SI APPENDIX: Laterally transferred gene recruited as a venom in parasitoid wasps

SUPPLEMENTAL MATERIALS AND METHODS

Venom and whole body transcriptome assembly

Transcriptomes were sequenced at the University of Rochester Genomics Research Center (URGRC) on an Illumina HiSeq2500; venom transcriptomes of 10 wasp species were sequenced on 1/10th of a lane and whole body transcriptomes of 13 wasp species were sequenced on 1/6th of a lane. Pre-processing of the raw Illumina sequences included seqClean adaptor, uniVec database filtering, and poly-A tail trimming, as well as end quality trimming using the FASTx-toolkit (fastq_quality_trimmer) with the following parameters: "-t 13 -l 25 -Q 33" (Pearson, et al. 1997). *De novo* assembly was performed via the Trinity software package v.r2013-02-25 (Grabherr, et al. 2011). To obtain expression values, raw reads were mapped using the Burrows-Wheeler Aligner (BWA v.0.7.8) allowing for two mismatches per raw read (-n 2) (Li and Durbin 2009) and fragments per kilobase per million (FPKM) values were calculated in Cufflinks v.2.2.0 (Trapnell, et al. 2012; Trapnell, et al. 2013). Genes identified as GH19 chitinase were deposited in GenBank under accessions KT359534-KT359552.

Venom proteomes

For each species (*N. vitripennis*, *N. giraulti*, *N. longicornis, T. sarcophagae, U. rufipes*, *M. uniraptor, S. cameroni, Melittobia sp,* and *T. zealandicu*s), 75-100 reservoirs were collected and venom was extracted by centrifuging at 12,000 g for 15 minutes at 4°C. Proteomic analysis and peptide mapping to the *de novo* venom transcriptomes was performed by the Proteomics and Mass Spectrometry Facility at Cornell University.

SUPPLEMENTAL RESULTS

Intron and copy number in mosquitoes

We found a single copy of the GH19 chitinase gene with no introns in *A. aegypti*, whereas GH19 chitinase appears four times in the genome of *C. quinquefasciatus* (SI Appendix, Fig. S2). In two *C. quinquefasciatus* proteins, tandem duplication has resulted in a single protein with two concatenated GH19 chitinase domains, separated by an intron (SI Appendix, Fig. S6). The third copy has developed an intron within the GH19 chitinase domain. In the final *C. quinquefasciatus* copy, the GH19 chitinase domain has become attached to the mosquito transmembrane protein,

tartan, and is transcribed into a single mRNA (EDS42856). In *Drosophila melanogaster*, *tartan* is involved in cell surface interactions during the organization of epidermal structures (Chang, et al. 1993).

GH19 chitinase in bacteria and plants

Phylogenetic reconstruction does not provide definitive support, but the likely point of origin for the GH19 chitinase gene appears to be bacteria or plants. Among bacteria, GH19 chitinase has an extremely sporadic occurrence, for example, the gene is readily found in Actinomycetales, yet is absent among the sister lineage Bifidobacteriales and the remaining taxa of the phylum Actinobacteria (supplementary fig. S8, Supplementary Material online). This patchy distribution suggests horizontal transfer has influenced the distribution of GH19 chitinase among bacteria. Alternatively, GH19 chitinase could have arisen as a fungal defense mechanism in Embryophytes (land plants) during their migration to land. GH19 chitinases are present in the earliest Embryophytes (*i.e*. mosses and spike mosses), Gymnosperms, and Angiosperms, yet absent in the immediately branching taxa (supplementary fig. S7, Supplementary Material online). Embryophyte GH19 chitinases underwent at least three duplications relatively early in plant evolution and numerous copies are present in individual genomes (*e.g*. *Zea mays* has 19 copies) (Prakash, et al. 2010).

Supplemental references

Chang Z, Price BD, Bockheim S, Boedigheimer MJ, Smith R, Laughon A. 1993. Molecular and genetic characterization of the *Drosophila tartan* gene. Dev. Biol. 160:315-332.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644-652.

Klopfstein S, Vilhelmsen L, Heraty JM, Sharkey M, Ronquist F. 2013. The hymenopteran tree of life: evidence from protein-coding genes and objectively aligned ribosomal data. PLoS ONE 8:e69344.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754-1760.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763-767.

Pearson WR, Wood T, Zhang Z, Miller W. 1997. Comparison of DNA sequences with protein sequences. Genomics 46:24-36.

Peters RS, Meusemann K, Petersen M, Mayer C, Wilbrandt J, Ziesmann T, Donath A, Kjer KM, Aspöck U, Aspöck H. 2014. The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. BMC Evol. Biol. 14:52.

Prakash NU, Jayanthi M, Sabarinathan R, Kangueane P, Mathew L, Sekar K. 2010. Evolution, homology conservation, and identification of unique sequence signatures in GH19 family chitinases. J. Mol. Evol. 70:466-478.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431-437.

Roy RS, Price DC, Schliep A, Cai G, Korobeynikov A, Yoon HS, Yang EC, Bhattacharya D. 2014. Single cell genome analysis of an uncultured heterotrophic stramenopile. Sci. Rep. 4.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-Seq. Nat. Biotechnol. 31:46-53.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. Nat. Protoc. 7:562-578.

Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim J-W, Lambkin C, Bertone MA, Cassel BK, Bayless KM, Heimberg AM. 2011. Episodic radiations in the fly tree of life. Proc. Natl. Acad. Sci. USA 108:5690-5695.

SUPPLEMENTAL FIGURES

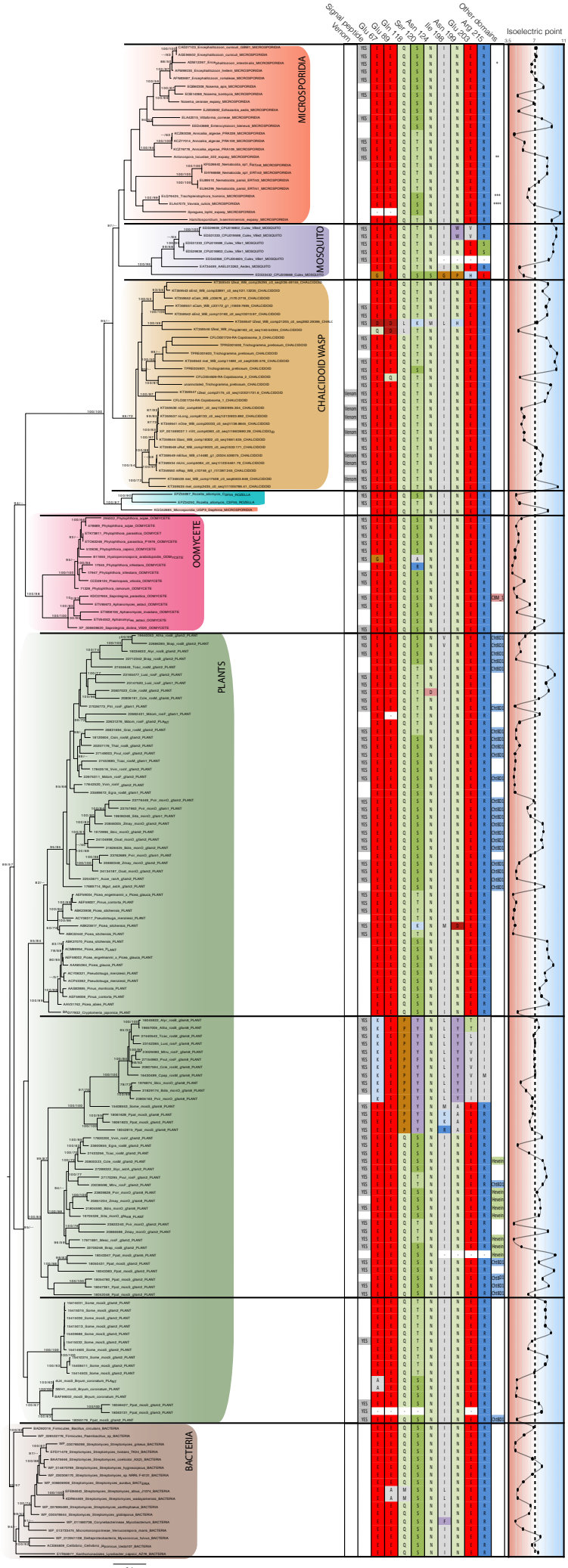


Figure S1. Full phylogeny of GH19 chitinases. A maximum likelihood tree with posterior probabilities from a Bayesian analysis (>90) and bootstrap support (>50) colored coded by taxa. Accession numbers are from NCBI with exception of the plants, which have Phytozome accession numbers. To the right of the tree is information for each protein sequence including: Pteromalid venom component, existence of a signal peptide, GH19 catalytic residue conservation, occurrence of additional protein domains, and isoelectric point. Components of Pteromalid venom were predicted based on the presence of GH19 in the venom proteome and venom transcriptome. Signal peptides were predicted with the online server, SignalP 4.1. Catalytic sites were identified in the protein sequence alignment used to generate the phylogenetic tree, with reference to the *Hordeum vulgare* chitinase 2BAA, P23951 (amino acids are colored by their underlying properties). Protein domains, including carbohydrate-binding modules, were predicted with the NCBI's Batch Web CD-Search online server (http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi). The 'other domains' category displays the additional protein domains predicted, which include: Blue- cd00035, Type 1 chitin binding domain. Green- cd06921, Hevein or Type 1 chitin binding domain. Red- pfam00734, CBM_1 Carbohydrate-binding module Family 1 (cellulose). *- cl14602, glideosome-associated protein 45; PTZ00423 superfamily. **- cl09347, DNA polymerase phi; DNA_pol_phi superfamily. ***- cl09930, Replication protein A, class 2b aminoacyl-tRNA synthetases; RPA_2b-aaRSs_OBF_like superfamily. ****- cl22451, ASF1 like histone chaperone; ASF1_hist_chap superfamily. Isoelectric point for each sequence in the phylogeny was predicted with the online Sequence Manipulation Suite: Protein Isoelectric Point (http://www.bioinformatics.org/sms2/protein_iep.html).

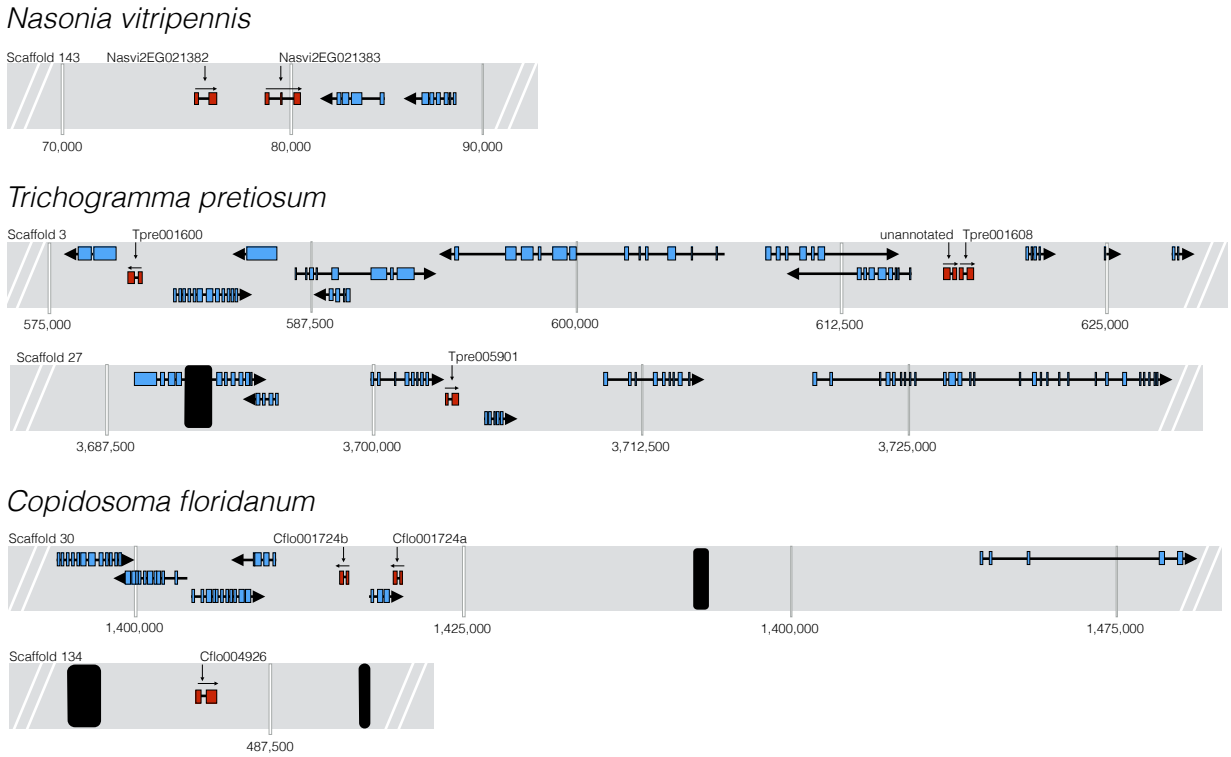


Fig. S2. Genomic position of GH19 chitinase genes in chalcidoid wasps.

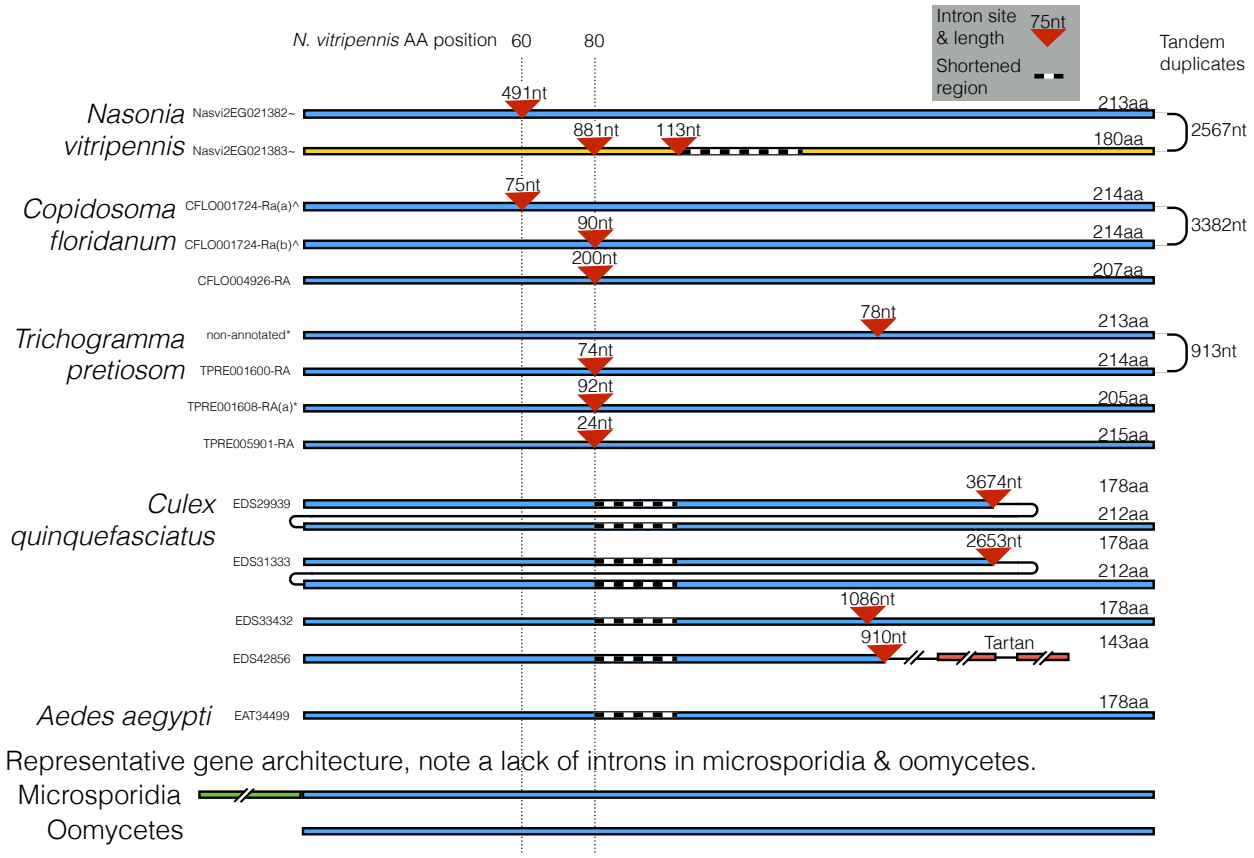


Fig. S3. GH19 chitinase gene copies in chalcidoid wasps and mosquitoes contain introns, which are not present in micrsporidia or oomycetes. Red arrows indicate intron position and the length of the intron. Yellow indicates a nonfunctional pseudogene.

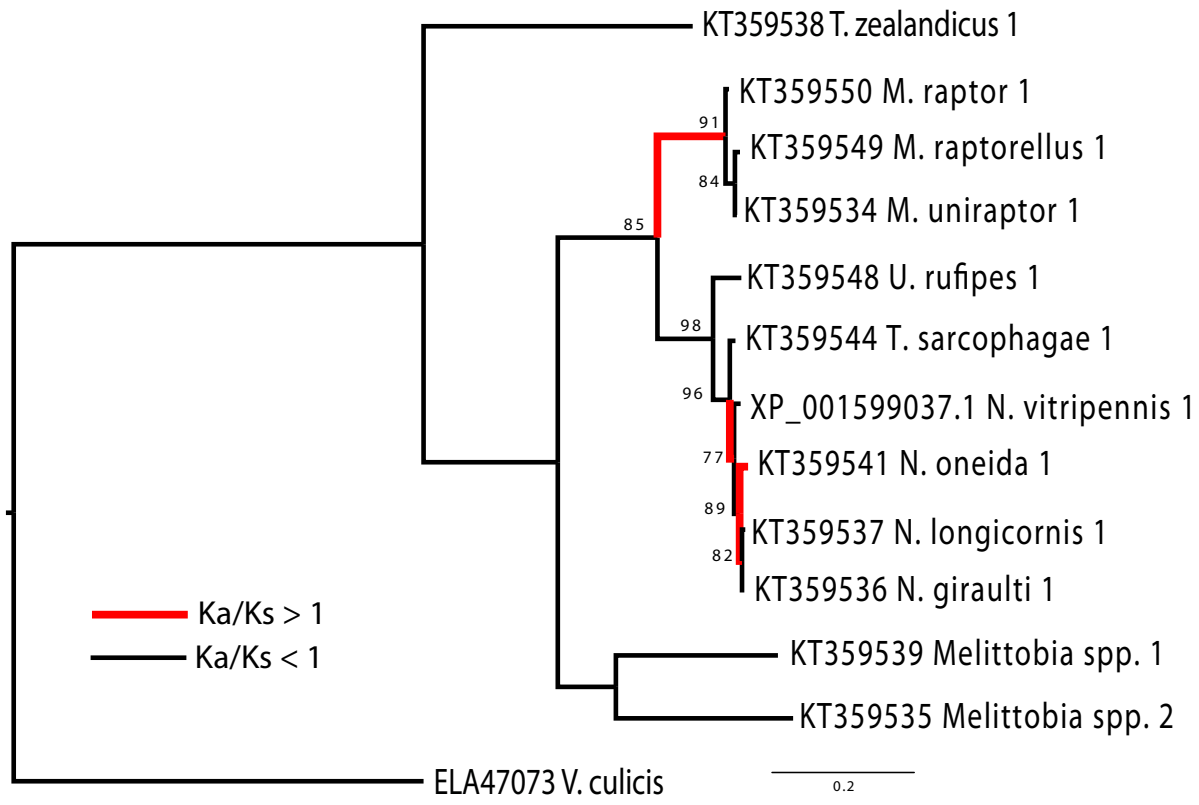
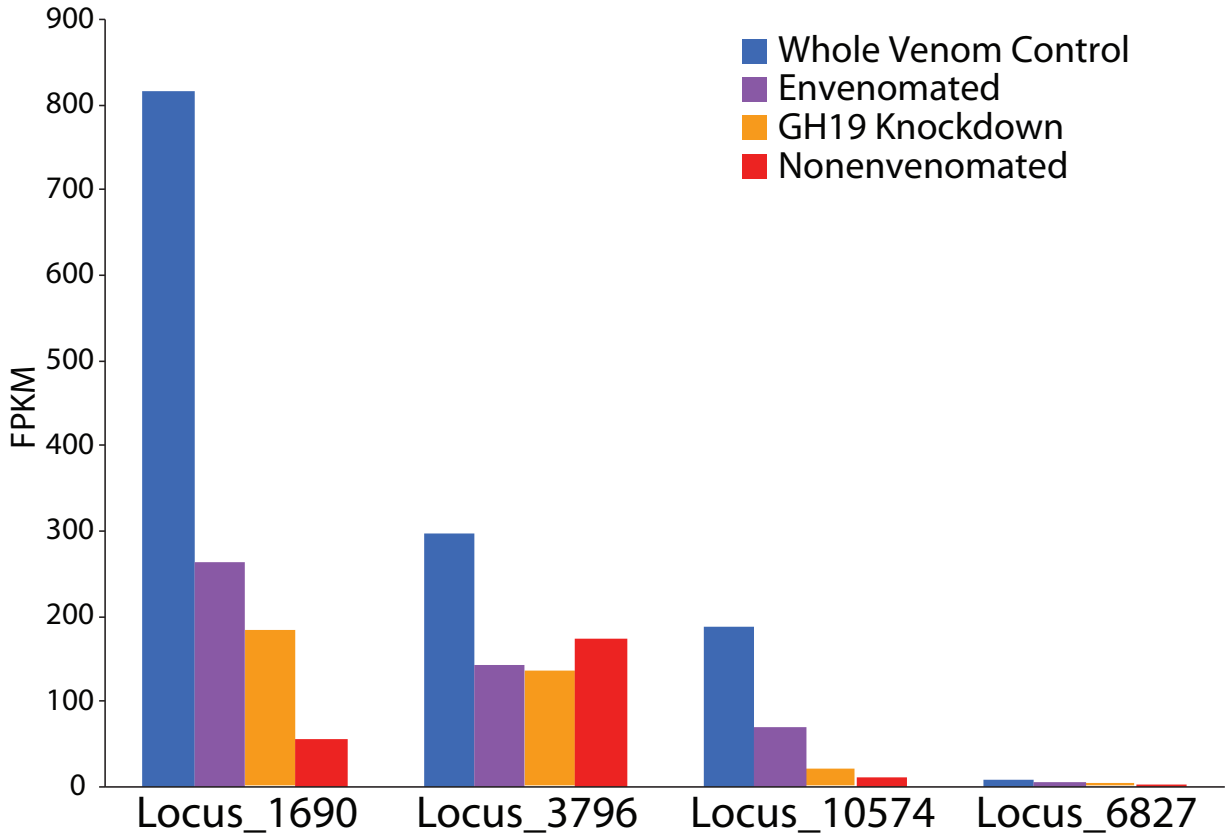
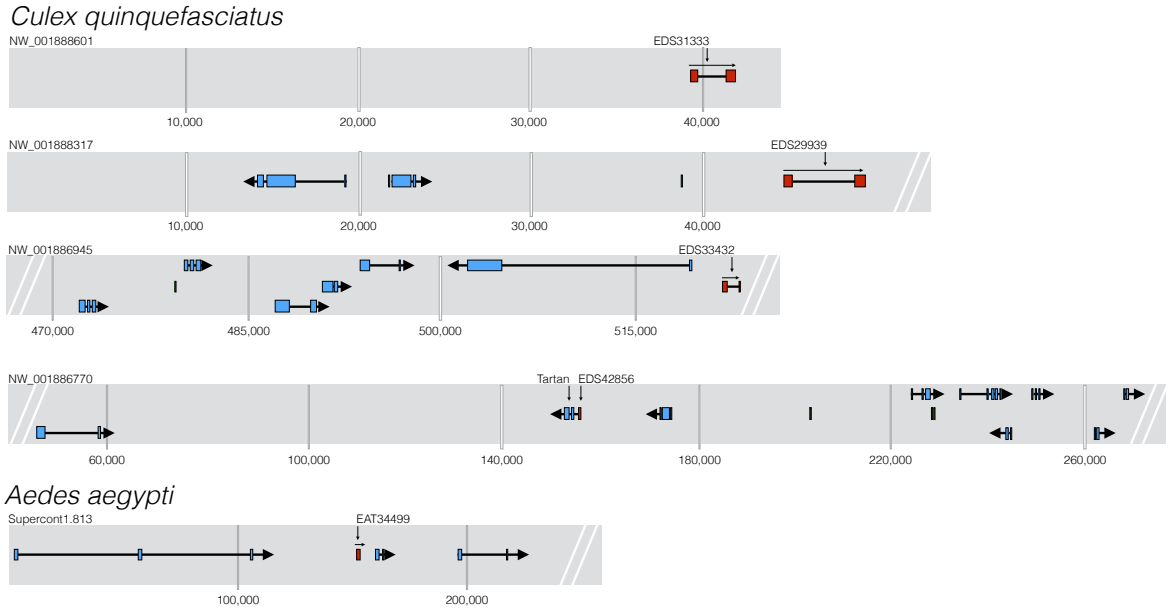


Fig. S4. Sequences and maximum likelihood tree used for the CodeML analysis in PAML.



121
122 Fig. S5. **Genes without annotation of *Sarcophaga bullata*, the fly host, that are significantly**
123 **differentially expressed upon envenomation by *N. vitripennis* with GH19 chitinase knocked**
124 **down via RNAi.** Comparison of host gene expression 72 hours after envenomation with the
125 complete venom repertoire "Envenomated", with *LacZ* RNAi control venom "Whole Venom
126 Control", and venom depleted of GH19 chitinase by RNAi "EC Knockdown". The
127 "Nonenvenomated" represents gene expression in normally developing hosts at the same age.
128 Expression is measured in fragments per kilobase per million (FPKM). The "Envenomated" and
129 "Nonenvenomated" data are from Martinson et al. (2014).
130
131
132

133



134 Fig. S6. Genomic position of GH19 chitinase genes in mosquitoes.

135

136

137

138

139

140

141

142

143

144

145

146

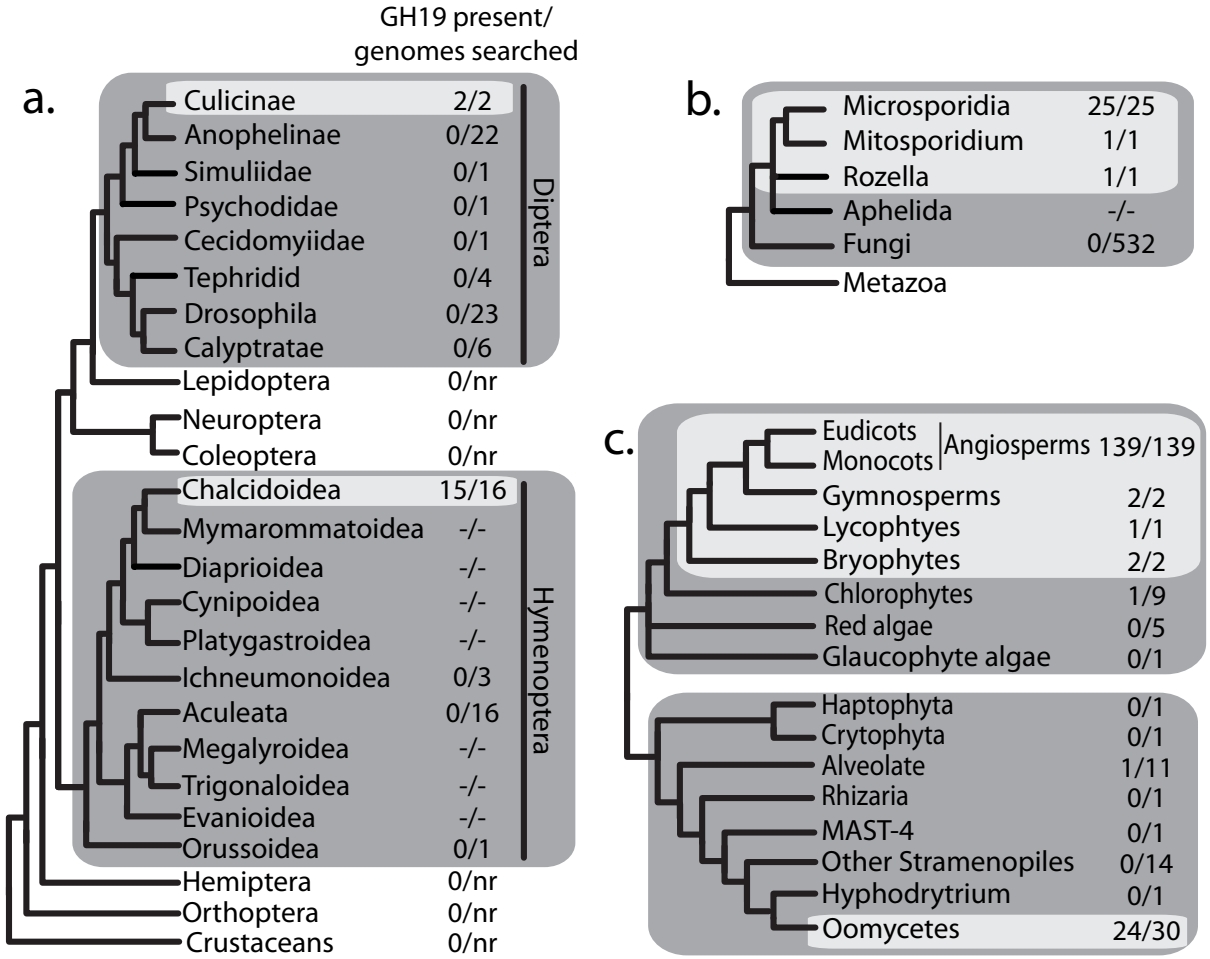


Fig. S7. The number of species in which a GH19 chitinase copy was found compared to the number of genomes that were searched (or the NCBI nr database) across different sections of the tree of life. Phylogenies adapted from (Wiegmann, et al. 2011; Klopfstein, et al. 2013; Misof, et al. 2014; Peters, et al. 2014; Roy, et al. 2014).

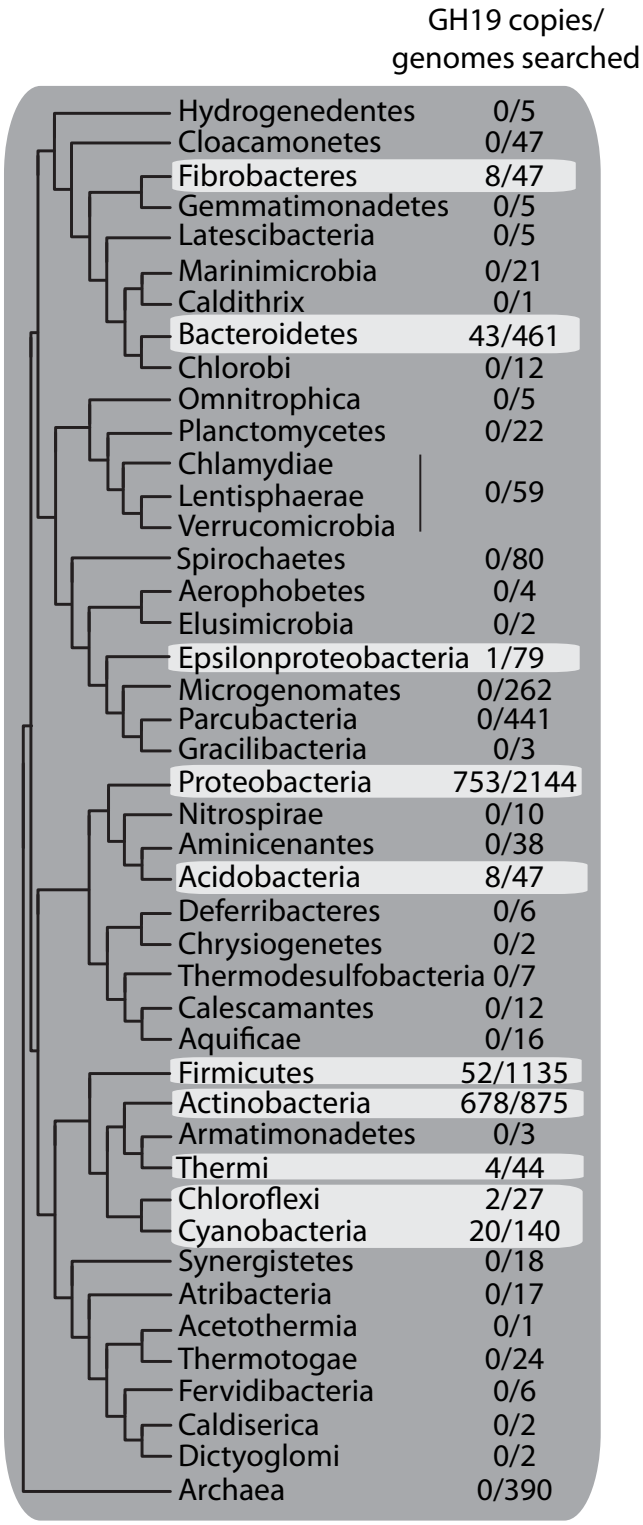


Fig. S8. The number of GH19 chitinase copies that were found compared to the number of genomes that were search across Bacteria and Archaea. Phylogeny adapted from (Rinke, et al. 2013).

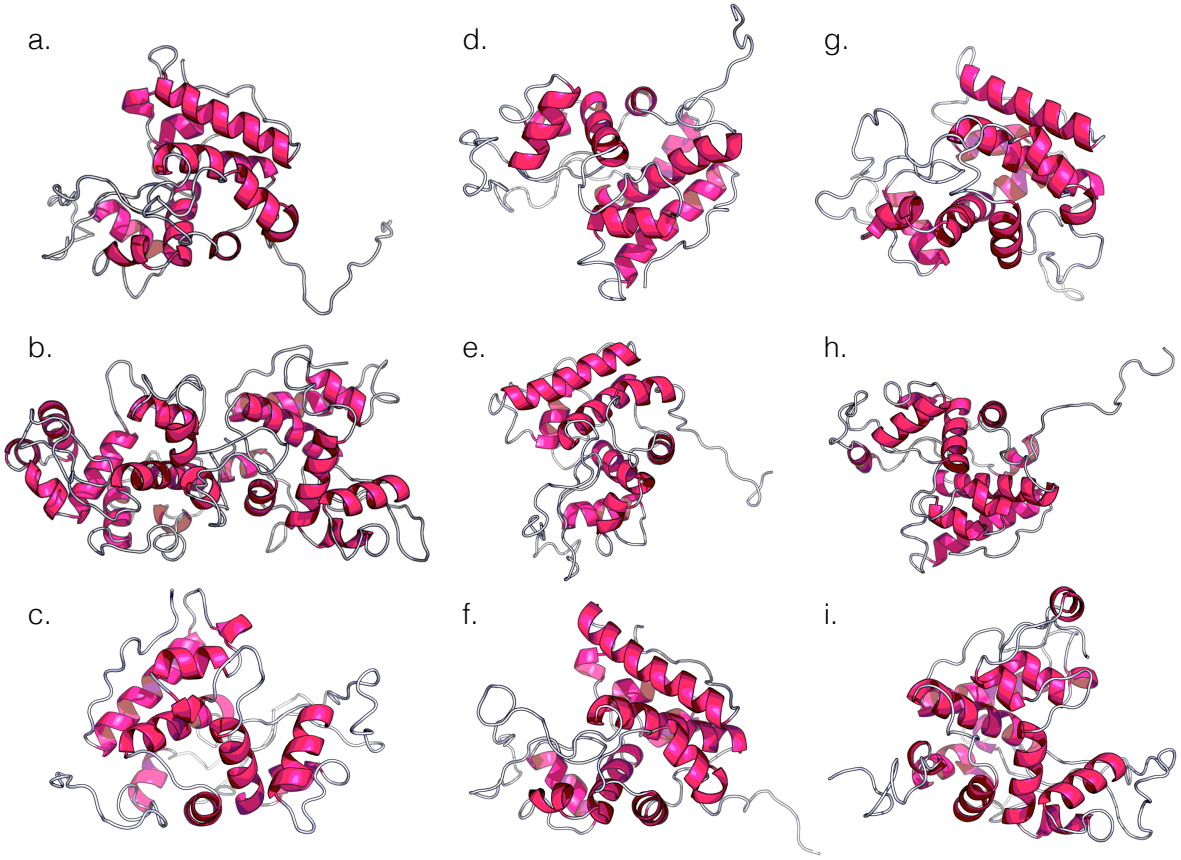


Fig. S9. Predicted protein structures for *Nasonia vitripennis* (a), *Culex quinquefasciatus* (b), *Rozella allomycis* (c), *Nosema apis* (d), *Vavraia culicis* (e), *Phytophthora parasitica* (f), *Physcomitrella patens* (g), *Streptomyces coelicolor* (h), *Amborella trichopoda* (i). Each sequence has a strong match (p-value < 3e-08) to a GH19 chitinase structural models in either bacteria or plants (2dkv, 2cjl, 3wh1, 1wvu, 4dwx).

173 **SUPPLEMENTAL TABLES**

Table S1: Intron number and size in the GH19 chitinase gene copies in three available chalcidoid genomes. Columns indicate name of the GH19 chitinase copy in each genome assembly, the scaffold number of the gene, the scaffold length, number and length of intron, whether there are immediate genes with top hits to Hymenoptera, and the amino acid position in the *N. vitripennis* venom GH19 chitinase does the intron start. Superscripts indicate that the copies are located close together with *913bp, ^3382bp, and ~2567bp apart.

	Scaffold	Scaffold length	Intron number	Avg. Intron length	Flanking wasp genes	Start aa of intron
Trichogramma pretiosum						
TPRE001600-RA	3	7060343	1	35	yes	80
TPRE001608-RA(a)*	3	7060343	1	74	yes	80
non-annotated*	3	7060343	1	60	yes	158
TPRE005901-RA	27	4034374	1	92	yes	80
Copidosoma floridanum						
CFLO001724-Ra(a)^	30	3281913	1	75	yes	60
CFLO001724-Ra(b)^	30	3281913	1	75	yes	80
CFLO004926-RA	134	2271108	1	209	no	80
Nasonia vitripennis						
Nasvi2EG021382~	143	399086	1	491	yes	60
Nasvi2EG021383~	143	399086	2	497	yes	80, 103

174

175