Supplementary Figures and Legends for: "Parasitoid Jewel Wasp Mounts Multi-Pronged Neurochemical Attack to Hijack a Host Brain"

Ryan Arvidson, Maayan Kaiser, Sang Soo Lee, Jean-Paul Urenda, Christopher Dail, Haroun Mohammed, Cebrina Nolan, SongQin Pan, Jason E. Stajich, Frederic Libersat, Michael E. Adams.

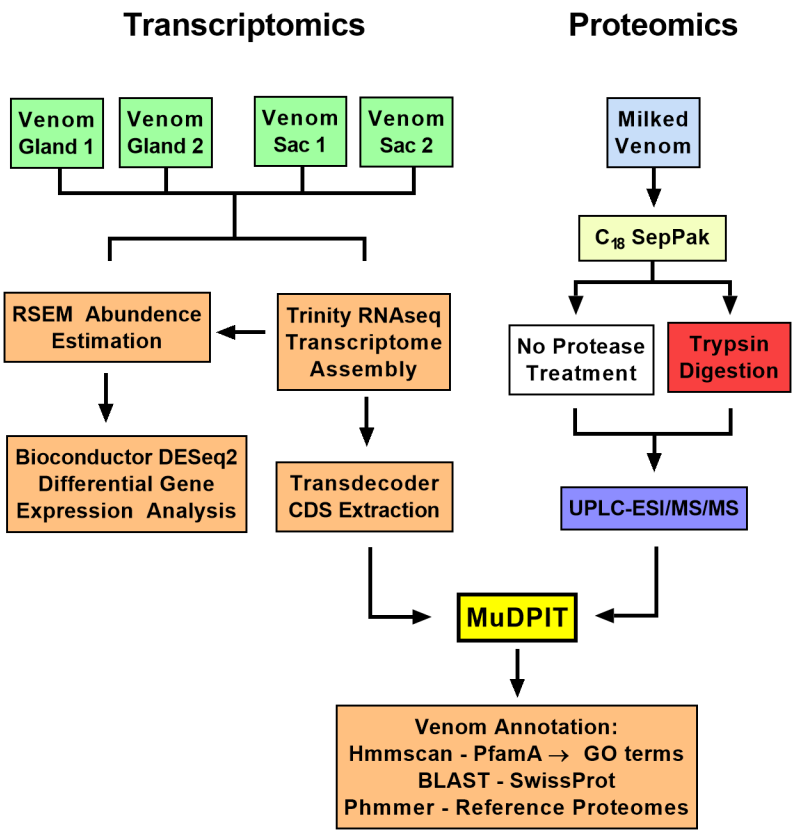


Figure S1. Experimental Design. Transcriptomics: RNA was sequenced from two distinct components (the venom gland and the venom sac) of the venom apparatus in biological and technical replicates. RNA sequencing reads were combined and assembled into a transcriptome using the Trinity pipeline. Transcript levels for each replicate were estimated using RSEM, and differential gene expression analysis between tissue types was examined using DEseq plugins for Trinity. Probable coding sequences were extracted from the transcriptome by Transdecoder. Proteomics: Protein was C_{18} SepPak-purified, concentrated by Speed-Vac, and analyzed by mass spectrometry, either by trypsinizing the sample, or direct analysis. Venom and venom apparatus proteins were identified by multiple dimension protein identification technology (MudPIT) analysis using MASCOT interrogation of the Transdecoder database with mass spectrometry data. Identified proteins were annotated using Hmmscan against PfamA, also from which GO terms were extracted; by BLAST against SwissProt database; and by Phmmer, against Uniprot's reference proteomes.

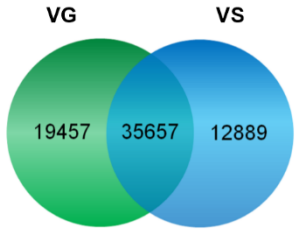
23

24

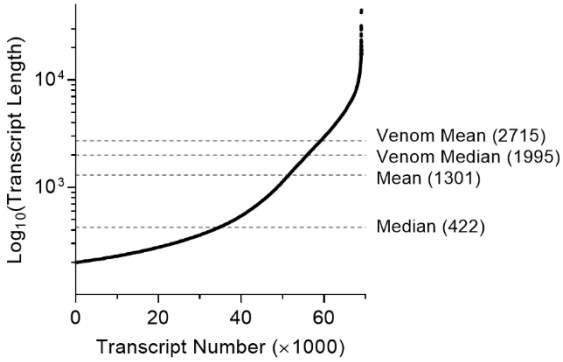
A

Trinity Stats		CEGMA	
Total Trinity Transcripts	N50	Complete %	Partial %
69009	3529	98	99.6

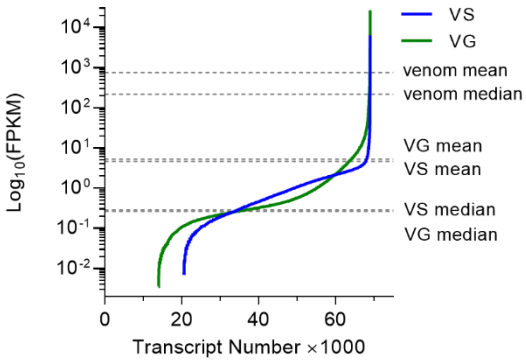
B



C



D



25

26 **Figure S2. *A. compressa* venom apparatus transcriptome statistics.**

27 **A**. Transcriptome assembly statistics and CEGMA analysis of assembly completeness. Trinity
28 assembly generated 69009 transcripts. Core Eukaryotic Gene Mapping Approach (CEGMA)
29 identified 98 ultra-conserved, core eukaryotic genes (CEGs) indicating completeness of the
30 assembly. **B**. Venn diagram of quantifiable transcripts found only in the VG, only in the VS, and
31 transcripts common to both. Of the 35,657 transcripts shared between tissue types only 2,222 are
32 differentially expressed. **C**. Transcript length in base-pairs by transcript number. Transcripts range
33 in length from 200 (assembly minimum) to 44,751 bp. Venom-specific transcripts tend to be longer
34 on average than the mean assembled transcript length. **D.** Transcript abundance estimation as
35 fragments per kilobase per million mapped reads (FPKM) by transcript number. The average
36 transcript abundance for identified venom transcripts (mean = 756) is two orders of magnitude
37 higher than the average for all transcripts (VS mean = 4.61, VG mean = 5.21)

38

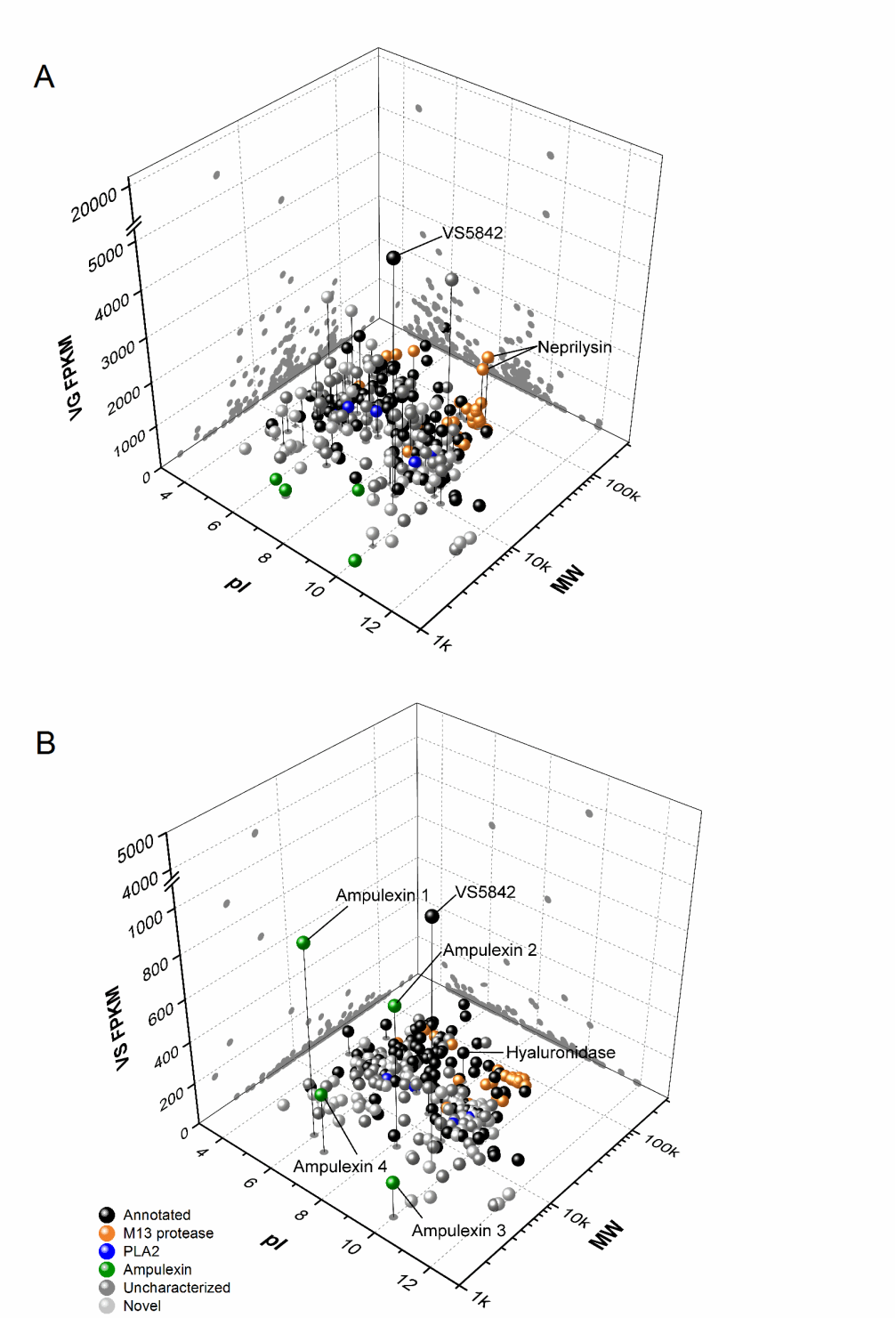

A
VG FPKM
20000
5000
4000
3000
2000
1000
0
VS5842
Neprilysin
4
6
8
10
12
pI
1k
10k
100k
MW
B
VS FPKM
5000
4000
1000
800
600
400
200
0
Ampulexin 1
VS5842
Ampulexin 2
Hyaluronidase
Ampulexin 4
Ampulexin 3
Annotated
M13 protease
PLA2
Ampulexin
Uncharacterized
Novel
4
6
8
10
12
pI
1k
10k
100k
MW

Figure S3. 3D 'gel' analysis of the venom proteome by gland expression

Theoretical values of isoelectric point (pI) and molecular weight (MW) of venom proteins plotted against relative abundance of transcripts (FPKM, fragments per kilobase per million mapped reads). **A**. Venom proteins plotted against venom gland (VG) transcript abundance. **B**. Venom proteins plotted against venom sac (VS) transcript abundance. The theoretical gel results correlate with the 1D SDS-PAGE (Figure 2B) with higher molecular weight protein expression in the venom gland and higher expression of low molecular peptides in the venom sac. The M13 family peptidases (red) and phospholipase A2 (blue) fall onto similar mass ranges but different pI's. The most abundantly expressed venom proteins in each gland are highlighted; notably M13 peptidases (neprilysin in the venom gland, and hyaluronidase, serpin, and the ampulexin peptides in the venom sac. VS5842 is highly expressed in both glands. All other proteins identified in global databases are in black. Uncharacterized proteins (grey) are differentiated from novel proteins (white) in that uncharacterized proteins were found represented in NCBI-nr database as "putative" or "uncharacterized", whereas novel proteins did not return any significant hits (E-value < 10^{-5}) from Uniprot, or PfamA databases.

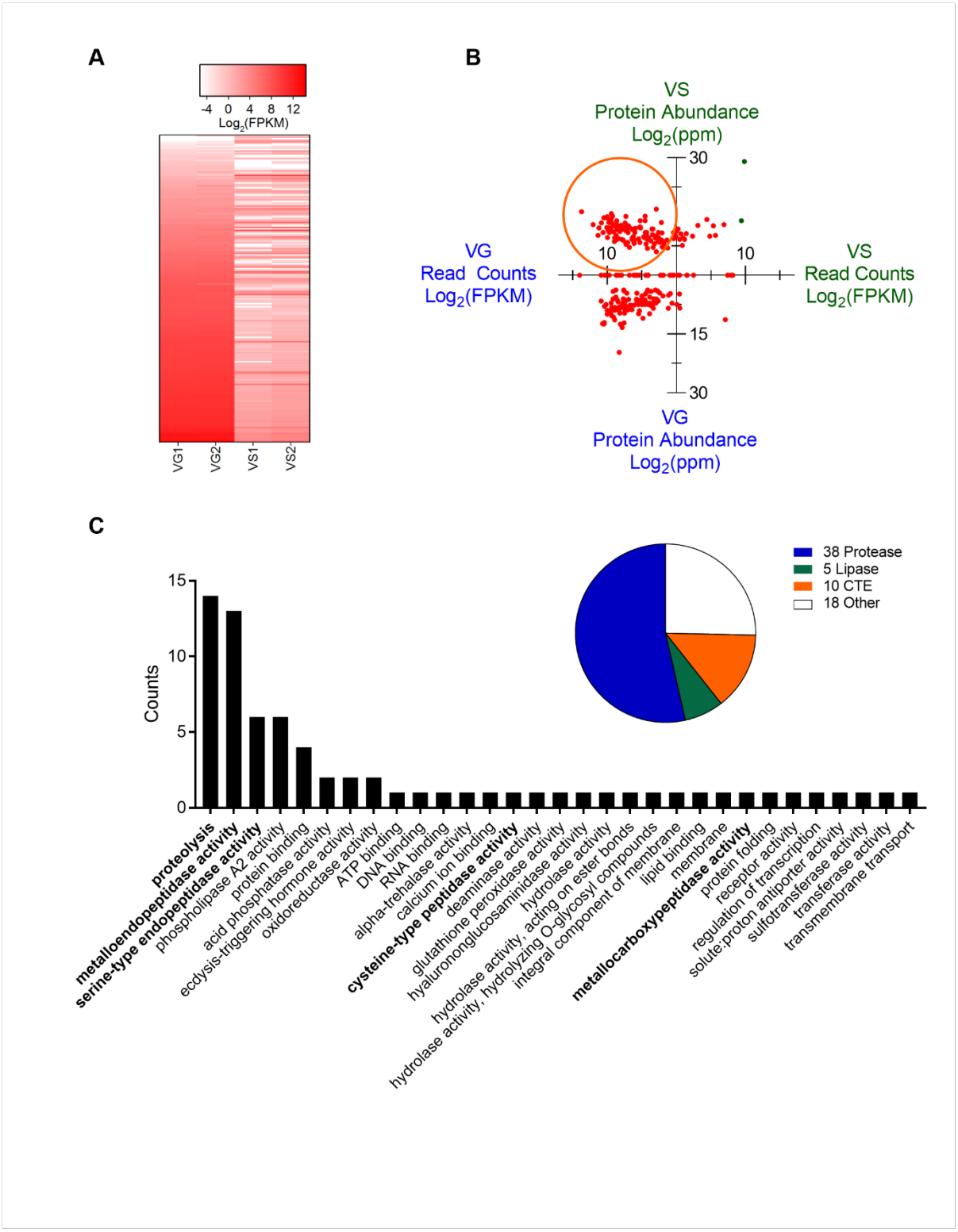

A
-4 0 4 8 12
Log2(FPKM)
VG1
VG2
VS1
VS2
B
VS
Protein Abundance
Log2(ppm)
30
10
10
15
30
VG
Read Counts
Log2(FPKM)
VS
Read Counts
Log2(FPKM)
VG
Protein Abundance
Log2(ppm)
C
38 Protease
5 Lipase
10 CTE
18 Other
Counts
15
10
5
0
proteolysis
metalloendopeptidase activity
serine-type endopeptidase activity
phospholipase A2 activity
protein binding
acid phosphatase activity
ecdysis-triggering hormone activity
oxidoreductase activity
ATP binding
DNA binding
RNA binding
alpha-trehalase activity
calcium ion binding
cysteine-type peptidase activity
deaminase activity
glutathione peroxidase activity
hyaluronoglucosaminidase activity
hydrolase activity
hydrolase activity, acting on ester bonds
hydrolase activity, hydrolyzing O-glycosyl compounds
integral component of membrane
lipid binding
membrane
metallocarboxypeptidase activity
protein folding
receptor activity
regulation of transcription
solute:proton antiporter activity
sulfotransferase activity
transferase activity
transmembrane transport

Figure S4. Bioinformatic analysis of the venom proteome.

A. Heat map representing Log_2 fold change in transcript abundance of 201 mass spectrometry identified venom transcripts between venom gland (VG) and venom sac (VS) sorted by VG1 form low to high. **B.** Venom transcript read counts and protein abundance per tissue type. For each venom protein identified in extracts of the venom gland or venom sac, transcript abundance (FPKM) is plotted vs protein abundance; emPAI values for VG and VS were normalized as parts-per-million (ppm) of the total for each tissue. On the axis of which the values are greatest, such that if both FPKM and protein abundance values are greatest in the venom sac, it would be plotted in quadrant 1, or if greater in the venom gland, plotted in quadrant 3. In a case where expression is greater in the venom gland, but protein abundance is greater in the venom sac, these values would appear in quadrant 2, and the converse plotted in quadrant 4. Clustering of values in quadrant 2 (orange circle) is consistent with the hypothesis that the majority of venom proteins are expressed in the venom gland, and stored in the venom sac. Green points in quadrant 1 are ampulexins, which we show in this work to be more highly expressed in the venom sac than venom gland. **C.** Hierarchical distribution of GO terms associated with identified venom proteins. The venom proteome contains representatives of many protein activities, but proteases are a dominant fraction. The pie chart summarizes and represents venom proteins with enzymatic activity by broad enzymatic classification. Protease activity is the dominant activity, representing more than half of venom GO terms. Proteases in the venom include serine, cysteine, and metalloproteases. A significant fraction of enzyme activity target carbohydrates or glycans as substrates and are referred to as carbohydrate-targeting-enzymes (CTE). Lipases include almost exclusively phospholipase A2 isoforms, which may not have lipase activity.

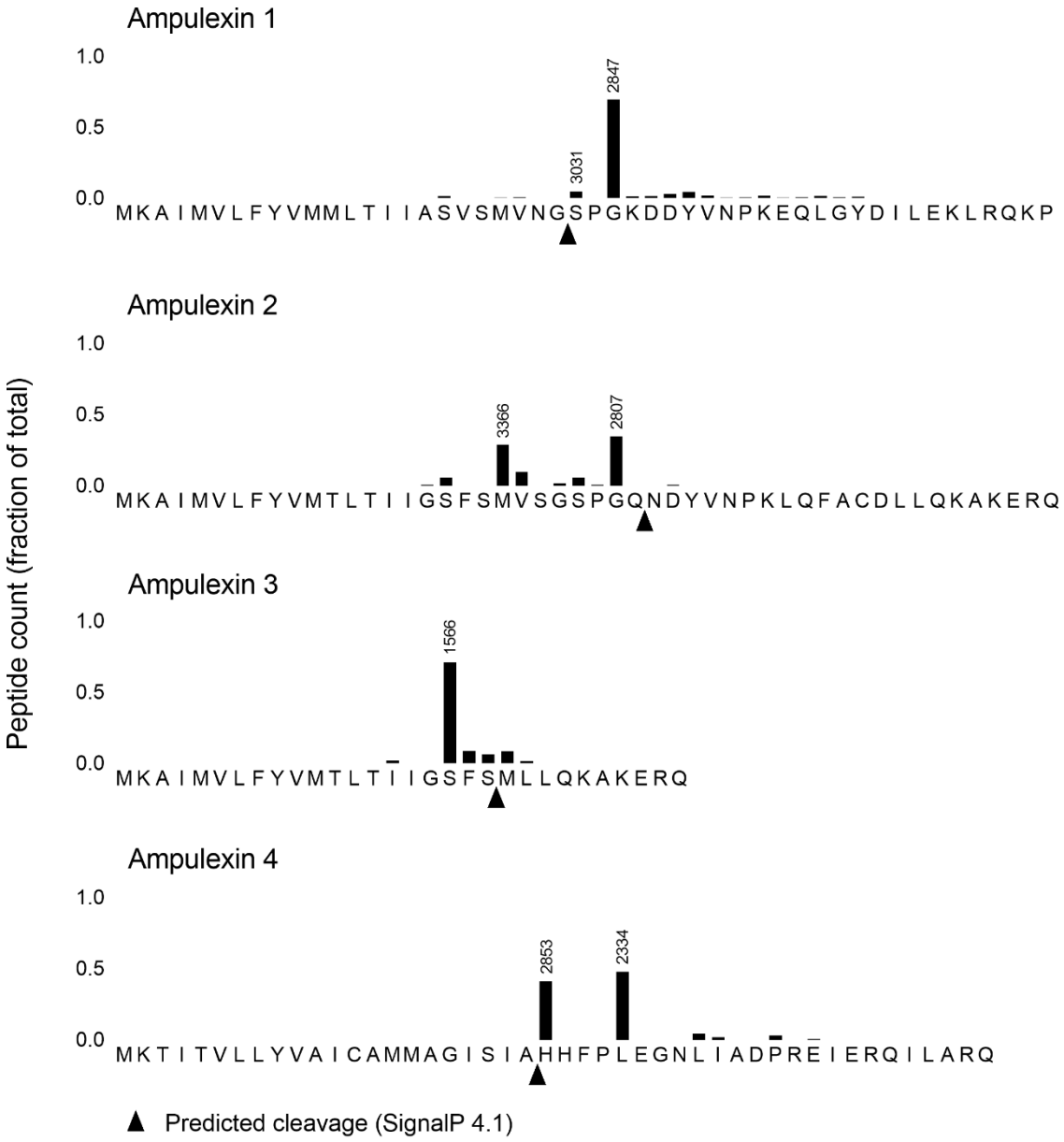


Figure S5. Ampulexin peptide counts and signal cleavage

The ampulexin family peptides cleavage sites as identified by mass spectrometry. The relative abundance of a peptide's spectral count is plotted as a histogram above the most N-terminal residue of the detected peptide. Predicted cleavage sites are indicated with an arrow. The molecular mass of the peptide is given above the histogram.

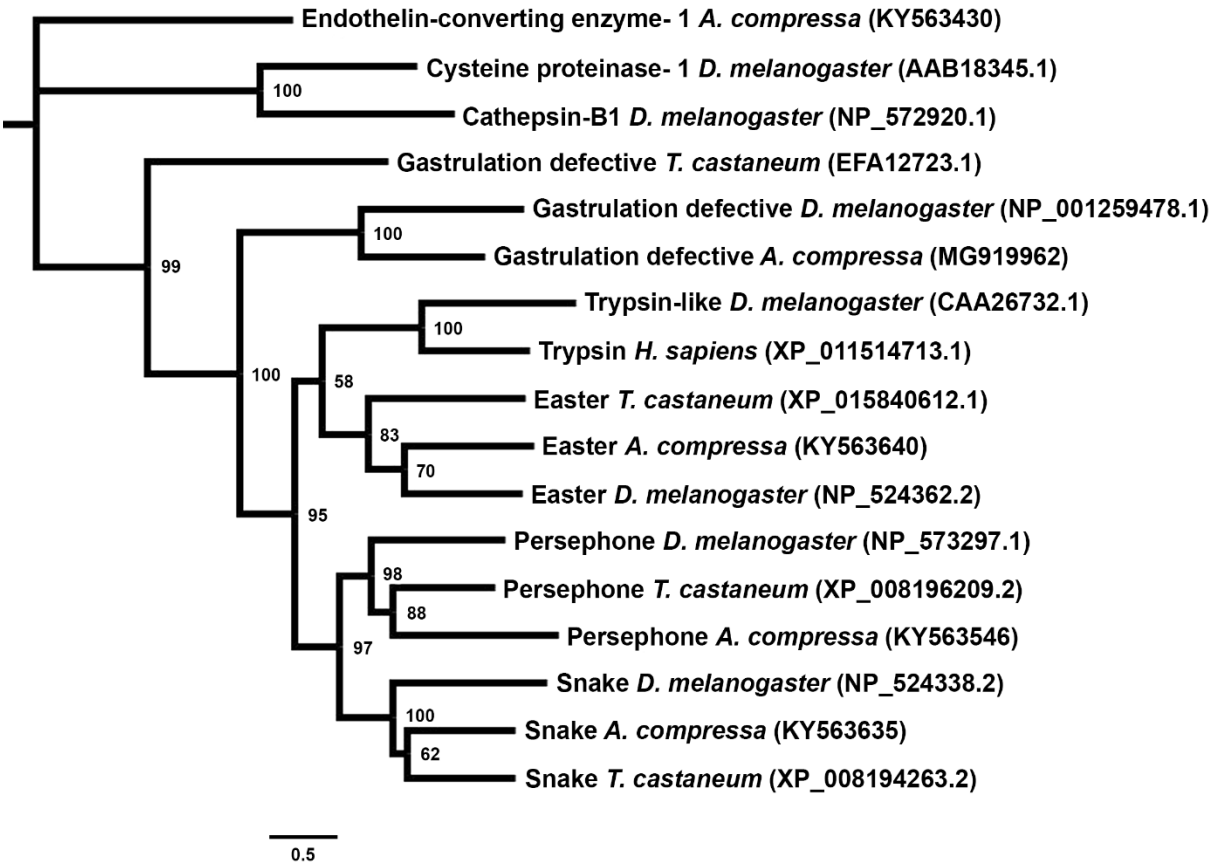


Figure S6: Maximum-likelihood phylogenetic analysis of the Spätzle activation cascade

Venom serine proteases involved in activation of the Toll ligand Spätzle cluster with their respective homologs in *Drosophila* and *Tribolium*. Spätzle is activated by a protease cascade beginning with Easter, which is activated by Snake, which in turn is activated by Gastrulation defective. Persephone is also known to be involved in a separate activation cascade of Spätzle. Cysteine proteinase-1, Cathepsin-B (a cysteine protease), and endothelin-converting enzyme (a zinc protease) were used as outgroups. Supports are bootstrap values. NCBI accession numbers are in parentheses.

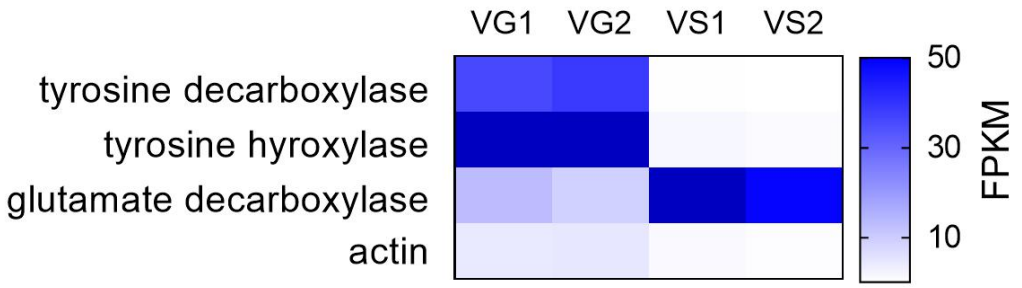


Figure S7: Expression of dopamine and GABA synthesis enzymes in the venom apparatus.

The VG and VS appear to differentially synthesize known venom components dopamine and GABA. Read counts of enzymes in the biosynthetic pathway of dopamine, tyrosine decarboxylase and tyrosine hydroxylase, are expressed highly in the VG, but not in the VS. Conversely, the biosynthetic enzyme that generates GABA form glutamate, glutamate decarboxylase, is highly expressed in the venom sac, but not the venom gland.

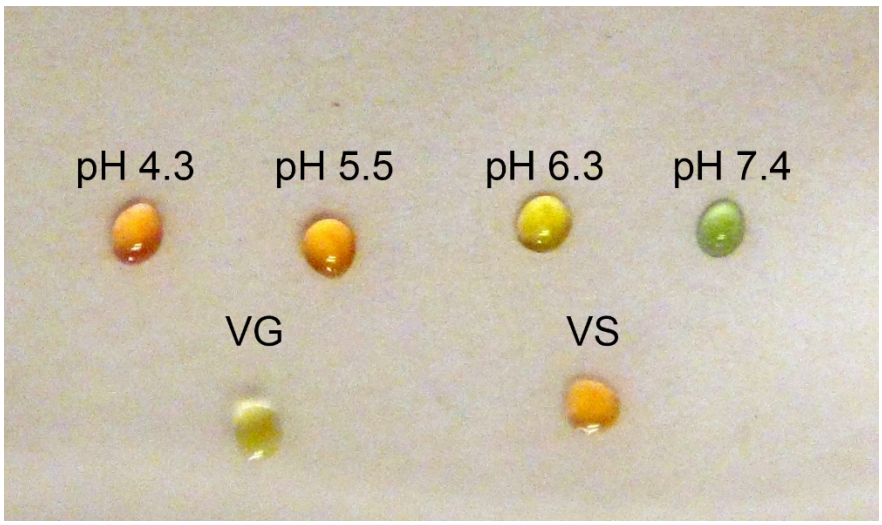


Figure S8. Venom sac contents are acidic

Phosphate-buffered saline at varying pH values was used as reference standards with a universal pH indicator solution (top row). Venom sac (VS) contents are acidic with respect to venom gland contents (VG).