SB, Choate JR, Genoways HH. 1986. Blarina brevicauda. Mamm Species. 261:1–9.
Gibbs HL, Rossiter W. 2008. Rapid evolution by positive selection and gene gain and loss: PLA2 venom genes in closely related *Sistrurus* rattlesnakes with divergent diets. J Mol Evol. 66(2):151–166.
Gilbert D. 2013. Gene-omes built from mRNA-seq not genome DNA. 7th Annual Arthropod Genomics symposium. Notre Dame. F1000Research. 5:1695.
Grabherr MG, et al. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 29(7):644–652.
Greene HW. 1983. Dietary correlates of the origin and radiation of snakes. Am Zool. 23(2):431–441.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52(5):696–704.
Haas BJ, et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8(8):1494–1512.
Hackl T, Hedrich R, Schultz J, Förster F. 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 30(21):3004–3011.
Hamilton WJ. 1941. The food of small forest mammals in eastern United States. J Mammal. 22(3):250–263.
Hargreaves AD, Mulley JF. 2015. Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. PeerJ 3:e1441.
Harris JB, Scott-Davey T. 2013. Secreted phospholipases A2 of snake venoms: effects on the peripheral neuromuscular system with comments on the role of phospholipases A2 in disorders of the CNS and their uses in industry. Toxins 5(12):2533–2571.
Hatton MN, Loomis RE, Levine MJ, Tabak LA. 1985. Masticatory lubrication. The role of carbohydrate in the lubricating property of a salivary glycoprotein-albumin complex. Biochem J. 230(3):817–820.
Henry MS, Gendron L, Tremblay ME, Drolet G. 2017. Enkephalins: endogenous analgesics with an emerging role in stress resilience. Neural Plast. 2017:1–11.
Heo SM, et al. 2013. Host defense proteins derived from human saliva bind to *Staphylococcus aureus*. Infect Immun. 81(4):1364–1373.

Holding M, et al. 2018. Evaluating the performance of *de novo* assembly methods for venom-gland transcriptomics. Toxins 10(6):249.
Huang YC, et al. 2007. The flexible and clustered lysine residues of human ribonuclease 7 are critical for membrane permeability and antimicrobial activity. J Biol Chem. 282(7):4626–4633.
Jackson BC, et al. 2011. Update of the human secretoglobin (SCGB) gene superfamily and an example of 'evolutionary bloom' of androgen-binding protein genes within the mouse Scgb gene superfamily. Hum Genomics. 5(6):691.
Jönsson D. 2018. Antimicrobial peptides: roles in periodontal health and disease. In: Bostanci N, Belibasakis G, editors. Pathogenesis of periodontal diseases. Cham: Springer. p. 97–110.
Kita M, et al. 2004. Blarina toxin, a mammalian lethal venom from the short-tailed shrew *Blarina brevicauda*: isolation and characterization. Proc Natl Acad Sci U S A. 101(20):7542–7547.
Kita M, et al. 2005. Purification and characterisation of blarinasin, a new tissue kallikrein-like protease from the short-tailed shrew *Blarina brevicauda*: comparative studies with blarina toxin. Biol Chem. 386(2):177–182.
Kochva E, Bdolah A, Wollberg Z. 1993. Sarafotoxins and endothelins: evolution, structure and function. Toxicon 31(5):541–568.
Kordiš D. 2011. Evolution of phospholipase A2 toxins in venomous animals. Acta Chim Slov. 58(4):638–646.
Kowalski K, Marciniak P, Rosiński G, Rychlik L. 2017. Evaluation of the physiological activity of venom from the Eurasian water shrew *Neomys fodiens*. Front Zool. 14:46.
Kowalski K, Rychlik L. 2018. The role of venom in the hunting and hoarding of prey differing in body size by the Eurasian water shrew, *Neomys fodiens*. J Mammal. 99(2):351–362.
Kreil G. 1995. Hyaluronidases—a group of neglected enzymes. Protein Sci. 4(9):1666–1669.
Krowarsch D, Cierpicki T, Jelen F, Otlewski J. 2003. Canonical protein inhibitors of serine proteases. Cell Mol Life Sci. 60(11):2427–2444.
Kumar S, et al. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 35(6):1547–1549.
Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol. 29(6):1695–1701.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–359.
Laver T, et al. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 3:1–8.
Ligabue-Braun R. 2015. Venom use in mammals: evolutionary aspects. Evol Venomous Anim Toxins. 2015:1–23.
Mackessey. 2010. Handbook of venoms and toxins of reptiles. Boca Raton: CRC Press.
Magi A, et al. 2018. Nanopore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform. 19:1256–1272.
Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27(21):2957–2963.
Maroney SA, Mast AE. 2015. New insights into the biology of tissue factor pathway inhibitor. J Thromb Haemost. 13:S200–S207.
Martin IG. 1981. Venom of the short-tailed shrew (*Blarina brevicauda*) as an insect immobilizing agent. J Mammal. 62(1):189–192.
Martin IG. 1984. Factors affecting food hoarding in the short-tailed shrew *Blarina brevicauda*. Mammalia 48(1):65–72.
Nussbaum RA, Maser C. 1969. Observations of *Sorex palustris* preying on *Dicamptodon ensatus*. Murrelet 1:23–24.
Nyffeler M, Sterling WL, Dean DA. 1994. How spiders make a living. Environ Entomol. 23(6):1357–1367.
Olsson A, Lilja H, Lundwall Å.2004. Taxon-specific evolution of glandular kallikrein genes and identification of a progenitor of prostate-specific antigen. Genomics 84(1):147–156.
Pearson OP. 1942. On the cause and nature of a poisonous action produced by the bite of a shrew (*Blarina brevicauda*). J Mammal. 23(2):159–166.
Pearson OP. 1950. The submaxillary glands of shrews. Anat Rec. 107(2):161–169.
Pekár S, et al. 2018. Venom gland size and venom complexity—essential trophic adaptations of venomous predators: a case study using spiders. Mol Ecol. 27(21):4257–4269.
Phuong MA, Mahardika GN, Alfaro ME. 2016. Dietary breadth is positively correlated with venom complexity in cone snails. BMC Genomics 17(1):15.
Reed HC, Landolt PJ. 2019. Ants, wasps, and bees (Hymenoptera). In: Mullen GR, Durden LA, editors. Medical and veterinary entomology. Cambridge: Academic Press. p. 459–488.
Rhoads A, Au KF. 2015. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13(5):278–289.
Rimphanitchayakit V, Tassanakajon A. 2010. Structure and function of invertebrate Kazal-type serine proteinase inhibitors. Dev Comp Immunol. 34(4):377–386.
Robertson G, et al. 2010. De novo assembly and analysis of RNA-seq data. Nat Methods. 7(11):909–912.
Robinson DE, Brodie ED. ED. 1982. Food hoarding behavior in the short-tailed shrew *Blarina brevicauda*. Am Midl Nat. 108(2):369–375.
Rode-Margono EJ, Nekaris AK. 2015. Cabinet of curiosities: venom systems and their ecological function in mammals, with a focus on primates. Toxins 7(7):2639–2658.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12):1572–1574.
Sanggaard KW, et al. 2015. Characterization of the Gila monster (*Heloderma suspectum suspectum*) venom proteome. J Proteomics. 117:1–11.
Santos-Filho NA, Santos CT. 2017. Alpha-type phospholipase A(2) inhibitors from snake blood. J Venom Anim Toxins Incl Trop Dis. 23:19.
Schrödinger, LLC. 2015. The {PyMOL} molecular graphics system, Version~1.8.
Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28(8):1086–1092.
Silla-Martínez JM, Capella-Gutiérrez S, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.
Singhal S. 2013. *De novo* transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. Mol Ecol Resour. 13(3):403–416.
Smith-Unna R, et al. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 26(8):1134–1144.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.
Starr CK. 1985. Enabling mechanisms in the origin of sociality in the Hymenoptera—the sting's the thing. Ann Entomol Soc Am. 78(6):836–840.
Stern R, Jedrzejas MJ. 2006. Hyaluronidases: their genomics, structures, and mechanisms of action. Chem Rev. 106(3):818–839.
Sunagar K, Moran Y. 2015. The rise and fall of an evolutionary innovation: contrasting strategies of venom evolution in ancient and young animals. PLoS Genet. 11(10):e1005596.
Sunagar K, Morgenstern D, Reitzel AM, Moran Y. 2016. Ecological venomics: how genomics, transcriptomics and proteomics can shed new light on the ecology and evolution of venom. J Proteomics. 135:62–72.
Tan MH, et al. 2018. Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly. GigaScience 7(3):gix137.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures Math Life Sci. 17:57–86.
Tegoni M, et al. 2000. Mammalian odorant binding proteins. Biochim Biophys Acta Protein Struct Mol Enzymol. 1482(1–2):229–240.
Thompson JD, Plewniak F, Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27(13):2682–2690.
Tomasi TE. 1978. Function of venom in the short-tailed shrew, *Blarina brevicauda*. J Mammal. 59(4):852–854.
Tomee JF, et al. 1998. Secretory leukoprotease inhibitor: a native antimicrobial protein presenting a new therapeutic option. Thorax 53(2):114–116.
Torres AM, et al. 2003. Identification of a novel family of proteins in snake venoms purification and structural characterization of nawaprin from *Naja nigricollis* snake venom. J Biol Chem. 278(41):40097–40104.
Torres SM, et al. 2018. Salivary proteomics of healthy dogs: an in depth catalog. PLoS One 13(1):e0191307.
Tu AT, Hendon RR. 1983. Characterization of lizard venom hyaluronidase and evidence for its action as a spreading factor. Comp Biochem Physiol B Biochem Mol Biol. 76(2):377–383.
Vander Wall SB. 1990. Food hoarding in animals. Chicago: University of Chicago Press.
Verdes A, et al. 2016. From mollusks to medicine: a venomics approach for the discovery and characterization of therapeutics from Terebridae peptide toxins. Toxins 8(4):117.
Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. Genome Res. 27(5):757–767.
Williams SE, et al. 2006. SLPI and elafin: one glove, many fingers. Clin Sci. 110(1):21–35.
Wollberg Z, Bdolah A, Kochva E. 1989. Vasoconstrictor effects of sarafotoxins in rabbit aorta: structure-function relationships. Biochem Biophys Res Commun. 162(1):371–376.
Wood JP, Ellery PE, Maroney SA, Mast AE. 2014. Biology of tissue factor pathway inhibitor. Blood 123(19):2934–2943.
Wu S, Zhang Y. 2008. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72(2):547–556.
Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.
Yang Z, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol. 19(1):49–57.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22(12):2472–2479.
Zhang Y, et al. 2012. BmK-YA, an enkephalin-like peptide in scorpion venom. PLoS One 7(7):e40417.
Zhu J, et al. 2002. Conversion of proepithelin to epithelins: roles of SLPI and elastase in host defense and wound repair. Cell 111(6):867–878.

Associate editor: Sabyasachi Das