Title: **Horizontal transfer of prokaryotic cytolethal distending toxin B genes to eukaryotes**

Authors: Kirsten I. Verster[1,*], Jennifer H. Wisecaver[2], Rebecca P. Duncan[1], Marianthi Karageorgi[1], Andrew D. Gloss[3], Ellie Armstrong[4], Donald K. Price[5], Aruna R. Menon[1], Zainab M. Ali[1], Noah K. Whiteman[1,*]

Affiliations:

[1] Department of Integrative Biology, University of California – Berkeley, Berkeley, CA

[2] Department of Biochemistry, Purdue University, West Lafayette, IN

[3] Department of Ecology and Evolution, University of Chicago, Chicago, IL

[4]Department of Biology, Stanford University, Stanford, CA

[5]School of Life Sciences, University of Nevada – Las Vegas, Las Vegas, NV

*Correspondence to: kiv@berkeley.edu (Kirsten I. Verster); whiteman@berkeley.edu (Noah K. Whiteman

Abstract:

Cytolethal distending toxins (CDTs) are tripartite eukaryotic genotoxins encoded in diverse bacterial and phage genomes. The cdtB subunit is a DNAse that causes eukaryotic cell cycle arrest and apoptosis, and in one context, is associated with resistance against parasitoid wasp infections. Here we report the discovery of functional *cdtB* copies in the nuclear genomes of insect species from two distantly related insect orders, including fruit flies (Diptera: Drosophilidae) and aphids (Hemiptera: Aphididae). Insect cdtB copies are most closely related to bacteriophage copies, were horizontally transferred to insect genomes > 40 million years ago and encode a protein that retains ancestral DNase activity. This phage-derived toxin has been domesticated by diverse insects and we hypothesize that it is used as a defensive weapon against parasitoid wasps.

One Sentence Summary: We report horizontal transfer of the gene *cytolethal distending toxin B*, which encodes a DNase, into eukaryotic genomes from bacteriophage.

Significance: Cytolethal distending toxins (CDTs) are secreted by diverse pathogenic bacterial species to kill animal cells. The cdtB subunit enters cell nuclei, damaging the DNA and leading to mitotic arrest and apoptosis. In the pea aphid, a bacterial endosymbiont provides protection against wasp attack, possibly via *cdtB*. We discovered that this same endosymbiont-encoded lineage of *cdtB* was transferred to the genomes of Diptera and Hemiptera species and retains ancestral DNase activity. This is the first report of *cdtB* outside of bacteria or phages. A toxin that first evolved to kill eukaryotic cells has been co-opted by insects, potentially to their benefit.

Main Text

Cytolethal distending toxins (CDTs) are widespread intracellular-acting eukaryotic genotoxins encoded by a gene family restricted to Actinobacteria, Proteobacteria and bacteriophage genomes (*1*). CDTs are found in diverse pathogens, including *Campylobacter jejuni*, *Escherichia coli, Salmonella enterica*, and *Yersinia pestis* and may be a cause of irritable bowel syndrome (*1*). CDT holotoxin is an AB_2 toxin typically encoded in a three-gene operon (*cdtA, cdtB,* and *cdtC*) (*2*) and cdtB is the catalytic subunit necessary for DNase activity (*3*, *4*). CdtB nicking leads to DNA damage in eukaryotic cells followed by cell cycle arrest, cellular distention and death (*5*).

Although cdtB is a eukaryotic genotoxin, in one context it is associated with increased fitness of eukaryotes. Some strains of the bacterium *Candidatus* Hamiltonella defensa, a secondary endosymbiont of the pea aphid (*Acyrthosiphon pisum*), are infected with strains of the lysogenic bacteriophage APSE (*6*, *7*). APSE-positive *Ca.* H defensa strains confer protection from attack by parasitoid braconid wasps that insert eggs into aphids (*8*). Comparative genomic studies point to *cdtB*, which is encoded in the genome of phage strain APSE-2, as a likely candidate underlying this protective effect (*6–8*).

We used a sequence similarity-based screen (*9*) to identify a *cdtB* homolog as a horizontal gene transfer (HGT) candidate in a *de novo* genome assembly of the drosophilid fly *Scaptomyza flava*. To identify *cdtB* copies in genomes of other eukaryotes, we executed TBLASTN (*10*) searches of the NBCI refseq database (which includes all eukaryotes), all NCBI '*Drosophila*' genomes, and the genomes of 11 unpublished Hawaiian *Drosophila* species. We found high-confidence hits to *cdtB* homologs in the drosophilid species *Dr. ananassae*, *Dr. bipectinata* (both in the ananassae subgroup) and *Dr. biarmipes*, the Hawaiian *Dr. primaeva*, and

the aphid species *Myzus persicae* (**Table S1a)**. We also discovered *cdtB* orthologs in the transcriptomes of two other species in the ananassae subgroup, *Dr. pseudoananassae* and *Dr. ercepeae* (*11*). We subsequently searched all available AphidBase genomes and found single high-confidence hits to *cdtB* homologs in the Russian wheat aphid (*Diuraphis noxia*) and the black cherry aphid (*M. cerasi*), both in the Macrosiphini (**Table S1b**).

Putative HGT events can be due to microbial contamination arising from low-quality genome assemblies (*12*), so we used several methods to address these possibilities (*9*). First, *cdtB* was identified on scaffolds in species with high-quality genome assemblies (**Table S2)**. The presence of *cdtB* was verified by PCR and Sanger sequencing of both genomic and complementary DNA **(Table S3; Figure S1**). *CdtB*, when present, was found in all transcriptomes except that of *Di. noxia* (**Table S1**). The transcriptome libraries we searched were enriched for polyadenylated mRNA, suggesting insect *cdtB* was not due to bacterial contamination since bacteria typically lack 3'- polyA tails (*13*). Additionally, mRNA sequences of *cdtB* from all insect species (other than *S. flava*) contain at least three exons separated by intronic splice sites (*14*), which are rare in bacteria. The absence of *cdtB* transcripts in *Di. noxia,* coupled with a frame-shifting deletion and stop codon in the first (and only) predicted exon suggests that this *cdtB* fragment is a pseudogene in this species.

Phylogenetic conflict between gene tree and species tree topologies provides additional support for HGT (*15*). To evaluate this and determine the potential source of insect-encoded *cdtB*, we reconstructed a cdtB protein phylogeny using all available sequences (*9*). Viral, bacterial, and metazoan cdtB sequences were downloaded from the NCBI refseq protein database, aligned and used to create a protein tree (**Figure 1**, full phylogeny in **Figure S2**). The cdtB phylogeny reveals that all insect cdtB sequences form a clade with cdtB sequences from

Ca. H. defensa and APSE-2. A HGT event from an APSE-2 ancestor to eukaryotes is further supported by the case of *Dr. bipectinata,* in which two *cdtB* copies are present in tandem array. One of the two *cdtB* copies in *Dr. bipectinata* is fused with a homolog of an unrelated AB toxin, *apoptosis inducing protein 56,* found immediately downstream of *cdtB* in *Ca.* H. defensa. This chimeric *ctdB+aip56* sequence is expressed as mRNA in *Dr. bipectinata*. Synteny between *Dr. bipectinata* and *Ca.* H. defensa suggests the two genes were horizontally transferred together (see **Supplementary Text**) from a bacterial or phage ancestor prior to the divergence of the extant ananassae spp. subgroup *Drosophila* species. This hypothesis is supported by the presence of homologous *cdtB+aip56* chimeric sequences in two other ananassae subgroup species, though it has been lost in *D. ananassae*.

Our data suggest two independent acquisitions of intron-bearing and intronless insect *cdtB*. The cdtB phylogeny resolves two insect-encoded sub-clades, one containing all cdtB sequences encoded by insect-encoded, intron-bearing *cdtB* copies (*Myzus* spp.*, Dr. biarmipes,* and ananassae spp.) and the other containing all intron-less insect-encoded cdtB copies (*Scaptomyza* spp. + *D. primaeva*), which is in turn sister to the clade containing cdtB from *Ca.* H. defensa and APSE-2 genomes. Furthermore, an approximately unbiased test forcing monophyly of drosophilid cdtB is slightly worse (p=0.059) than the recovered cdtB phylogeny, suggesting that the intronless cdtB and the intron-bearing *cdtB* were independently transferred into insects.

In order to understand the number and timing of horizontal transfer of *cdtB* in insects, we reconstructed drosophilid and aphid species phylogenies and mapped *cdtB* evolution on these trees (*9*). We first constructed a drosophilid species phylogeny including all *Drosophila* and *Scaptomyza* genomes scanned for *cdtB*. We performed ancestral state reconstruction (ASR) for

the origin of *cdtB* to estimate the number and timing of HGT events across the drosophilid species phylogeny. This analysis, coupled with a clear pattern of conserved synteny within clades, suggests that *cdtB* was acquired three times in drosophilids: (1) prior to the divergence of the ananassae subgroup (94% posterior clade probability, or PP) ca. 21 million years ago (mya) (*16*), (2) following the split between *Dr. biarmipes* and *Dr. suzukii* (98% PP) ca. 7.3 ± 2.5 mya (*17*), and (3) in an ancestor common to *S. flava* and *Dr. primaeva* (13% PP) ca. 24 ± 7 mya (*18*) **(Figure 2A**). While the likelihood that *cdtB* was present in the common ancestor of *Dr. primaeva* and *S. flava* is low based on ASR, synteny suggests that a single HGT event occurred in a common ancestor of these two species. None of the genomes (out of 10 surveyed) from the more recently derived Hawaiian *Drosophila* species sister to *Dr. primaeva* were found to encode a *cdtB* copy. Thus, *cdtB* was most likely lost prior to the divergence of the picture wing clade, ca. 7 ± 4 mya (*18*). We did not perform ASR in aphids due to limited availability of sequenced aphid genomes. However, *cdtB* was syntenic in *Di. noxia, M. cerasi* and *M. persicae,* distantly related members of the Macrosiphini. We hypothesize that *cdtB* was horizontally transferred into a common ancestor of these three aphid species (41 ± 5 mya (*19*)). While a functional copy was retained in *M. persicae* and *M. cerasi,* it was pseudogenized in *Di. noxia* and lost completely in *A. pisum* **(Figure 2B**).

Interestingly, *cdtB* copies with three exons (*Myzus* spp.*, Dr. biarmipes,* and ananassae spp.) share identical splice junctions (**Figure S3**), which indicates either convergent origins of a modular exonic structure or that intron-bearing *cdtB* copies share a common ancestor and have been transferred horizontally between these distantly related insect lineages after an initial HGT event into one insect (*20*). HGT within eukaryotes could be mediated by several mechanisms,

including predaceous mites (*21*), bracovirus (by parasitoid wasps intermediaries) and helitrons (*22*). We illustrate hypotheses on the order and timing of *cdtB* HGT in **Figure 2C**.

There are many examples of genes derived from prokaryote-to-eukaryote HGT events stably integrating into nuclear genomes, and this process often involves optimizing the transferred genes for expression in eukaryotic cells (*23*). All insect-encoded *cdtB* copies exhibit features common to eukaryotic transcription initiation and termination (**Figure S4**; **Supplementary Text**). Additionally, insect *cdtB* copies have polyadenylated mRNA, 5' and 3' untranslated regions, and introns (except for *Scaptomyza* spp. + *Dr. primaeva*), which may modulate eukaryotic transcription/translation (*13*, *24*, *25*).

Expression patterns of HTGs often evolve to become finely tuned to eukaryotic cellular environments (*24*, *26*). We evaluated if *cdtB* shows differential expression patterns throughout development in two drosophilids that represent species with intron-bearing and intronless *cdtB* (*9*). Consistent with a potential role in parasitoid resistance, we predicted that *cdtB* expression would be highest in larvae, the developmental stage most prone to parasitoid attack in drosophilids (*27*). We used RT-qPCR in larvae, pupae, and adult males and females of *S. flava* and *Dr. ananassae* and found that *cdtB* expression was indeed highest in larvae of both species **(Figure 3)**.

A critical aspect of cdtB cytotoxicity is its DNase activity, which induces double-strand breaks that can lead to cell cycle arrest, cellular distention and death (*5*). Residues in cdtB involved in enzyme catalysis, DNA binding, and metal ion binding are critical in causing mitotic arrest in eukaryotic cells and are homologous to those in DNase I. To determine if insect-encoded cdtB is a DNase, we aligned cdtB from insect lineages and other bacterial species whose DNase and cytotoxic activity are well-characterized and found that residues necessary for DNase

activity are highly conserved in all insect copies (**Figure 4A, Figure S7**). To determine if these conserved residues correspond to DNase activity, we heterologously expressed and purified His-tagged cdtB from *Dr. ananassae* (**Figure S5**) in *E. coli* and utilized an agarose gel-based assay to determine its nuclease activity *in vitro* (*9*). We incubated *Dr. ananassae* cdtB (and *E. coli* cdtB as a positive control) with supercoiled plasmid pGEM-7zf(+) (Promega) at both 28°C and 37°C for 2 h. Supercoiled plasmid migrates more rapidly through a gel than nicked plasmid, which has greater surface area from relaxed superhelical tension (*4*). We predicted incubation of supercoiled (sc) plasmid with cdtB would result in a greater proportion of nicked plasmid (open coiled, or oc) isoforms. As expected, purified *Dr. ananassae* cdtB showed DNase activity *in vitro* (**Figure 4B**). Incubation at 28°C resulted in higher *Dr. ananassae* cdtB activity than *E. coli* and vice versa at 37°C, which may be a consequence of adaptation to insect and mammalian body temperatures, respectively (see **Figure S6**).

The maintenance of *cdtB* in diverse insect genomes for millions of years suggests that it has an adaptive function. One clear possibility is that *cdtB* plays a role in parasitoid wasp resistance, as it does in the bacterial secondary symbionts of pea aphids (*7*, *8*). Given that many drosophilid and aphid species are at high risk of parasitoid wasp attack (*27*), cdtB may facilitate protection, through DNase activity against the parasitoid wasp egg or larva. In a parasitization assay, 100% of *Dr. ananassae* and *Dr. biarmipes* survived attack by both the generalist *Leptopilina heterotoma* and specialist *L. boulardi* (*28*). It is possible, although speculative, that this unusual level of resistance is facilitated by cdtB.

To our knowledge, this is the first report of the horizontal transfer of *cdtB* from prokaryotes to eukaryotes. The domestication of *cdtB* in insects is remarkable given that the toxin originally evolved to destroy, not benefit, eukaryotic cells. Given the wealth of genetic and

genomic resources available within drosophilids and aphids, horizontally transferred *cdtB* promises to be an exciting experimentally tractable system in which to explore the biology of a novel eukaryote-adapted toxin, which also has potential in targeting and killing tumor cells in humans (*29*).

References and Notes

1. L. Guerra, X. Cortes-Bratti, R. Guidi, T. Frisan, The biology of the cytolethal distending toxins. *Toxins* **3**, 172-190 (2011).
2. M. Lara-Tejero, J. E. Galán, CdtA, CdtB, and CdtC form a tripartite complex that is required for cytolethal distending toxin activity. *Infect. Immun.* **69**, 4358–4365 (2001).
3. M. Lara-Tejero, J. E. Galán, A bacterial toxin that controls cell cycle progression as a Deoxyribonuclease I-like protein. *Science* **290**, 354–357 (2000).
4. C. A. Elwell, L. A. Dreyfus, DNase I homologous residues in CdtB are critical for cytolethal distending toxin-mediated cell cycle arrest. *Mol. Microbiol.* **37**, 952–963 (2000).
5. R. N. Jinadasa, S. E. Bloom, R. S. Weiss, G. E. Duhamel, Cytolethal distending toxin: a conserved bacterial genotoxin that blocks cell cycle progression, leading to apoptosis of a broad range of mammalian cell lineages. *Microbiology* **157**, 1851–1875 (2011).
6. N. A. Moran, P. H. Degnan, S. R. Santos, H. E. Dunbar, H. Ochman, The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proc. Natl. Acad. Sci.* **102**, 16919–16926 (2005).
7. P. H. Degnan, N. A. Moran, Diverse phage-encoded toxins in a protective insect endosymbiont. *Appl. Environ. Microbiol.* **74**, 6782–6791 (2008).

8. K. M. Oliver, P. H. Degnan, M. S. Hunter, N. A. Moran, Bacteriophages encode factors required for protection in a symbiotic mutualism. *Science* **325**, 992–994 (2009).

9. Materials and methods are available as supplementary materials.

10. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

11. S. Signor, T. Seher, A. Kopp, Genomic resources for multiple species in the *Drosophila ananassae* species group. *Fly* **7**, 47–57 (2013).

12. G. Koutsovoulos, S. Kumar, D. R. Laetsch, L. Stevens, J. Daub, C. Conlon, H. Maroon, F. Thomas, A. A. Aboobaker, M. Blaxter. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini. Proc. Natl. Acad. Sci.* **113**, 5053–5058 (2016).

13. M. Dreyfus, P. Régnier, The poly(A) tail of mRNAs: Bodyguard in eukaryotes, scavenger in bacteria. *Cell* **111**, 611–613 (2002).

14. S. Kitamura-Abe, H. Itoh, T. Washio, A. Tsutsumi, M. Tomita, Characterization of the splice sites in GT-AG and GC-AG introns in higher eukaryotes using full-length cDNAs. *J. Bioinform. Comput. Biol.* **2**, 309–331 (2004).

15. E. A. Gladyshev, M. Meselson, I. R. Arkhipova, Massive horizontal gene transfer in bdelloid rotifers. *Science* **320**, 1210–1213 (2008).

16. M. Nozawa, M. Kumagai, T. Aotsuka, K. Tamura, Unusual evolution of interspersed repeat sequences in the *Drosophila ananassae* subgroup. *Mol. Biol. Evol.* **23**, 981–987 (2006).

17. L. Ometto, A. Cestaro, S. Ramasamy, A. Grassi, S. Revadi, S. Siozios, M. Moretto, P.

Fontana, C. Varotto, D. Pisani, T. Dekker, N. Wrobrei, R. Viola, I. Pertot, D. Cavalieri, M. Blaxter, G. Anfora, O. Rota-Stabelli, Linking genomics and ecology to investigate the complex evolution of an invasive *Drosophila* pest. *Genome Biol. Evol.* **5**, 745–757 (2013).

18. K. N. Magnacca, D. K. Price, Rapid adaptive radiation and host plant conservation in the Hawaiian picture wing *Drosophila* (Diptera: Drosophilidae). *Mol. Phylogenet. Evol.* **92**, 226–242 (2015).

19. Z. Ren, A. J. Harris, R. B. Dikow, E. Ma, Y. Zhong, J. Wen, Another look at the phylogenetic relationships and intercontinental biogeography of eastern Asian – North American *Rhus* gall aphids (Hemiptera: Aphididae: Eriosomatinae): Evidence from mitogenome sequences via genome skimming. *Mol. Phylogenet. Evol.* **117**, 102–110 (2017).

20. M. Long, E. Betrán, K. Thornton, W. Wang, The origin of new genes: Glimpses from the young and old. *Nat. Rev. Genet.* **4**, 865–875 (2003).

21. M. A. Houck, J. B. Clark, K. R. Peterson, M. G. Kidwell, Possible horizontal transfer of *Drosophila* genes by the mite *Proctolaelaps regalis*. *Science* **253**, 1125–1128 (1991).

22. L. Gasmi, H. Boulain, J. Gauthier, A. Hua-Van, K. Musset, A. K. Jakubowska, J. Aury, A. Volkoff, E. Huguet, S. Herrero, J. Drezen, Recurrent domestication by Lepidoptera of genes from their parasites mediated by bracoviruses. *PLoS Genet.* **11**, 1–32 (2015).

23. Y. Sasakura, Y. Ogura, N. Treen, R. Yokomori, S. Park, K. Nakai, H. Saiga, T. Sakuma, T. Yamamoto, S. Fujiwara, K. Yoshida, Transcriptional regulation of a horizontally transferred gene from bacterium to chordate. *Proc. R. Soc. B*. **283**, 20161712 (2016).

24. F. Husnik, J. P. McCutcheon, Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.* **16**, 67–79 (2017).

25. L. Barrett, S. Fletcher, S. Wilton, Untranslated Gene Regions and Other Non-coding Elements (Springer, Basel, 2013).

26. N. Wybouw, Y. Pauchet, D. G. Heckel, T. Van Leeuwen, Horizontal gene transfer contributes to the evolution of arthropod herbivory. *Genome Biol. Evol.* **8**, 1785–1801 (2016).

27. Y. Carton, M. Poirié, A. J. Nappi, Insect immune resistance to parasitoids. *Insect Sci.* **15**, 67–87 (2008).

28. T. A. Schlenke, J. Morales, S. Govind, A. G. Clark, Contrasting infection strategies in generalist and specialist wasp parasitoids of *Drosophila melanogaster*. *PLoS Pathog.* **3**, 1486–1501 (2007).

29. C. Bachran, R. Hasikova, C. E. Leysath, I. Sastalla, Y. Zhang, R. J. Fattah, S. Liu, S. H. Leppla, Cytolethal distending toxin B as a cell-killing component of tumor-targeted anthrax toxin fusion proteins. *Cell Death Dis.* **5**, e1003 (2014).

30. H. Kim, S. Lee, Y. Jang, Macroevolutionary patterns in the Aphidini aphids (Hemiptera: Aphididae): diversification, host association, and biogeographic origins. *PLoS One*. **6**, e24749 (2011).

31. B. N. Singh, Species and genetic diversity in the genus *Drosophila* inhabiting the Indian subcontinent. *J. Genet.* **94**, 351–361 (2015).

32. H. F. van Emden, V. F. Eastop, R. D. Hughes, M. J. Way, The ecology of *Myzus persicae*. *Annu. Rev. Entomol.* **14**, 197–270 (1969).

Acknowledgments:

We thank Dr. Chris Jeans and Brooks Bond-Watts for their preparation of purified cdtB. Timothy O'Connor provided bioinformatics advice, Raoul O. Martin provided biochemistry advice, Julianne N. Pelaez assisted in phylogenetic reconstruction, and Anthony T. Iavarone performed mass spectrometry analysis. Coco Verster provided field assistance in acquiring specimens. Dr. Artyom Kopp and Dr. Doris Bachtrog provided *Drosophila* specimens. The *Myzus persicae* transcriptome assembly was provided by Alex Wilson and Honglin Feng. Dr. Nancy Moran ran BLAST searches of unpublished aphid genomes that corroborated our conclusions from this study. Dr. Naomi E. Pierce and Dr. Frederick M. Ausubel provided early support for obtaining a genome sequence from *S. flava.* **Funding:** KIV was supported by a National Science Foundation Graduate Research Fellowship and grants from Sigma Xi (University of California – Berkeley chapter) and the Animal Behavior Society. RPD was supported by the Miller Institute for Basic Research in Science at the University of California, Berkeley. Research was supported by the National Institute of General Medical Science of the National Institutes of Health award number R35GM119816 to NKW. **Author contributions:** KIV, ADG, MK, RPD, and NKW were involved in conceptualization of the project. KIV, JHW, RPD, MK, ZMA, EA, DKP and NKW conducted the investigations. KIV, RPD and NKW wrote the paper. All authors edited and approved the manuscript. **Conflict of interest:** KIV, ADG and NKW are inventors on a pending patent application related to this work, entitled "Cytolethal distending toxin B from insects for human cancer treatment". **Data and materials availability:** *CdtB* sequences from *Scaptomyza* species and *D. primaeva* were deposited to NCBI GenBank

under accession numbers MH884655-MH884659. *CdtB* codon-optimized oligos used for nuclease assays were deposited under GenBank accessions MH891796-MH891799.

Supplementary Materials:

Materials and Methods

Figures S1-S8

Tables S1-S9

References (*33-105*)

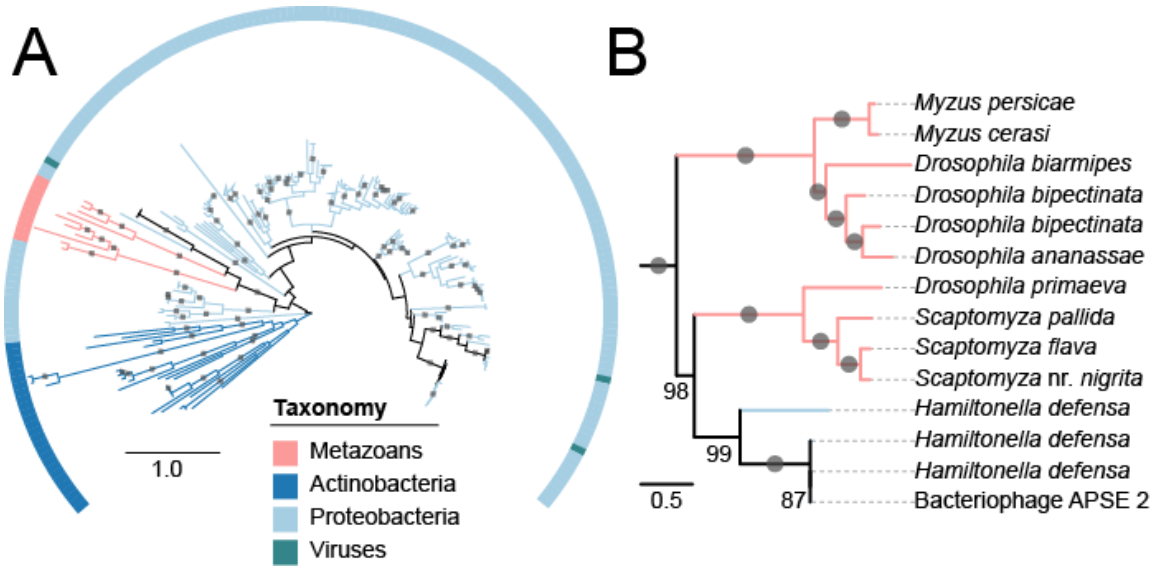


Fig. 1. CdtB protein phylogeny indicates HGT into insects from bacteria.

(A) ML phylogeny of cdtB from across the tree of life. Tree is midpoint rooted and branches with 100% bootstrap support are indicated by grey circles. Four clades consisting of highly similar sequences from Proteobacteria were collapsed for clarity. The full phylogeny is available in **Fig. S2.** (B) Detailed view of insect cdtB clades. Numbers below branches indicate percent bootstrap support when < 100.

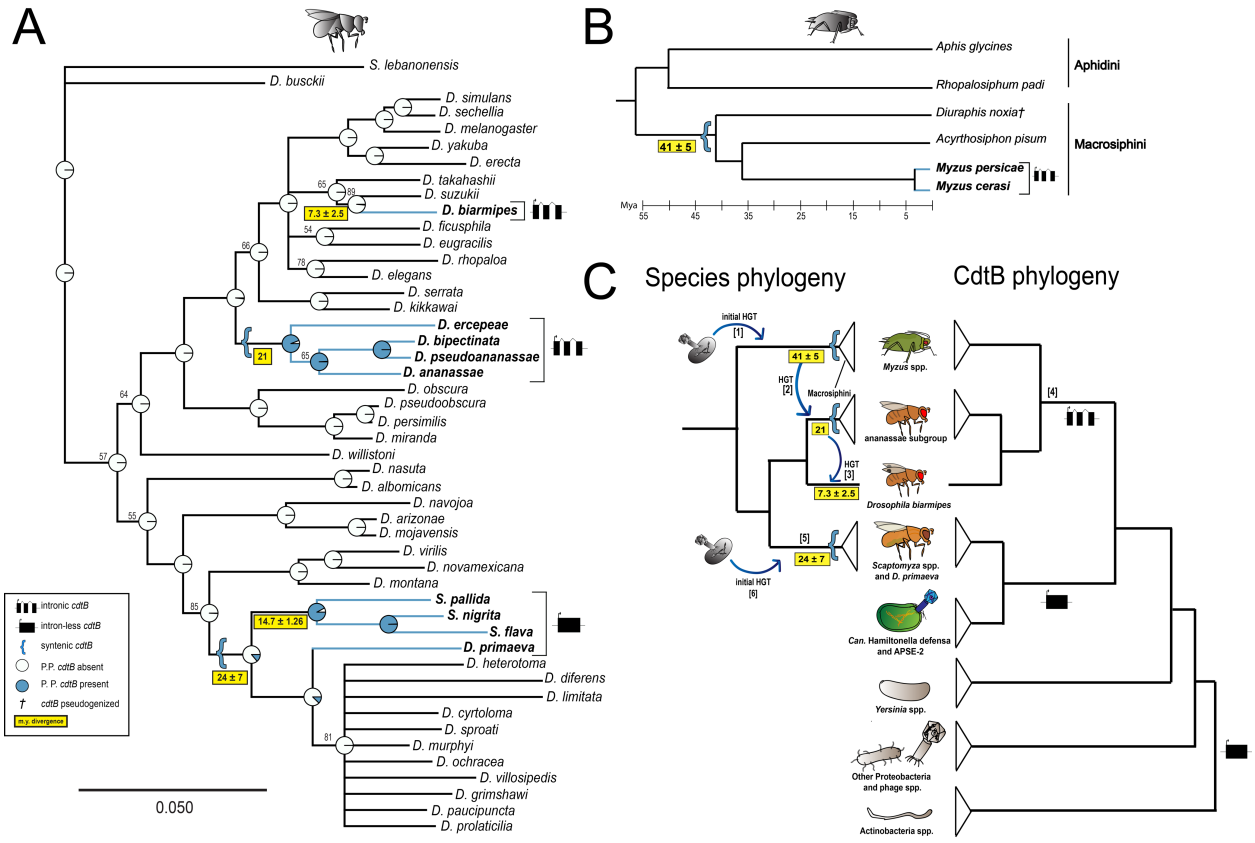


Fig. 2. Species phylogenies show *cdtB* was transferred into, and possibly between, genomes of distant insect lineages.

A. ML phylogeny of drosophilid species. Node labels indicate bootstraps if <90% or are collapsed to polytomies if <50%. ASR shows posterior probability (P.P.) of *cdtB* at nodes. **B.** Phylogeny of Aphidinae species. Branch lengths drawn approximately to scale using divergence dates from (*19*, *30*). **C.** Simplified paired cdtB and species phylogenies. Blue arrows suggest possible HGT directions and bracketed numbers are described here. Possible initial prokaryote-eukaryote HGTs are [1,6]. We hypothesize an initial HGT of *cdtB* from bacteria or phage integrated into an aphid nuclear genome [1] and was lost or pseudogenized in some aphid lineages (**2B**). We then posit an inter-ordinal transfer [2] from a *Myzus* spp. ancestor to an ananassae subgroup spp. ancestor, followed by inter-specific transfer [3] to a *D. biarmipes* ancestor. This transfer sequence is supported by subclade ages, conserved intron splice sites in [4], and the regional co-occurrence of these subclades (*31*, *32*). However, conserved exon structure in [4] could also arise from convergence. CdtB in [5] could have evolved independently, or was derived from the same HTG as [4] but failed to acquire introns.

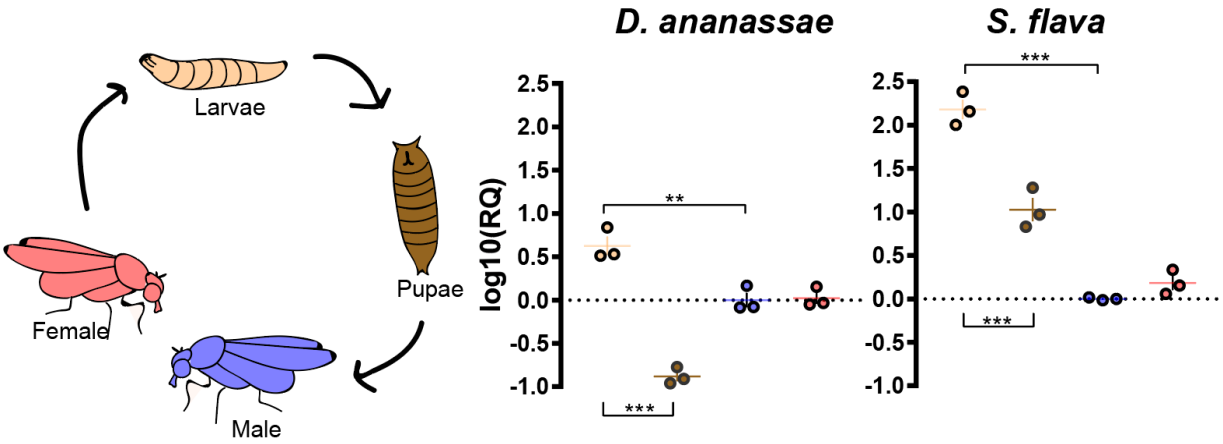


Fig. 3. *CdtB* is expressed most highly in the *Drosophila* larval stage.

Fold changes in expression of *cdtB* in two representative insect lineages (*Dr. ananassae* and *S. flava*) across development. Colors correspond to developmental stages in the left panel. Fold change is standardized against *rpl32* mRNA expression in males. Each dot represents one biological replicate. $P < 0.005$**, $P < 0.0001$***. All pairwise comparisons (except those between males and females) are significantly different, but are not marked for simplicity.

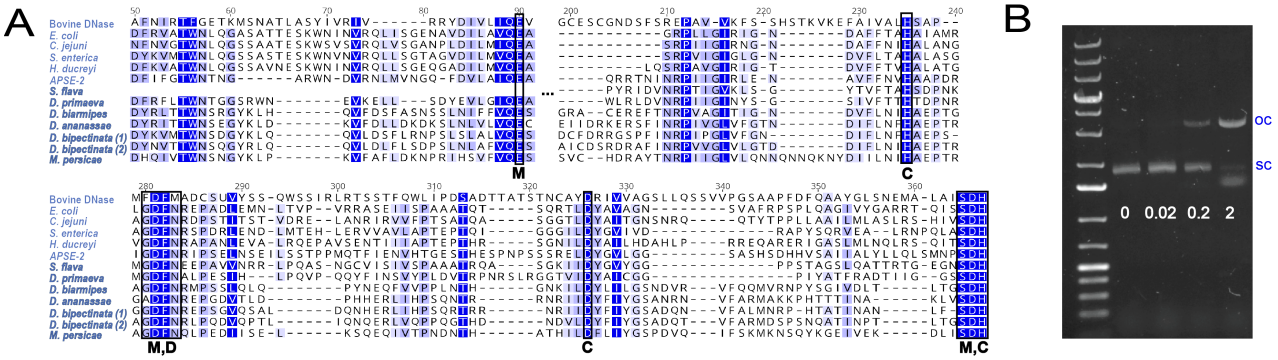


Fig. 4. Critical DNase residues are conserved in insect-encoded cdtB and confer DNase activity *in vitro*.

A. MUSCLE aligned amino acid sequence of DNase I and cdtB across taxa. Boxed residues are necessary for DNase activity of cdtB. Blue scale corresponds to similarity under the Blosum62 scoring matrix. Numbers correspond to alignment residue. Breaks in alignment are indicated by brackets. Species names in bold are eukaryotic. M = metal-ion binding residues, C=catalytic residues, D = DNA contact residues (*4*). **B.** Plasmid degradation following exposure to variable quantities (in μg) of cdtB from *Dr. ananassae* over 2 hrs. 0.8% agarose 1X TBE gels were stained with 0.01% SYBR™ Safe. OC = open-coil isoform, SC = supercoiled isoform.

Supplementary Materials for

Horizontal transfer of prokaryotic cytolethal distending toxin B genes to eukaryotes

Kirsten I. Verster*, Jennifer H. Wisecaver, Rebecca P. Duncan, Marianna Karageorgi, Andrew D. Gloss, Ellie Armstrong, Donald K. Price, Aruna R. Menon, Zainab M. Ali, Noah K. Whiteman*

Correspondence to: kiv@berkeley.edu; whiteman@berkeley.edu

This PDF file includes:

Materials and Methods

Materials and methods are described in the order they appear in the main text.

Initial detection of horizontally transferred protein coding genes in insects: *Scaptomyza flava*

From the annotated genome assembly of *S. flava* (GenBank Accession RKRM00000000.1), all predicted protein sequences were queried against a local copy of the NCBI refseq protein database (downloaded May 5, 2017) using phmmer, in the HMMER3 software suite (*33*), with acceleration parameters --F1 1e-5 --F2 1e-7 --F3 1e-10. A custom perl script sorted the phmmer results based on the normalized bitscore (*nbs*), where *nbs* was calculated as the bitscore of the single best-scoring domain in the hit sequence divided by the best bitscore possible for the query sequence (i.e., the bitscore of the query aligned to itself). The top ≤ 10,000 hits were retained for further analysis, saving no more than three sequences per unique NCBI Taxonomy ID.

The alien index score (*AI*) was calculated for each query protein (modified from Gladyshev et al., 2008). The *AI* is given by the formula: AI=nbsO-nbsM, where *nbsO* is the normalized bitscore of the best hit to a non-metazoan species, *nbsM* is the normalized bitscore of the best hit to a metazoan (skipping all hits to the Drosophilini tribe NCBI:txid46877). *AI* can range from 1 to -1 and is > 0 if the gene has a better hit to a non-metazoan, which is suggestive of either HGT or contamination in the assembly. To reduce the risk of contamination, genes were considered potential HGT candidates if they were assembled on scaffolds with ≥ 5 protein coding genes and the average *AI* of the scaffold was < 0.

Phylogenetic trees of protein sequences were constructed for all potential HGT candidates with *AI* > 0. Full-length proteins corresponding to the top 200 hits (E-value < 1×10^{-3}) to each query sequence were extracted from the local database using esl-sfetch (*33*). Protein sequences were aligned with MAFFT v7.310 using the E-INS-i strategy and the BLOSUM30 amino acid scoring matrix (*34*) and trimmed with trimAL v1.4.rev15 using its gappyout strategy (*35*). Proteins with trimmed alignments < 150 amino acids in length were excluded. The topologies of the remaining genes were inferred using maximum likelihood as implemented in IQ-TREE v1.5.4 (*36*) using an empirically determined substitution model and rapid bootstrapping (1000 replications). The phylogenies were midpoint rooted and branches with local support < 95 were collapsed using the ape and phangorn R packages (*37*, *38*). Phylogenies were visualized using ITOL version 3.0 (*39*) and inspected manually to identify phylogenetically supported HGT candidate proteins. The cdtB phylogeny was the only one that passed this manual inspection.

Identification of *cdtB* in aphid genomes and transcriptomes

An initial TBLASTN search using *S. flava* cdtB against NCBI nr resulted in hits to *Myzus persicae*, an aphid species, as well as other drosophilids (discussed below). We therefore further searched for *cdtB* in genomes and transcriptomes from representatives of Aphididae. Aphids were sampled based on availability of published or unpublished genomic resources, and included 11 species from three tribes and three subfamilies. Representatives from the subfamily Eriosomatinae, which is sister to the rest of aphids (*40*) were included in our sampling: *Pemphigus obesinymphae* and *Pemphigus populicaulis* (Subfamily: Eriosomatinae, Tribe: Pemphiginae), *Tamalia coweni* and *Tamalia inquilinus* (Subfamily: Tamaliinae), *Myzus persicae*, *Myzus cerasi, Diuraphis noxia* and *Acyrthosiphon pisum* (Subfamily: Aphidinae, Tribe: Macrosiphini), and *Aphis glycines, Aphis nerii,* and *Rhopalosiphum padi* (Subfamily: Aphidinae, Tribe: Aphidini). Genomes were sampled from *M. persicae* (*41*), *M. cerasi* (available on

aphidbase.com), *A. pisum* (*42*), *Di. noxia* (*43*), *Ap. glycines* (available on aphidbase.com), and *R. padi* (available on aphidbase.com). We sampled published transcriptomes from the remaining aphid species (*44–47*).

We searched genome or transcriptome assemblies for the presence of *cdtB* with TBLASTN searches using two different cdtB proteins as the query: cdtB from *M. persicae* (XP_022163116.1) and cdtB from the *Candidatus* Hamiltonella defensa phage APSE-2 (C4K6T7), since it is infects aphid species (*48*). CdtB full or partial hits were only found in three aphids with genome sequences (*M. persicae*, *M. cerasi*, and *Di. noxia*), so to assess if *cdtB* was expressed in those species, we searched transcriptome assemblies for each species with TBLASTN searches using the same query proteins (**Table S1c**). For *M. persicae*, we used the assembly from the previously published transcriptome (*49*), and for *M. cerasi* and *Di. noxia* we conducted *de novo* assemblies from previously published RNAseq data. We downloaded raw RNAseq reads for *M. cerasi* (BioProject PRJEB9912, runs ERR983165 (head), ERR983166 (head), ERR983167 (head), ERR983168 (whole body), ERR983169 (whole body), ERR983170 (whole body) (PRJEB9912) and *Di. noxia* (BioProject PRJNA233413, runs SRR1999270 (whole body) and SRR1999279 (whole body) (*43*) from the Sequence Read Archive on GenBank. All runs for each species were combined into a reference transcriptome in Trinity v. 2.4.0 (*50*) using the built in Trimommatic pipeline for quality trimming (default parameters) and *in silico* normalization.

Demonstrating *cdtB* is encoded in the nuclear genome of drosophilid species

Analysis of possible contamination by coverage depth analysis in S. flava.

For *S. flava,* we aligned long PacBio reads to the genome via Burrows-Wheelers alignment (*51*) to search for unusual coverage depth relative to neighboring genes, which can be a reflection of contamination (*12*). The region containing *S. flava cdtB* did not exhibit unusual coverage depth (Grubbs' test, $p>0.05$) (**Table S4**).

PCR and RT-PCR reaction conditions.

PCR reaction conditions were composed of: 4.2 µL nuclease-free water, 7.5 µL Failsafe Premix E (Epicentre), 1.2 µL each of F and R primers (IDT), 0.8 µL DNA, and 0.12 µL of *Taq* polymerase (New England Biolabs). Thermal cycler settings were: 5 m at 95°C and 30 cycles of 95°C for 30 s, Ta for 30 s, and 68°C for 30 s, followed by 5 m of extension at 68°C. The exception to this was with *S. flava* Intergenic PCR amplification. PCR reaction conditions were composed of: 12 µL nuclease-free water, 4 µL 5X Phusion HF buffer, 0.4 µL 10 mM dNTPs, 1 µL 10 µM Intergenic F and R primers (IDT), respectively, 0.6 µL DMSO, 0.2 µL Phusion DNA polymerase (New England Biolabs), and 0.8 µL template gDNA. Thermal cycler settings were: 30 s at 98°C and 30 cycles of 98°C for 10 s, 64.1°C for 30 s, 72°C for 2 m 40 s, followed by 10 m extension at 72°C.

1% agarose 1X TBE gels were prepared with Apex Agarose in 1X TBE buffer with 1 µL SYBR™ Safe staining gel per 10 mL of gel solution. 4 µL PCR product was mixed with 1 µL ThermoScientific 6X Loading Dye and run on 1% gels in Owl ™ EasyCast ™ B1 Mini Gel Electrophoresis System rigs at 120 V for 30 m. 5 µL of ladder was used (O'Gene Ruler 100 bp or O'Gene Ruler 1kb). Gels were visualized using AlphaImager ™ Gel Imaging System (Alpha Innotech). PCR amplicons were Sanger sequenced in both directions at the UC Berkeley DNA

Sequencing Facility using ABI dye terminator chemistry. Relevant gel images and primers are shown in **Figure S1** and **Table S3**.

CdtB phylogeny reconstruction and topology test

All insect-encoded cdtB protein sequences translated from nucleotides were queried against an updated local copy of the NCBI refseq protein database (downloaded August 1, 2018) using phmmer (*33*) and default parameters, saving no more than one sequence per unique NCBI Taxonomy ID. Full-length proteins were extracted from the local database using esl-sfetch (*33*), and results from each insect cdtB search were combined to yield a final cdtB sequence set for phylogenetic analysis. Sequences were aligned with MAFFT v7.310 using the L-INS-i strategy and the BLOSUM30 amino acid scoring matrix (*34*). A total of 15 proteobacterial hits were excluded due to poor alignment and the remaining sequences were trimmed to include only the conserved cdtB domain. MAFFT was then repeated. The topology of cdtB was inferred using maximum likelihood as implemented in IQ-TREE v1.5.4 (*36*) and RAxML v8.2.9 (*52*) using empirically determined substitution models. Ten independent searches with different starting trees were carried out using each program as recommended by (*53*). The likelihood scores of all trees were re-calculated using RAxML and the tree with the highest likelihood was selected as the best cdtB phylogeny. Lastly, 1000 non-parametric bootstrap replicates were performed in IQ-TREE on the final phylogeny (**Figure S2**).

Constrained phylogenetic trees in which the insect-encoded cdtB were forced to be monophyletic were also constructed. As with the best tree, ten independent searches with different starting trees were carried out using RAxML and IQ-TREE, and the tree with the highest likelihood given the constraint was selected as the constrained cdtB tree. The best and constrained trees were then compared in CONSEL v1.2 (*54*) using the approximately unbiased (AU) test (*55*).

Ancestral state reconstruction of *cdtB* in *Drosophila*

To construct a drosophilid species tree, DNA sequences from *adh, marf, COI, COII, 16s, cytb, gpdh, nd1*, and *nd2* (see **Table S5** for sources of phylogenetically informative genes) were aligned individually using default settings in MUSCLE (*56*) as implemented in Geneious (*57*). Alignments were visually inspected, manually trimmed and then concatenated. The final alignment included 48 species with 7479 nucleotide sites. The concatenated alignment was used to infer a drosophilid species phylogeny using maximum likelihood with the TN93 (*58*) model of nucleotide substitution in MEGA v10.0.4. The tree with the highest log likelihood (-60735.94) is shown with *Scaptodrosophila lebanonensis* as the outgroup. Branch lengths are drawn to scale and are measured in number of substitutions per site. Bootstraps values are shown (n=500) and those less than 50% were collapsed into polytomies. Maximum-likelihood ancestral state reconstruction of *cdtB* HGT occurrence was performed using the "rerooting Method" function under an equal-rates model in phytools (*59*) in R v. 1.1.456 (*60*).

Expression of *cdtB* throughout development

Sample preparation and collection.

Ca. 100 male and female *Dr. ananassae* (13-17 days old, the optimal egg-laying age as determined by a pilot study) were left in small embryo collection cages (Genesee Scientific, #59-100) for 6 hours with 60 x 15 mm Falcon polystyrene petri dishes filled with 3% agar in organic apple juice with a dab of Fleischmann's active dry yeast paste. After egg-laying, eggs were

cleaned and isolated in Corning Netwells inserts (#3477) and transferred onto 100 x 15 mm petri dishes with Nutri-Fly media (Genesee Scientific #66-112) prepared using standard protocols. For *S. flava,* >25 male and females (5-7 days old) were staged as above except petri dishes were filled with 3% agar and Arrowhead water with 5-9 Col-0 *Arabidopsis thaliana* leaves from adult plants submerged in the agar. For each species, we collected L2 (assessed by FBdv:00005338) (L), P-2 (assessed by (*61*)) (P), and virgin females (F) and males (M). For *Dr. ananassae* L, P, F and M, we collected 10, 5, 5, and 5 individuals, respectively, per replicate. For *S. flava,* we collected 4, 3, 3, and 3 individuals, respectively, per replicate. Samples were submerged in Ringer's solution prior to collection. Each species and developmental stage had three replicates. Experiments occurred at 25ºC under 14 h light:10 h dark cycles.

RNA extraction and cDNA synthesis.

Samples were washed in Ringer's solution again prior to RNA extraction. We performed RNA extraction using the Promega ReliaPrep RNA Tissue MiniPrep System following guidelines of protocol for samples < 5 mg. Final elution volume was 10 µL in nuclease-free water. RNA concentration and purity were quantified using NanoDrop ND-1000 (Thermo Fisher). We performed cDNA synthesis following standard protocols with the ProtoScript cDNA synthesis kit (NEB) using 1 µg of RNA for each sample. Synthesis of cDNA was confirmed via Qubit Fluorometer 2.0 (Invitrogen) using dsDNA HS Assay Kit (Thermo Fisher).

RT-qPCR primer design.

RT-qPCR primers for *cdtB* and *rpl32* were designed using GenScript Real- time PCR (TaqMan) Primer Design tool (https://www.genscript.com/tools/real-time-pcr-tagman-primer-design-tool). Default primer settings were used with selection for primers. Efficiencies were determined via standard curves. Four serial 1:10 dilutions were prepared starting at 20 ng, two technical replicates and two controls with nuclease-free water in lieu of template cDNA. Melt curves showed all primer sets had high specificity. Since in most cases primers could not be designed to span exon/exon boundaries (with the exception of *S. flava rpl32*), we confirmed there was no genomic DNA contamination by loading the RNA product in a 1% agarose 1X TBE gel and by conducting RT-PCR of a no RT control and running out products on 1% agarose 1X TBE gels. Primer sequences, efficiencies, and concentrations used are shown in **Table S6.**

RT-qPCR cycling conditions.

RT-qPCR reactions were run on StepOne ™ Real-Time PCR System (ThermoFisher Scientific). Reaction volumes were as follows: 10 µL 2X DyNAmo HS SYBR Green qPCR Kit, 0.15 µl ROX Passive Reference Dye, 0.5 µl of 40 µM forward and reverse primers, and 20 ng cDNA to a total reaction volume of 20 µl. All run cycles included initial 10 minute denaturation at 95° C, 40 cycles of: 95°C for 15 s, 60° C for 1 m, followed by a melt curve ramp from 60° C to 95° C where data was collected every +3°C. Nuclease-free water was used for no template controls.

Data analysis.

Relative quantification was calculated according to the Pfaffl model (*62*) using primer efficiencies described in the supplement. Multiple comparisons were analyzed by two-tailed t-tests and visualized in GraphPad Prism v7.04 (GraphPad Software, San Diego, USA).

Evaluating insect cdtB DNase activity

MUSCLE alignment of cdtB residues across taxa.

DNase and cdtB amino acid residues were found from the following sources: Bovine DNase P00639, *E. coli* Q46669, *C. jejuni* A0A0E1ZJ81, *S. enterica* G5MJJ6, *H. ducreyi* G1UB80, *APSE-2* C4K6T7, *Dr. biarmipes* XP_016950904.1, *Dr. ananassae* XP_014760894.1, *Dr. bipectinata* (1) XP_017099970.1, *Dr. bipectinata* (2) XP_017099943.1, *M. persicae* XP_022163116.1. *S. flava* and *Dr. primaeva* sequences were translated from CDS in MH88465 and MH884659, respectively. Sequences were aligned using MUSCLE (*56*) using a maximum number of 50 iterations and visualized in Geneious (*57*) with a custom blue-scale color scheme based on the Blosum62 scoring matrix (with a threshold of 1). Thus, darker blue colors correspond to higher similarity of a residue in the alignment.

Cloning cdtB.

CdtB oligos from *E. coli, Dr. ananassae, S. flava* and *Ca.* H. defensa were codon-optimized for *E. coli* expression and synthesized by GenScript Codon Optimization Services (deposited under GenBank accession #s MH891796-MH891799). *CdtB* was cloned into the pET His6 TEV vector 2B-T (a gift from Scott Gradia, Addgene plasmid #29666) using sequence and ligation-independent cloning (SLIC) (*63*).

Phobius (*64*) predicted signal peptides in cdtB from *Dr. ananassae* and *E. coli* and a transmembrane domain in *Ca.* H. defensa cdtB. In order to facilitate protein expression and purification, these domains were removed by amplifying GenScript oligos with the following SLIC-compatible primers: *E. coli* F: 5'-TACTTCCAATCCAATgcaGACCTGACCGATTTTCGTGTGG-3'; *E. coli* R: 5'-ACGACGGCTAACACCAACCGGATAGTGATCGCTGCTCATCTGGGTACGACGCGCAC CATAAACAATGCCCGCTTGCAGC-3'; *Dr. ananassae* F: 5'-TACTTCCAATCCAATgcaGACGTTACCGATTACCGTATTACCAC-3'; *Dr. ananassae* R: 5'-TTATCCACTTCCAATgttattaGCCACGCGGCGCC-3'; *Ca.* H. defensa F: 5'-TACTTCCAATCCAATgcaAGCCAAAGCCACAACCACAAC-3'; *Ca.* H. defensa R: 5'-TTATCCACTTCCAATgttattaGTTAAATTTAACCGGCTTGTGGTCG-3'. For *S. flava*, SLIC was performed using the following primers: *S. flava* F: 5'-TACTTCCAATCCAATgcaATGGCGATCATTACCCGTGAGC-3'; *S. flava* R: 5'-TTATCCACTTCCAATgttattaGCCGTTCATCGGCGCC-3'. *CdtB* was cloned following University of California – Berkeley QB3 SLIC protocols (available at: http://qb3.berkeley.edu/macrolab/lic-cloning-protocol/ [accessed March 28, 2018]).

CdtB expression and purification.

Clones were transformed into Rosetta™ 2(DE3)pLysS competent cells (Novagen) following manufacturer protocols. Freshly transformed cells were grown in 2xYT medium at 37°C to an OD_{600} of approximately 0.6, at which point the incubation temperature was lowered to 16°C. After 20 m growth at this temperature, IPTG was added to a final concentration of 0.5 mM. Cells were harvested by centrifugation after overnight growth at 16°C, resuspended in Nickel Buffer A (25 mM HEPES pH 7.5, 400 mM NaCl, 5% glycerol, 20 mM imidazole), then frozen at -80°C. Proteins were purified by Ni affinity chromatography, followed by removal of the His-tag with TEV protease, and size-exclusion chromatography.

After removal of the His-tag from the *E. coli* cdtB, a subtractive Ni affinity step was used to separate the untagged protein from the TEV protease and other contaminant proteins. The untagged protein was concentrated and loaded onto a HiPrep 16/60 Sephacryl S-200 HR size-exclusion column equilibrated in 25 mM HEPES-NaOH pH 7.5, 400 mM NaCl, 10% glycerol. Fractions containing cdtB were pooled and concentrated, assayed by UV absorption, and frozen in aliquots at -80ºC.

The His-tag could not be removed from *Dr. ananassae* cdtB by TEV protease, and so the His-tagged protein was further purified by size-exclusion chromatography on a HiPrep 16/60 Sephacryl S-300 HR column equilibrated in 25 mM Tris-HCl pH 8.0, 200 mM NaCl, 10% glycerol, 2 mM EDTA, 5 mM DTT. Fractions containing cdtB were pooled and concentrated, assayed by UV absorption, and frozen in aliquots at -80ºC.

We failed to purify cdtB from *S. flava* due to its aggregation into inclusion bodies. *Ca.* H. defensa cdtB was expressed at low levels and the final product contained multiple bands. Thus, these proteins were not included in the analysis of DNAse activity.

Mass spectrometry.

Since there were several faint bands on the SDS-PAGE gel in addition to cdtB (**Fig. S5**), we analyzed trypsin-digested protein by LC-MS/MS and determined they were degradation products of cdtB and not from contaminating nucleases. Trypsin-digested *Dr. ananassae* cdtB was submitted to QB3/Chemistry Mass Spectrometry Facility at University of California - Berkeley for LC/MS analysis. 10 µg cdtB was denatured and reduced in 6.27 M urea and 9.8 mM DTT at 55ºC for 20 m. Denatured protein was alkylated by incubation of 19 mM iodoacetamide for 20 m in the dark. The reaction was quenched by addition of 37.3 mM DTT and followed by overnight trypsin digestion following standard protocols for Trypsin-ultra™, Mass Spectrometry Grade (New England Biolabs). Trypsin-digested protein sample was analyzed using a Thermo-Dionex UltiMate3000 RSLCnano liquid chromatography system (LC) that was equipped with a C18 column (length: 250 mm, inner diameter: 0.075 mm, particle size: 3 µm, pore size: 100 Å) and a 1-µL sample loop. The LC was connected in-line with an LTQ-Orbitrap-XL mass spectrometer that was equipped with a nanoelectrospray ionization source and operated in the positive ion mode (Thermo Fisher Scientific, Waltham, MA). Data acquisition and analysis were performed using Xcalibur (version 2.0.7) and Proteome Discoverer (version 1.3, Thermo) software. Peptides from expressed and purified protein were measured by tandem mass spectrometry. The number of measured peptides can be used to roughly gauge the relative amounts of the different proteins in the sample. The abundance of contaminating native *E. coli* protein was negligible compared to that of heterologous *Dr. ananassae* cdtB. We searched UniProt for the contaminant, low-abundance proteins and determined none had known nuclease activity likely to lead to false positives of cdtB nuclease activity.

Nuclease assay.

To determine DNase activity, supercoiled pGEM-7zf+ (Promega) plasmid DNA was incubated with purified cdtB. Reaction volumes were 25 mM HEPES, 5 mM $MgCl_2$, 5 mM $CaCl_2$ (*vis* (*65*)), 500 ng pGEM-7zf(+) incubated with variable amounts of cdtB from *E. coli* and *Dr. ananassae* in 20 µL volume. For negative controls, cdtB storage buffer was used. After 2 hours incubation in a 28ºC water bath, reactions were quenched with the addition of 10 mM EDTA following protocols in (*66*). Samples were loaded onto a 0.8% agarose 1X TBE gel (premixed

with SYBR Safe) and subjected to electrophoresis for 1.5 h at 90V. Images were visualized with AlphaImager ™ Gel Imaging System (Alpha Innotech).

Supplementary Text

Domestication of *cdtB* following horizontal gene transfer from prokaryotes to eukaryotes

The presence of eukaryotic motifs in putative HTGs after transfer from prokaryotes may indicate adaptive optimization to eukaryotic transcription, translation, and cellular function. Here we summarize how we determined if insect-encoded *cdtB* is potentially adapted to eukaryotic machinery.

Transcription and translation initiation elements. Thomas and Chiang 2006 (*67*) provides a comprehensive list of core promoter elements and their consensus sequences identified by transcription initiation factors (TF) TFIID and TFIIB. Other transcription or translation initiation elements we searched for included the Kozak sequence (*68–70*), the GC box (*71*), and the CAAT box (*72*, *73*). Additionally, a Shine-Dalgarno sequence (a ribosomal binding site in bacterial RNA (*74*)), can help assess if a putative HTG is actually due to bacterial contamination.

Transcription termination elements. Transcription termination elements are summarized in Proudfoot 2011 (*75*) and include polyadenylation signals, cleavage site (CA), and upstream and downstream sequence elements (USE and DSE, respectively).

Elements of post-translational processing. Motifs involved in recognition of cargo by accessory proteins of COP and clathrin coated vesicles are described in (*76*) and were searched using the web-based database LOCATE (*77*). Additional motifs included mannosylation sites (*78*, *79*), sulfation sites (*80*), nuclear localization signals (*81*, *82*), and signal peptides, which were predicted using Phobius and SignalP (*64*, *83*).

For the sake of brevity, we here only consider the transcriptional motifs identified bioinformatically in insect *cdtB* nucleotide sequences. This list is not exhaustive and all elements will not necessarily be found in all eukaryotic genes (*84*). We did not conduct experiments to confirm the function of the candidate elements. A visual representation of these eukaryotic motifs is shown in **Figure S4.**

Legend

- Predicted exons are highlighted light blue while predicted introns are yellow. For *Dr. biarmipes* and *Dr. bipectinata*, exons and introns were predicted by (*85*) as part of the modENCODE project. *CdtB* regions in *Dr. ananassae* and *M. persicae* were annotated based on a Gnomon gene prediction set provided by the NCBI (*D. ananassae:* FBrf0227294; *M. persicae*: LOC111028693). For *S. flava, cdtB* was annotated by aligning *S. flava* transcriptome (*86*) to an unpublished *S. flava* genome assembly.
- Coding sequence is light blue and underlined. Thus, 5' or 3' UTRs are light blue, non-underlined nucleotides.
- polyA signals or cleavage sites are highlighted in turquoise.
- Intergenic regions (between the two copies of *cdtB* in *Dr. bipectinata*) are lowercase.

- TATA box motifs are written in orange text, initiator sequence is in purple text, USEs are in grey text and DSEs are in blue text.

Dr. biarmipes

ATAAATAAGGAGAATTTCTTTCTTTTCAGTTTATTATTGAGCATCAAGATGAGAAGAATAAT
TTTGAGCCTAGCGTTTCTGACTCGTGTAATGAGTTTAGTTACCGACTACAGACTAACGACAT
GGAATTCTCGGGGATATAAATTACATCAAGTTTTCGATTCATTCGCTAGTAATTCATCGTTGA
ATATTTTCTTTGTACAAGAAAGTGGAAATTTGGCCGATAAACGTTTAATTTCAATACAACAA
AATTTACCAGTAAGTTCAAGGAATATTTTTAAGTATTATACTATATTTTTTTATTATTTTCATT
TTTTCTAGTTTTATTTGAATGATGGTAGTAATTCTTATCTATGCGGCGCTTCTGATTTTGTGAA
AGTGTACCAATATCAAGATCAAAGGGTTAATTTATATATATATAACTTTTTCCCGCCTCCAAA
TGGTATGACACGTTTCGTCATTATACTGATCTTTAATTAATTATATTAATAATACATTTATATT
TTCTTCTGCAGTTTTATCAACACTTAGTCTGACAATTGTAAGTCGAGTACCAGCCGATAGTAT
AATTTACTTTGCTTCTCTAACGCGGCGACCTGGACGTGCATGCGAGCGAGAGTTTACAAATC
GCCCCGTTGCCGGAATAACTATTGGCAACGATGTTTTTCTCAATATTCATGCTGAACCTACGG
GCATCAGAAACGAAGTTCCGGATCAATTGGATGCCATTCGAAACCATATGCGCACACATGCT
CCGCTTTCATCTTGGTTGCTAGCTGGCGATTTCAACAGAATGCCGTCATCTCTACAATTACAA
CCATATAATGAACAATTTGTCGTCCCGCCCCTCAACACCCACGGCAACAAAATTTTGGACTA
CCTAATTTTAGGATCTAATGACGTGAGAGTTTTTCAACAAATGGTTAGAAACCCCTATAGCG
GAATTGTTGATCTCACTTTAACAGGTTCCGATCACAAGGCTGTACATTTTTCTCTTTGAATCT
CACACAATGCCACTATTTTGCCAATGTTTCAAATAATTGTAAAACTGAACACGACTAACACG
ATTTTTGAATAAACTCATGGTAAATTCAAA

Dr. ananassae

TATAGATATTTATACATGTTCCATGGCCATGTACTCATCATTCACTAAAAATTGTCCGAATCG
CGCAGAACAATGAATAGAGTGCTTTCGTTATTAATTCCAGTTTTACTGAATCAGAATCTCGTT
TCTAGTGATGTTACGGATTATAGAATAACGACTTGGAATTCAGAGGGTTATAAACTAGATAA
AGTTTTTGACTTATTGGATAAAGACAAGTCCTTAAATTTGGTCTTAGTGCAAGAATGTGGAA
ATATTGCAGACAAAAACCCAGGCAGCATTATTAATCCACCTGTACAGGTACATAATTCATAC
GTAAACAGTTTCAAGGAATAGATTTATTCAAAATGTACACTTATCTTTTTTTTCTCTAATCTG
TCGAAAATAGTTTATAATGATTGACGGTGAAAATGAATACGACTCTGCCAATGATGGTAATT
ATGAAATCCGCGAGTATCGAACACGATCCACTCAATTGTTTATATATTATTTTCCGGCACCCA
AAAGTGGTAAGTATAAAATTGTTTTCAAGCGCATTTAAAATGCTGTGTGAAAAACTTGGGCA
CCATGTTCTCAAGCTTACCTTGTTTGACTAGTTGTTGTAACAGTTTTTTTCTAGAAAAGATAC
TAAAAAGTAATTCCTAGATTCGCTTTTTTTTTAATTTTTTAATCTGTTGATGTTTCTTGACTAC
AAATCAATGACATAATTCAGAAACTGGAAAATTGAATCCAGGTAAAGATTAATTTCTTGGTG
GAATTAATTTGCTATGAAATATTTGTTGTAATAAAATATAAAAGACAATAATGCTATATTAC
ATATAAGTTTTAGTTATAAAATGTTTTATCGCACATTTTTATTTCAGTTAATCAGCAATTTGG
ATTGGCTATTGTAACCAAACAACTGGCGTCAGAGATATTATACTTTGCATCTCTTCACAATCA
CCGAGAAATTATTGATCGCAAGGAACGTTCTTTCATTAATCGTCCTATTGTGGGATTGGTTTT
TGGCACTAATGATATTTTTCTTAATTTTCACGCTGAACCCACTAGAAACAACGAAGTTTTACT
TCAACTAAATGCAATTAAAACTTATATGAGCCGCTATAAACCCAATGCTTCCTGGATGCTAG
GCGCTGATTTCAACCGCGAGCCTGGAGATGTGACTTTGGATCCACATCATGAACGATTGATT
CACCCCTCGCAAAATACTCGCCGTAATAGAATAATAGATTACTTTATATATGGTTCTGCAAA
TAGAAATGTTTTCGCGCGAATGGCCAAAAAACCTCATACAACAACCATAAATGCTAAGTTGG
TCTCTGATCATAAGGCAGTAGATTTTAACCCTGCCCCGAGAGGGTAGTTGGATAAGGTTGTC
CAATTCCATCTTTGGCGCCCCCGGGTTAATTATAATTTAAAAAATCAAAGCAACTAGTGACA
GTAACACCTTTCGAATTATAACGTATAGCCGAGTCCATGCATTTTATTTTTCGTCGTTTTTGA
AAGTTTATAAAAGGCAGCCACCTTTCTTTTAATTTTGTTCAGATAAAAGCTAATCTCATTTAA

CATTTTGGGATTTAACTCATTTTACATGCATGCATTTTAATCTCTTATACAATTTATAATACA
ATGATTTATATACAATCACTATATACAATCATTTATATACAATGAATAATAAATGATAAATG
ATTTTTATTTTAAATTAAATTCTTGTGTTT

Dr. bipectinata

AAATATCATCCCAGTACTCATCATTCACTAAACCGTCTCTGATTCGCGCGGAACAATGAACA
CGGTGCTTTCATTAATTTTTGCGGTTCTACTGAATCGGAATCTCATTTCTGGTTTAGTTACGG
ACTATAAAGTAATGACTTGGAACTCAGATGGTTATAAATTGCAGCAAGTTTTGGATTCATTT
TTGAGGAACCCGTCCTTGAGTTTGGCTTTAGTGCAAGAAAGTGGAAATGTTGCTAGGCAAAA
CCCAGGCCAAGTCATTCAACAAAATTTAGAGGTACAAATACTCCTAATAAGAAAATTCAATA
ATAGTATCTTAATTGCCTTTAAGTAATAGTTTCAAATACTTAATCAAAGTTTAATCCTATTTT
TCATTTGAAATAGTTTACTATGGCTGATGGTGAAAATGCATTTCAATGCGCCAATGATGGAT
ATTATGAAGTCCGGCAGTATTTTAATCAAGGAACCCGTTTATATATTTATTTTTCCCCGCAG
CCGAAAATGGTAAATACAATAATCTCTCATATTCTATCGATACGTGCTCCAATCGGATTTCCT
AGTGAAGTGAATTTATACAATCACCCTTCTGAACAAATTTTTGCACGATTTTATTATACACTG
ATTTAATTCAAACTTTAATTTTAGTTCTACAAAAATTGGGATTGACCTTTGTAACCAGACAAC
CGGCAACAGAGATACTATACTTCGTATCTCTGCACAATCACCAAGACTGTTTTGATCGCCGC
GGAAGTCCTTTTATAAATCGTCCTATTCCGGGATTGGTTTTTGGCAACGACATTTTTCTCAAT
TTCCATGCTGAACCCTCTGCAAACAACGAAGTTTTAATTCAACTACGATCCATTAAATCTTTT
ATGAGTGTCTACAAACCCAATGCTTCCTGGATGCTAGGCGGCGATTTTAACCGCGAGCCTAG
TGGAGTACAATCTGCTTTGGACCAAAACCATGAGCGATTAATTCACCCTTCCCAAAGGACTC
GCCGTAATCGCATAATAGATTACTTTATATATGGATCCGCAGATCAAAATGTCTTCGCTCTA
ATGAACAGACCTCATACAGCAACTATTAATGCTAATTTGTTCTCAGACCACTATGCTGTAGA
TTTTAACCCTGCCCCAAGAAGAGTAGTAGGTTAATAAGTAGGTTAATAATAAATATAACTTA
CTCCACACCATACATgaattttaaactcaaacaatttataacaataaatactttgtttattaaaaattaagttctttaaattttgatttttttttcaatcact
aatctgaagtcgttaagtacttatcggccataagaaccctaggtcttgaaagaaatttcttcaacatttaaacttgataatagtgaaccgctcattgagttcacta
tcaagtgggagagtaaacaaaattctgtacaaatctggttttgattgtttgtacgagtgccgataagacaggtctgcagaaaaaaatccctcgaacaatttca
ctttctggggctcgcctgaaaaatatgatcacagtactcatcattcattaaacagtTTGTGATCTGCAGAAACAATGAACACAGT
GATTTCATTATTATTTGCCGTACTACTTAAACAGAATTTCGTTTATAGTTTAGTTACGGACTA
CAATGTTACGACTTGGAATTCACAAGGTTATAGACTACAGCAAGTTTTGGATTTATTTTTAAG
CGACCCGTCGTTAAATTTGGCTTTTGTGCAAGAAAGTGGAAATGTTGCTAGCGAAAACCCAG
GAAATGTTATTCATCAGAATTTAGAGGTACGGGTGGTTCGAACAAAAAACAAAAATTGGGA
ATTCACCTTCTTAAGTGTATTTAATTACACATTTATTGACATAAGAGATGAAATATTATGAAA
TGTGACGTGGGTACTGTGTATATTCAAGCTTTTTACTAGATCGTCGTTTTCTGAAAAGTCATT
TAATTTCCAACGTAATTTTTCATATTCAATCATTTGAAATAGTTTTACATAGCTGATTCTGAA
AATGCATTTGAATGCCGCAATGATGGATATTATGAAGTACGGGAGTATGTAAATCAAGGAA
CTCTATTGTATATATATTTTTTTCCCGCGCCTCAAAATGGTGAGTACGATTATCTGTATTATTT
TACCAATTTAAACTGCTGTACAATCCCAGTCTTCTATAGAAATTGTTGCCGGAATTGATTACA
CTTCACATTTTTATTTCAGTTCGACAAGCACTTGGATTGACTATTGTAAGCAGACATCCGGCG
ACAGGGATACTATACTTTGTATCTCAGCATAATCACCGGGCCATATGTGATTCCCGCGACCG
TGCATTTATTAATCGACCCGTTGTGGGATTGGTTCTTGGCACTGATGACATTTTCCTCAACAT
CCATGCTGAACCCACTCGAACAAGGAACGAAGTACTTACTCAACTAAGAGCCGTTAGATCCC
ATATGAGTATCCACAGACCCTCTGCTTCCTGGATGCTAGCAGGGGACTTCAATCGCTTACCT
CAAGACGTACAACCCACTTTGATTCAAAACCAAGAACGATTGGTCCAGCCTCCACAAGGTAC
TCACGATAACGATGTTCTGGATTACTTTATATATGGGTCCGCAGATCAGACTGTATTCGCGC
GAATGGACAGTCCATCAAATCAAGCAACTATTAACCCAACTCTGACCGGCTCGGATCATCAT
GCTGTTTATTTTTCCAACTATATTACATCAATTCAATACGATCGAAGCAACCTACAACATATT
TTAGATCAAGCATGCACCGTACGAAAAACAAAGTACCACGATTGGACGAACTTTTATGATG
AAGTTTACGATGATAAAGATAGGCTACCAAGCGCAGGTTTCATGCAGTTTCAGGACGGATCT
TACATGCAACCCGTCGTGGTTGAACGAGATACTGATATCAAAAAGAGCTCCAATTGGTACCC

GCTTGATTTGTCATTGTTAATTGGGCTTGAGGCCGCATATGCCTACAGTGTAAGGCGTAAAA
GGAGCCTTCGTGTGTCTTCTACAACTCATCACGGCTTGATTTTCATGAACCATAAAGGCTCTC
CAACCATGACGGTATTTTTAAGAAAAGATAAGTTGCTATGTATGACATGGACCAATAAAGG
ATGGAGCAGTCCGACAGAATGTTATCAACACATTTCCCCTGCACAGTTTTCAGCGTCAAGGA
ATGAACCGGTATTTGACCCTTAAAATACTAAACTTTACCATTTTAATTACATTCCTTTTTCTG
AATACAGACTTTTGATCGTTGTTATAACAAAGCTGAAGAATCAGTTATCAGCGAGAATTTGG
AAACATTGGACAAAAATATCCTCAAATCCCGTAACTTGCTTATATTGCAACGTGAGTTTTCA
AGAACCATGAGTAAGGATGGACTGTTCAACAATGAATACAATACATTTAAAGACTATTTTAA
TGCTAAATGGAATAACACGCTAGTCAGAGATGCTTACGGCATGTCAGGTAATTTACGAATAT
GGATCTTAATTGTACATGTTAATTTTACTATGAATCCTTACAAAAAGTCAAACTTTTTTCTAA
TTTTAGGACAAGAATTCTGGGTAAAGTCCCTCAGACACTGGAATTTTGGATACTACTTCAAA
GATGGATCTTTCTCCGTCGCATTCAGTTCTAAGCATATTCAAAACGGTGACTCCGATAATTGG
TTGTTTATAAACGGAAATGAAATTTCCGACTGGGATCCTTGGCTCAGAAACATCTATGGAAT
GTTTTTCTTCGACAAGTTCGGACGTCCAGTGATATTCTTTGCAGCTCTTAGTGATTACCCCCA
TTGCTGTATCTATAGGAGCCATTGGGTGTACGAATATCGTCCAGATAATAGTTGGGATTGGA
TCGAAAACTTTTCAGAATGGGAACCAAGCTTGTTGTCGAAAAACTCAGGCGCTTTGAGATTT
ATTGTTGATAAAAACCATATTCACAGTTCATA

S. flava

AGATTAACAAAGTTTTGCTGGTTTTTTTTCTATGCTTTGCCGTTGCGGCTGCTCAAAACAATT
ATAGATTTCTAACTTGGAACACCCAAGGACAACGCTGGCCTCAAGTAACCGCCTATTTGAAT
AGATACGATGTGCTATGTATTCAGGAAGCCGGTGCGTTGGGTTTGAGAGGTCTACAGGCTCT
TAACCAGAATAATATAGACTATCGAATTGTGGATGAGGACAACAGAGGAGAAATTGTCACA
ACATCTGGTTTTAATGGTGGCGTTGAGGCTTATACGTTTTCACATGGAAACGTACCGTATTAT
GTCTATTATTATAACCACTTGGTGGAAACTCACGCAGCAAGGGTAGCAAACAATGACCGCA
GAACGCGTAACATGGCTATTATAACCCGTGAGCGGGCAAGCAAGGTGTACATTATTCCAGC
ACATTCAAGGGATCCATATAGAATTGATGTCAATAGACCAACTATTGGAGTGAAGCTATCAG
GATATACGGTCTTCACGGCCCACTCAGATCCAAACAAAAACGAAATCGTCGATACCATTGGA
AAAGTGGCTCGTTTTATGGCCAGCGAAAATCAGTGCCAACAAACGAAATGGATTTTAATGG
GCGATTTCAACGAAGAGCCAGCAGTTGTAAATCGACGATTGCCTCAAGCTTCAAATGGTTGT
GTCATCAGTATAGTGTCTCCCGCAGCTGCCACCAGACAGGCCAGCGGCAAAATTATTGATTA
TGGAGTCTATGGCGGTCCACCTAGTACAGCAGGTTCGCTTCAAGCTACCACTAGAACGGGTG
AGGGCAACAGTGATCATTGGCCAGTTCAAATTATGCCTGCTCCTATGAATGGCTAAGAAAGC
AGTTTGTTTTTGAGAAAAAAAAACATTTTTTAATAAAAAATTGAATTGAAAAAAATGTATA

Interestingly, two long terminal repeats were found flanking *cdtB* in *S. flava* (see **Supplementary Text**), which may indicate retrotransposon insertion events (*87*). Repeats in *S. flava* were identified in Geneious (*57*) function "Find Repeats" using minimum repeat length of 100 bp tolerating 10% mismatches. Two long terminal repeats were found 1,126 bp upstream of the 5' end and 203 bp downstream of the 3' end of *cdtB* in *S. flava;* they are 379 bp long.

Repeat 1 (5' end):

TTTTTGGCTTGCCGTGCCGTGCTTTAATTATAATTTGGCTTCATTGCAGTTTAGCTGTTGAAA
ACAATACGCCAGTGGATGCTGCATTAATCATTCGAAAGCAATTTTTGCGGATTTACATGCGT
TGCAGAAAACGATAATTTTCATCCATCATTTTCAAGTGTTTCCGCCAGGATTATAATACCCAG
ATAGACATTGCATAGCATCCTGCTGGGCTGGAACAATCCTCTTTGTGTGTTTTTCTTCTACGC
TCATTATCCTTCGTTGTCGTCGTCAATTACATTGCAACATCTGAAGCTTTTAGCATTTAGTTG
ATGAGCTCGCCTTTTGGCGACCAACTTATCCTTCCATTCCCCGCTCCCGCCTCGAATGATTTA
TG

Repeat 2 (3' end):
TTTTTGGCTTGCCTTGCCGTGCTTTAATTATAATTTGGCTTCATTGCATTTAAGCTGTTGAAAA
CAATACGCCAGTGGATGCTGCATTAATCATTCGAAAGAAATTTTTGCGGATTTACATGAGTT
GCAGAAAACGATAATTTTCATCCATCATTTTCAATTGTTTCCGCCAGGATTATAATACCCAGA
TAGACATTGCACAGCATCCTGCTGGCCTGGAACAATCCTCTTTGTGTGTGTTTCTTCTACGCT
CATTATCCTTCGTTGTCGTCGTCAATTACATTGCAACATCTGAAGCTTTTAGCATTTAGTTGA
TGAGCTCGCCTTTTGGCGACCAACTTATCCTTCCATTCCCCGCTCCCGCCTCGAATGATTTAT
G

M. persicae
TATATAAGGCCGACTTTTAGCGTACTGTAGGTACTATAGGTTAGGTTAGGTACATTATAATA
TTGCTTATTATTTATCAACATATATATATATAACGTGAATTAAAAAAAATTTAACATTTAAAA
CTATAAGTTAGAAGCTGAAATAATATCTACAGAATGGCGACAATAGTCTTGCTATTATTAAT
TTCTCAGCTTATAAATTATAATTTAATTTCGTGTTTAGTTACTGATCACCAAATAGTAACTTG
GAATTCAAATGGCTACAAACTCCCAAAAGTTTTTGCTTTCTTGGATAAGAATCCACGTATAC
ATTCAGTTTTTGTGCAAGAAAGTGGAAATGTTGAATCTGAATCAAATAATGCAGGAACTCCA
GTACCGAAAACTAATTTACCAAAAGTATAAGGATTTTTAATTTTGATACAATTATTGTATAAT
ATATTTAAATATTTGATATTTTTTAATTTTTTAAAATATTTTAATTTATAGTTTAATGTCTCAA
ATATAATTTTTTTATATTTAATTAAATATAATAATCATATTTTTATAATTATTTTATATTAAAA
AAAATGATTTTTTTAAATAGTTTGTTATTGCTGACGTAGAAGGTGATATCGAATGTACGAATT
ATGCTGACTTTATTAAAGTCAAAACTGTAACGACATTTCAAAACGGGCCGTTTTATATATATT
ATTCACCTGCATCCCCCAATGGTAAGTAACATCAAATAATTTCAATAATACCTACAATTTTCA
TATTATTTTGTACTTCTATTATCTAGTTACTCCCAAAATTAGCATAACTATTGTAAGCAGACA
TTTAGCTAGAGAGATAATACTTTTTCCAAGTCAACACAATCATAAGAGTGTATGTCACGATC
GTGCTTATACTAACCGCCCTATTATTGGATTGGTTTTACAAAATAATCAAAATAATCAAAAA
AATTATGATATTATTTTAAATATTCACGCGGAACCAACTAGAAAACGTAACGAAGTGATAAC
ACAATTGAAAATTATCAGAACTTATATGAATACCATTAGAAAACCTACTTCATGGTTGTTAG
CCGGTGATTTTAATCAATTACCTGAAGACATTATAAGTGAATTAAAATCTCAAGAACAAATA
GTCACACCAAATGATAATACTCATGCTACTCATATTCTCGATTTTTTAATATATGGTTCCCCT
GATGTTCAAATTTTTTCGAAGATGAAAAATTCTCAATATAAAGGAGAAATAGTGGAAAAGG
ATTTGATTTCAGATCATAAAGCCGTGCATTTTTTTAAGTAGTTAGGTAATTTTTTATTAATGTT
TATTATTTTGGTGTTAGCGCATTGTATTCATCAAATATTTGTGAATTTTATTATTAAATTTATT
TTTATTTAACATAATAATTTATTAAGTAGGTACCTATATTTTACTATGGTCTGAATGACTGAA
TTTTGATGATTGATGATTGTCTATAATATAATATGATAGCCATTACACCATATTCAAAATAAT
AAGTACCTATTGTATATTTAAATAATTAAATTATATTTCATAGGTAACATGTATTTTTTAAGT
TATTTATTTAAGAATTTTAAATGTATTTTCACTCACAAAAAAAACAAATATGAATTGTACAAT
TGAGAAAACTCGGGGTCATAGAATAATCATTTTTGATAAAATAAATCGTTTATAATATTTTA
CTGTATACCGACTATTATAATTTGATAATTAATGTATTGTC

Apoptosis inducing protein 56 is fused at the C terminus with cytolethal distending toxin B in *Dr. bipectinata* and other ananassae subgroup species

Apoptosis inducing protein 56 is fused at the C terminus with cytolethal distending toxin B in Dr. bipectinata *and other ananassae subgroup species.* The full cdtB alignment (**Figure S7**) shows a large C-terminal region of the second *Dr. bipectinata* copy (with five exons) does not align to that of any other cdtB sequences except two other species in the ananassae species subgroup, *Dr. pseudoananassae nigrens* and *Dr. ercepeae.* To determine if this region was a bacterial contamination artifact, we amplified and sequenced this region from *Dr. bipectinata* via RT-PCR and Sanger sequencing, which confirmed that this region had introns (leading us to disfavor bacterial contamination) and overlaps the *cdtB* domain (disfavoring errant colocalization of the two genes by assembly error), which corroborated data from (*85*). We extracted this conspicuous region (residues 294-651 from XP_017099943.1) and submitted it as a BLASTP query **(Table S8a),** which showed the region has high homology to another *Ca.* H. defensa protein, hypothetical protein D (*88*). Interestingly in the *Ca.* H. defensa 5AT strain genome (NC_012751.1) the two genes *cdtB* (KF551594.1) and *ORF D* (DQ09613.1:2421-3185) are ca. 255 bp apart. Since we identified relatively few hits in the first BLASTP search and high divergence may have limited the number of identified homologs, we subsequently ran a BLASTP search using the *Ca.* H. defensa hypothetical protein D (WP_015874047.1) as a query (**Table S8b**). Our results show that the second half o*f D. bipectinata* cdtB has homology to the protein apoptosis inducing protein 56 (aip56), a key virulence factor of *Photobacterium damselae piscida*, one of the most important bacterial diseases in mariculture (*89*). Aip56 is secreted by the type II secretion system of *P. damselae piscida* (*90*). In infected fish, aip56 triggers apoptosis of macrophages and neutrophils, which leads to infection-associated necrotic lesion that can devastate a population (*91*).

Aip56 may have been transferred horizontally in other eukaryotic species. Interestingly, copies of *aip56* were also found in the genomes of two other insect species: *Operophtera brumata,* the winter moth, and *Danaus plexippus plexippus,* the monarch butterfly. However, the two *aip56* homologs (OWR44524.1, OWR45007.1) identified in *Da. plexippus* were located on short scaffolds containing only the gene of interest, suggesting these may be bacterial contaminants. While *aip56* homologs found in *O. brumata* were also found on relatively short scaffolds (1.6kb – KOB51764.1; 52 kb – KOB69574.1; 48kb – KOB68847.1, KOB68849.1) the two former identified scaffolds encoded other *bona fide* insect genes. Additionally, the latter scaffold had two *aip56* orthologs arranged in tandem, one of which contained an intron. Here, we only considered aip56 homologs to be insect-encoded if they were found on scaffolds containing other *bona fide* insect genes. Thus, we include *Dr. bipectinata aip56* and at least three *O. brumata aip56* copies. A previous study determined via phylogenetic network analysis that eukaryotic aip56 sequences cluster together, and are closely related to APSE-2 hypothetical protein D (*92*).

Aip56 is an AB-toxin, but only residues in the B domain appear to be conserved between eukaryotic and prokaryotic species. Aip56 is an AB-toxin (*89*). While relatively little is known about aip56, motifs have been discovered that facilitate their cytotoxicity, which manifests as induction of apoptosis in eukaryotic cells (*93*, *94*). Aip56 is composed of two domains linked by a disulphide bridge: the A domain is responsible for catalytic activity and the B domain facilitates cellular entry of the toxin. One of the key components of the A domain is a HEXXH

motif which is typical of zinc metalloproteases and is highly conserved within bacterial species (*95*). The A domain cleaves the transcription factor NF-KB 65, thus interfering with the regulation of inflammatory, anti-apoptotic genes (*95*), affecting bacterial pathogenicity. In an alignment of all aip56 sequences we identified in **Table S8,** an HEXXH motif was found in all bacterial species except in *Ca.* H. defensa, and was absent in all insect-encoded copies (**Figure S8**). In the B domain, unlike the A domain, we could identify several motifs that appeared to be conserved between eukaryotes and prokaryotes (e.g. $FD^{695\text{-}696}$, $GRP^{698\text{-}700}$). While the mechanisms behind B domain cellular entry are less defined, it is known that a deletion of this delivery module domain inhibits binding to target cells and reduces cytotoxicity, making it plausible that these residues are important in cellular uptake (*95*). Given that this domain is vital in cellular entry in diverse hosts (*89*, *91*), their conservation across domains of life may signal a vital role in facilitating clathrin-dependent endocytosis, the mechanism of aip56 and cdtB uptake (*91*, *96*). Both cdtB and aip56 also undergo endosomal maturation in host cells prior to inducing cytotoxic effects (*91*, *97*), suggesting compatibilities in their methods of cytoplasmic delivery that may have facilitated the *cdtB+aip56* fusion in the ancestor to *Dr.* ananassae subgroup species.

Hypotheses on the history, functions and mechanisms of cdtB + aip56 *fusion protein.* The fusion of *cdtB* and *aip56* in *Dr. bipectinata,* along with the proximity of those two genes in the *Ca.* H. defensa genome, strongly suggests that these two genes may have been horizontally transferred between ancestors of *Ca.* H. defensa APSE-2 and eukaryotes, either directly from a phage or via a bacterial intermediate (*98–100*). These two genes, which are encoded in an operon-like fashion in *Ca*. H. defensa, are close enough (within 300 bp) that small mutations could have led to read-through mutations or frameshift mutations and to the two individual proteins being expressed and translated as one larger protein (*101*). CDTs and aip56 are both AB toxins, and in *Dr. bipectinata* + *Dr. pseudoananassae* only the A subunit of the CDT, encoded by *cdtB*, and the B subunit encoded by *aip56* are found in these species. Thus, it is plausible that this ctdB+aip56 fusion was adaptive in some insect backgrounds because these two protein domains can work in concert to affect cellular internalization (aip56 B domain) followed by DNase and apoptogenic activity (cdtB). We speculate that the shared cellular internalization pathways of both cdtB and aip56 (clathrin-mediated endocytosis and endosomal maturation) are in fact synergistic. We hypothesize that the *cdtB+aip56* fusion represents a unique adaptation to the same problem *cdtB* may have evolved in response to. However, it is also clearly not the only viable way of affecting cdtB function, since we could not find *aip56* in *Dr. ananassae, Myzus* spp., or *Dr. primaeva* + *Scaptomyza* spp. genomes.

Analysis of *cdtB* synteny within drosophilid and aphid lineages

In order to assess the evolutionary history of horizontal transfer in *cdtB* within and among insect species, we analyzed synteny of *cdtB* in aphid and drosophilid species (*26*, *102–105*). We downloaded *Scaptomyza, Drosophila* and Macrosiphini aphid scaffolds and compared gene identity up- and down-stream of *cdtb* in SynMap using CoGe (*106*). SynMap revealed clear *cdtB* synteny between species in each of these three clades: *Dr. bipectinata* + *Dr. ananassae, Dr. primaeva* + *S. flava, Di. noxia* + *M. persicae* + *M. cerasi,* and none between *Dr. biarmipes* and any other species.

Additionally, since variability in scaffold size between species could limit syntenic inference, we used a complementary microsyntenic approach and manually identified genes flanking *cdtB* (see **Table S9**) in representative drosophilids, which corroborated the results from CoGe.

Comparison of *cdtB*-containing scaffolds to those in *Dr. melanogaster* indicate that *cdtB* is located on Muller element E (chromosome 4) in *S. flava* and *Dr. primaeva*. In contrast, *cdtB* is located on Muller element B (chromosome 3R) in *Dr. ananassae* and *Dr. bipectinata* as well as in the more distantly related *Dr. biarmipes* (*107*).

Results for the above SynPlot analyses can be regenerated at the following links:

- *Dr. ananassae* to *Dr. bipectinata*: https://genomevolution.org/r/qc7l
- *Dr. ananassae* to *S. flava*: https://genomevolution.org/r/s735
- *Dr. ananassae* to *Dr. biarmipes:* https://genomevolution.org/r/s3ze
- *Dr. bipectinata* to *Dr. biarmipes*: https://genomevolution.org/r/s3zk
- *Dr. biarmipes* to *S. flava*: https://genomevolution.org/r/s737
- *S. flava* to *Dr. primaeva*: https://genomevolution.org/r/135zl
- *M. persicae* to *M. cerasi*: https://genomevolution.org/r/12l4j
- *Di. noxia* to *M. persicae*: https://genomevolution.org/r/139ic

Fig. S1.

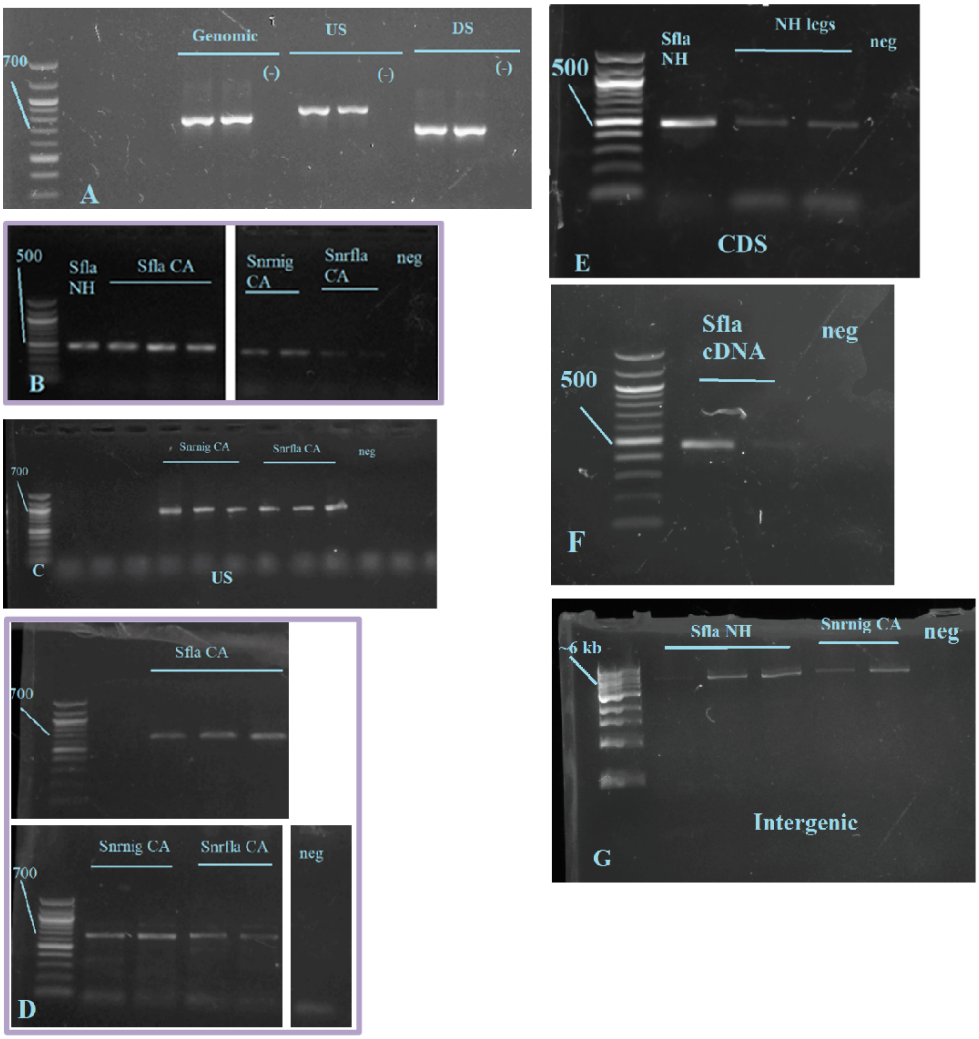
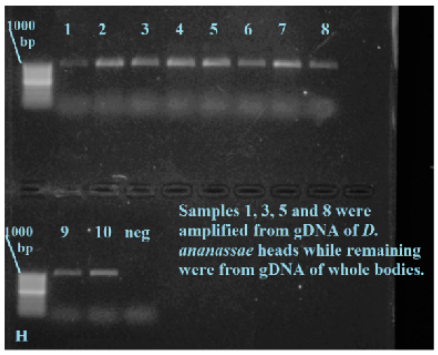
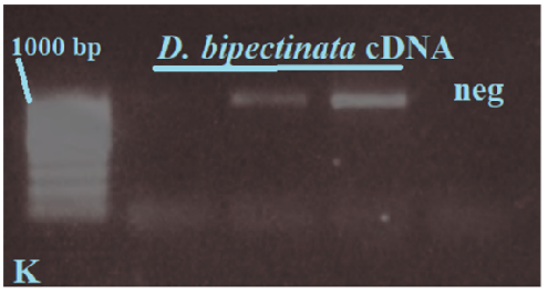
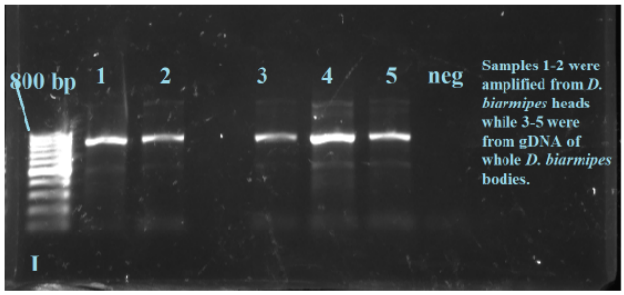


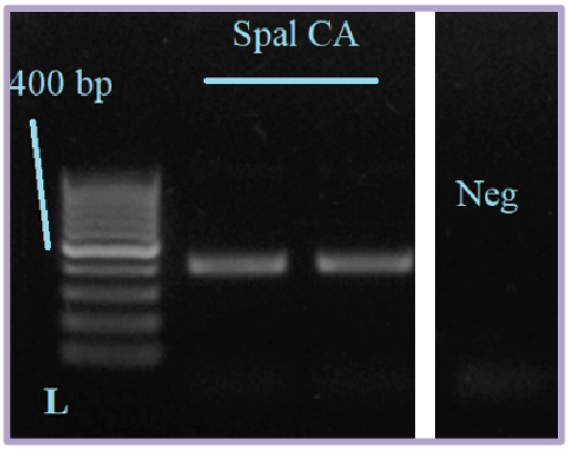
Fig. S1. Gel images of amplicons using primers from **Table S3.** Throughout, *Sfla* NH refers to *S. flava* colonies originally captured from New Hampshire, USA and maintained at UC-Berkeley. *Sfla* CA refers to *S. flava* collected from Berkeley, CA that are 100% identical at *COI* to *S. flava* from NH. *Snrnig* CA refers to *S.* nr. *nigrita* captured from Berkeley, CA. *Snrfla* CA refers to *S.* nr. *flava* collected from Berkeley, CA. *Spal* refers to *S. pallida.* 'Neg' or (-) refers to negative controls with nuclease-free water substituted for template DNA. Unless otherwise specified, *cdtB* was amplified from single whole bodies of the drosophilid species. In all figures O'Gene Ruler 100 bp Ladder (ThermoFisher) is used in the first well (except for item G, in which O'Gene Ruler 1kb Ladder (ThermoFisher) is used). Numbers adjacent to the ladder are approximate size in bp. 1% agarose 1X TBE gels were stained with 0.01% SYBR™ Safe. Amplification from heads and/or legs in E, H and I indicates that *cdtB* is unlikely to be from contamination of gut bacteria. In images B, D and L, different images from the same gel and primer set are stitched together for clarity, these divisions are indicated by white lines. Besides these concatenations, these images are not subject to any nonlinear adjustments, and are indicated by a purple border.

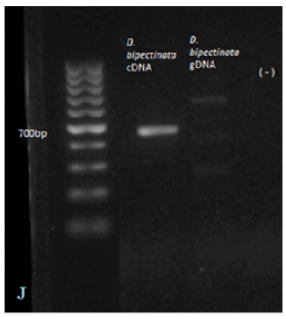
Fig S1 contd.

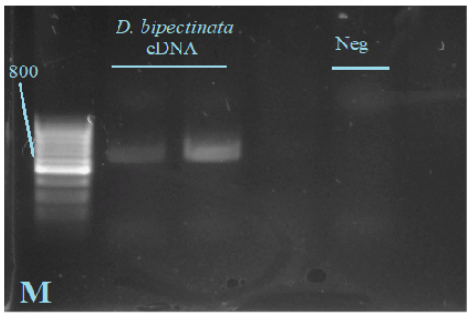












5

Fig. S2.

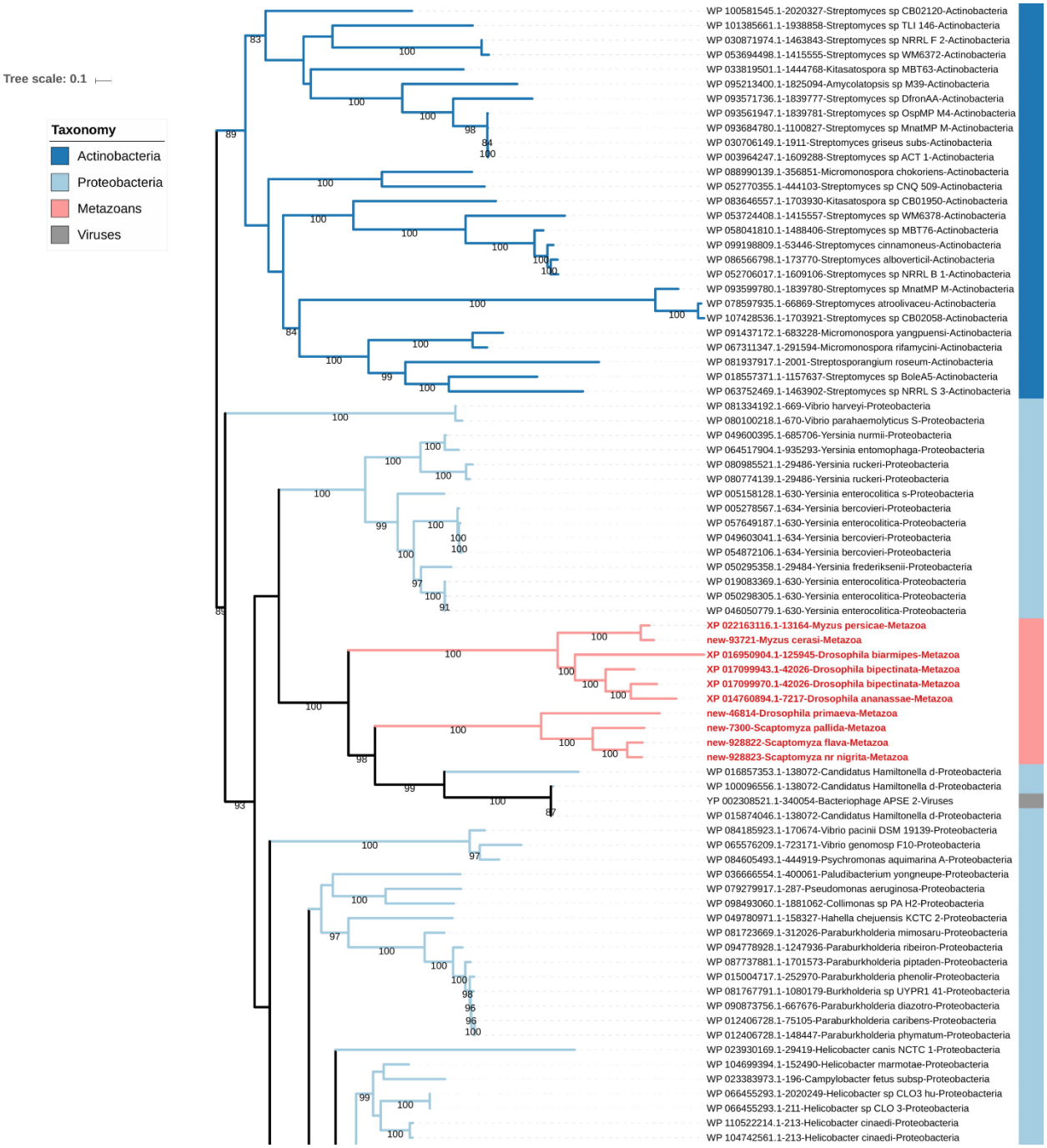
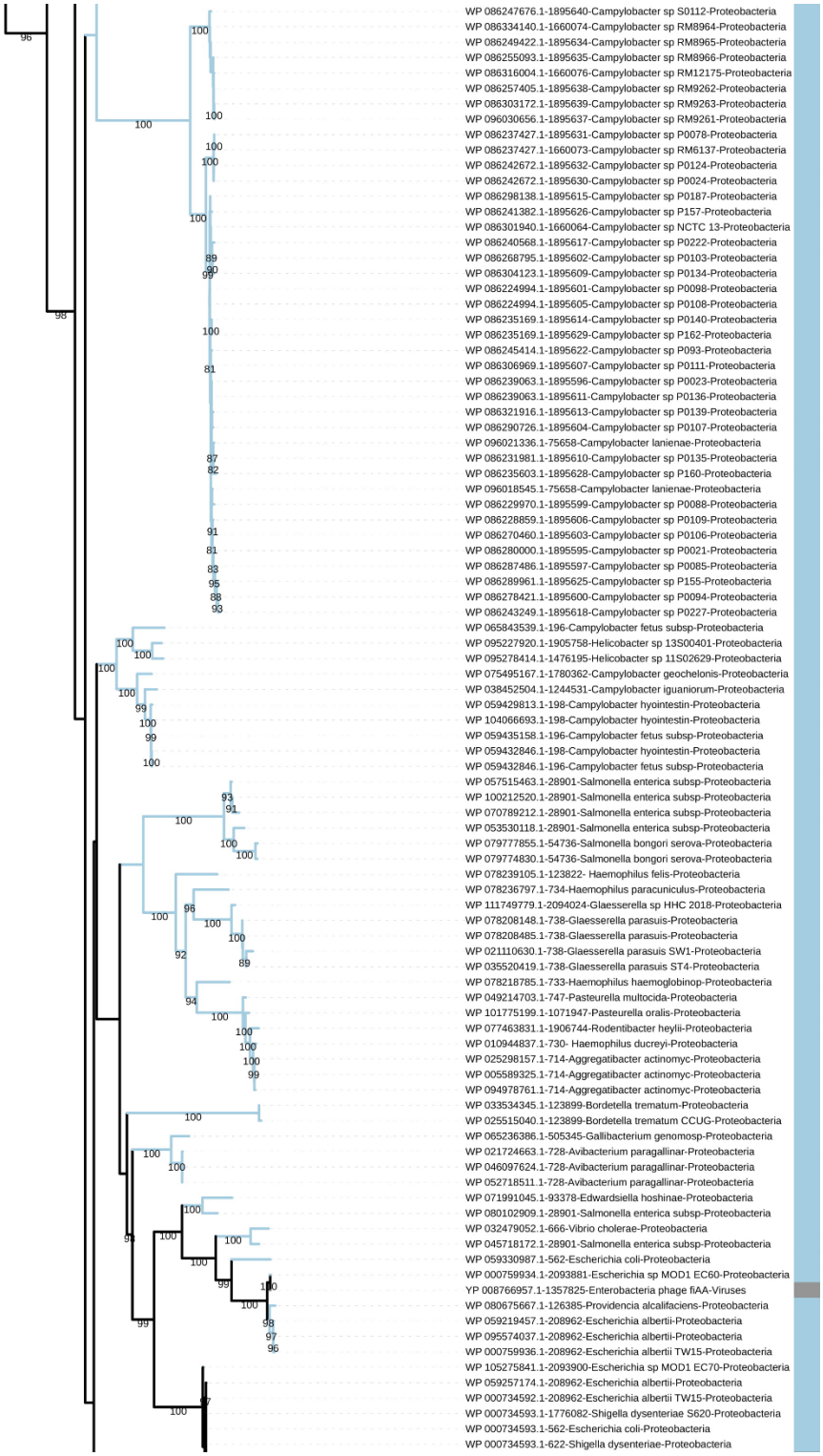


Fig. S2. CdtB protein maximum likelihood phylogeny incorporating sequences from NCBI refseq protein database and those found in this study. Numbers on branches are support values from 1000 bootstrap replicates. The names at the tips indicate the NCBI Protein ID (if not found in this study, in which case it is described as 'new'), species name and domain. For a further description of this phylogeny, please see **Materials and Methods**.

Fig S2 contd.



5

Fig S2 contd.

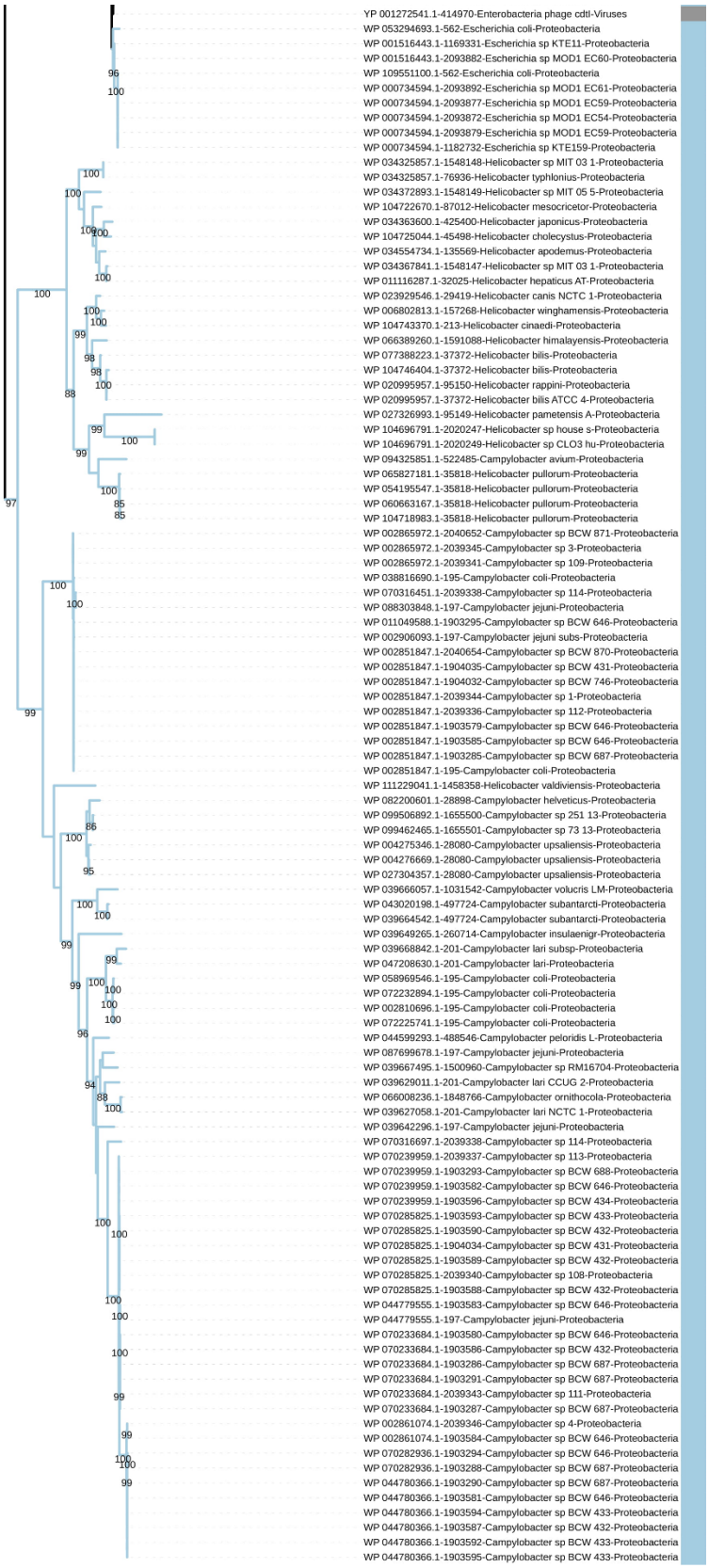


Fig. S3.

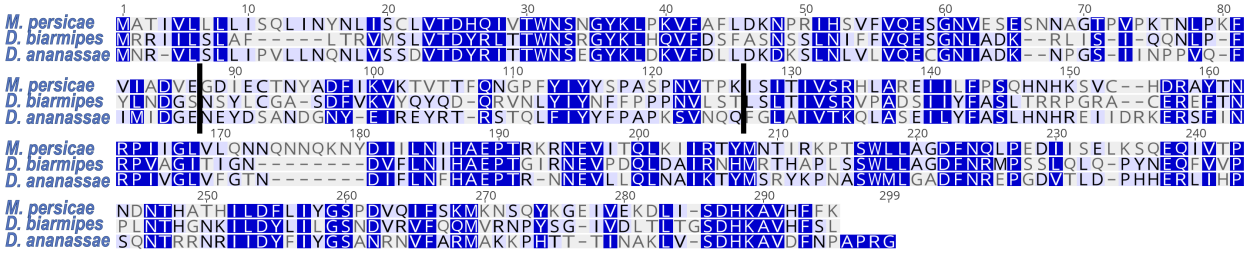


Fig. S3. MUSCLE cdtB amino acid alignment for representative intron-containing copies of cdtB. Splice junctions (indicated by black lines) are conserved in a MUSCLE alignment cdtB copies from *Dr. ananassae, Dr. biarmipes* and *M. persicae.* Blue scale corresponds to similarity under the Blosum62 scoring matrix with a threshold of 1 (where darker shading corresponds to higher similarity). Numbers indicate alignment residues.

Fig. S4.

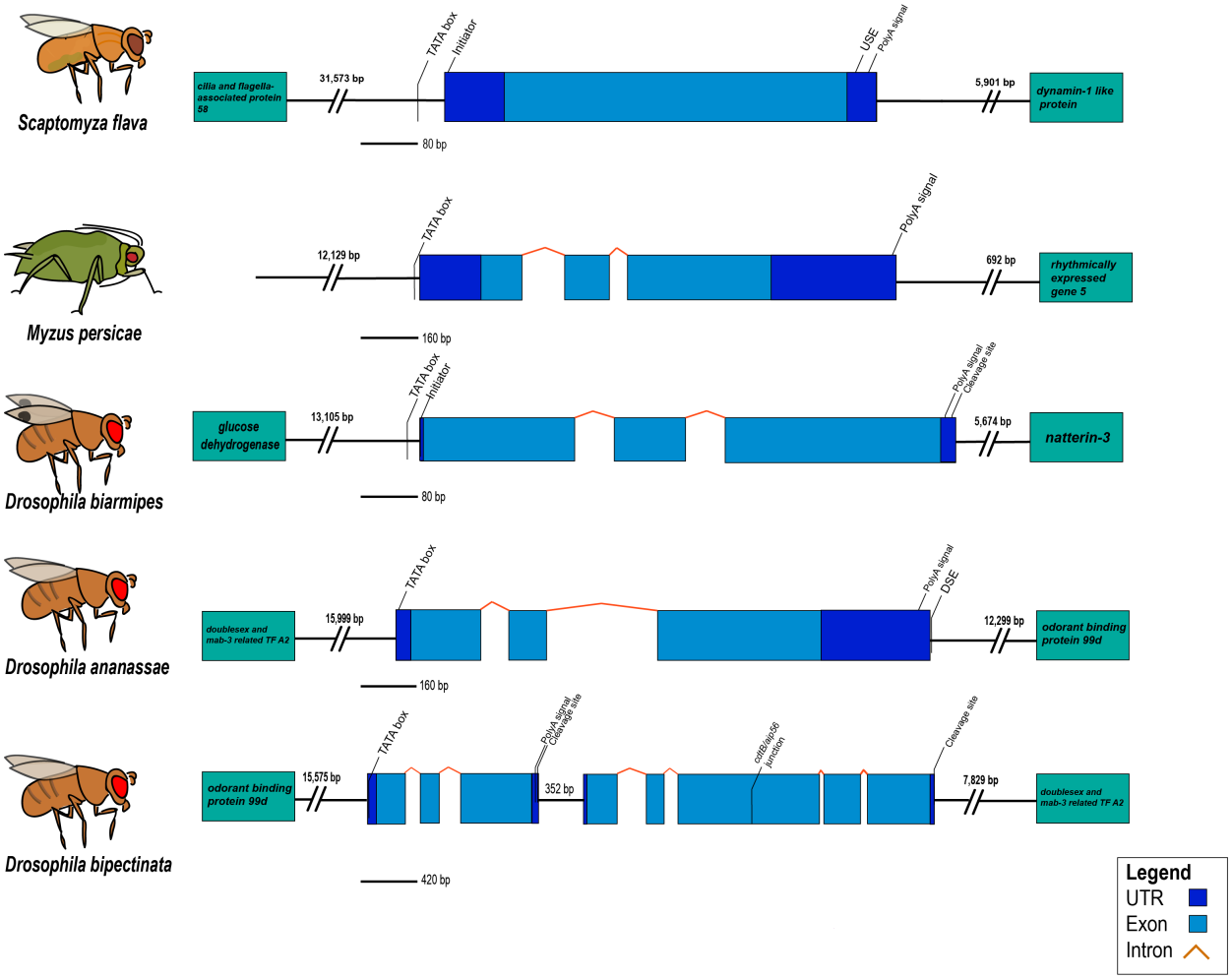


Fig. S4. Gene region and eukaryotic motifs of *cdtB* in representative insect species. UTRs are 5'-3' left to right. Dark blue boxes are UTRs, orange bent lines are introns, and light blue boxes are exons. Boxes to the left and right are nearest flanking genes and brackets indicate distance to nearest gene. Slanted lines with floating text indicate motifs described in the **Supplementary Text**. Gene representations are drawn approximately to scale with calibration legends.

Fig. S5

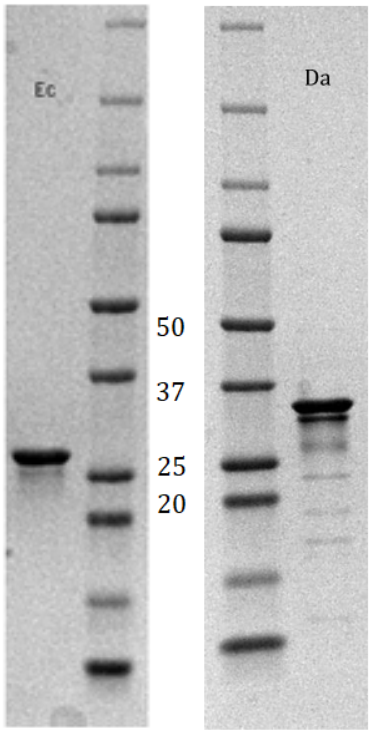


Fig. S5. CdtB from *E. coli* (Ec) and *Dr. ananassae* (Da) washed, separated by 10% SDS-PAGE, and visualized by staining with Coomassie blue. Molecular mass markers are in kDa.

Fig. S6

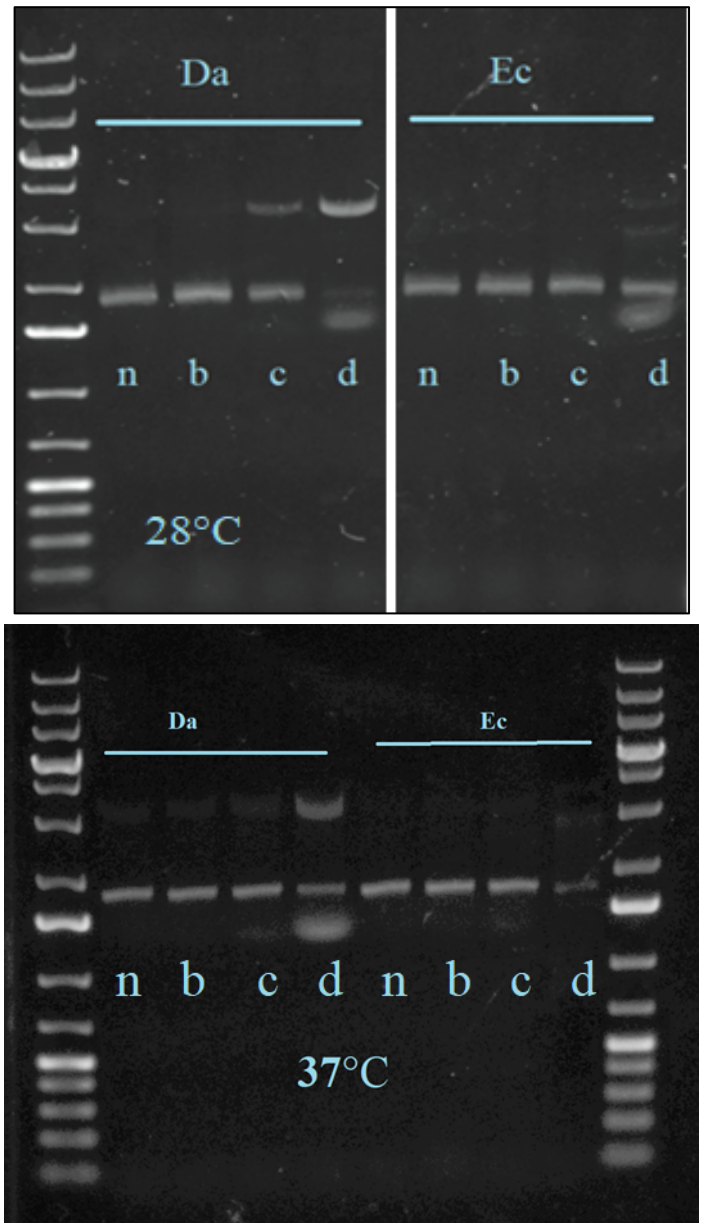


Fig S6. DNase activity assays of *Dr. ananassae* (Da) cdtB and *E. coli* (Ec) cdtB for 2 hours at two different temperatures, 28°C and 37°C. n = buffer control (no cdtB), b = 0.02 µg cdtB, c = 0.2 µg cdtB, d = 2 µg cdtB. 0.8% agarose 1X TBE gels were stained with 0.01% SYBR™ Safe. 5 µL of O'Gene Ruler 1kb Ladder (ThermoFisher) are in the first and last wells in the image. For clarity, the first image is stitched from two parts of the same gel, and this division is indicated by a white vertical line. There were otherwise no vertical manipulations or nonlinear adjustments. For information on incubation conditions refer to the **Materials and Methods.**

Fig. S7.

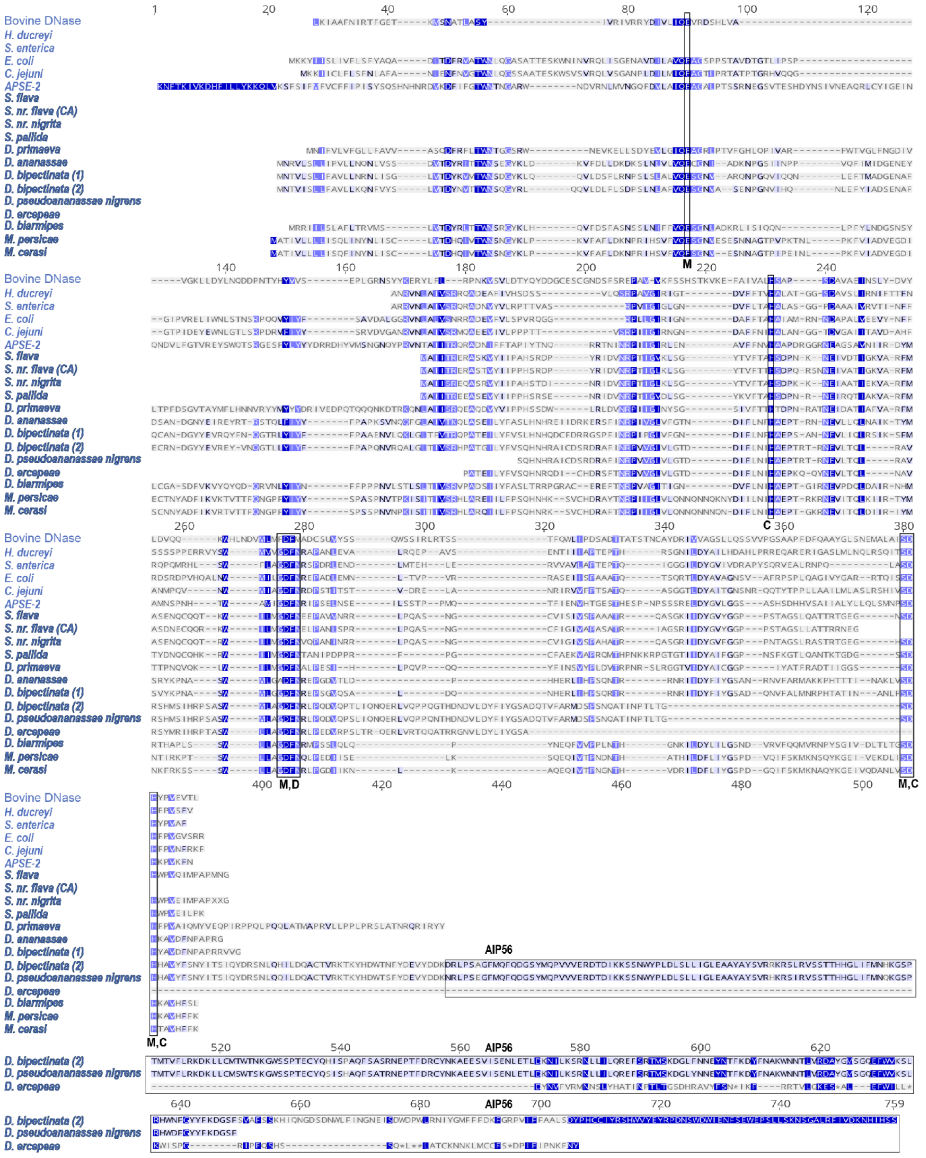


Fig. S7. MUSCLE alignment of cdtB from all identified eukaryotic species with non-pseudogenized *cdtB* copies. Vital DNase residues are highlighted in blue. Blue scale is based on BLOSUM62 similarity scores where darker residues are more similar. Bold species are eukaryotic. DNase and cdtB amino acid residues were from the following sources: Bovine DNase P00639, *E. coli* Q46669, *C. jejuni* A0A0E1ZJ81, *S. enterica* G5MJJ6, *H. ducreyi* G1UB80, *APSE-2* C4K6T7, *Dr. biarmipes* XP_016950904.1, *Dr. ananassae* XP_014760894.1, *Dr. bipectinata* (1) XP_017099970.1, *Dr. bipectinata* (2) XP_017099943.1, *M. persicae* XP_022163116.1. *Scaptomyza* spp. and *Dr. primaeva* were translated from CDS in GenBank sequences MH884655-MH884659. *Dr. pseudoananassae nigrens* and *Dr. ercepeae* sequences were translated from sequences found from their transcriptomes (*11*). *M. cerasi* sequence acquisition is detailed in **Methods.** Residues vital for DNase activity are described in the main text. AIP56 domain is indicated in a black box.

Fig. S8.

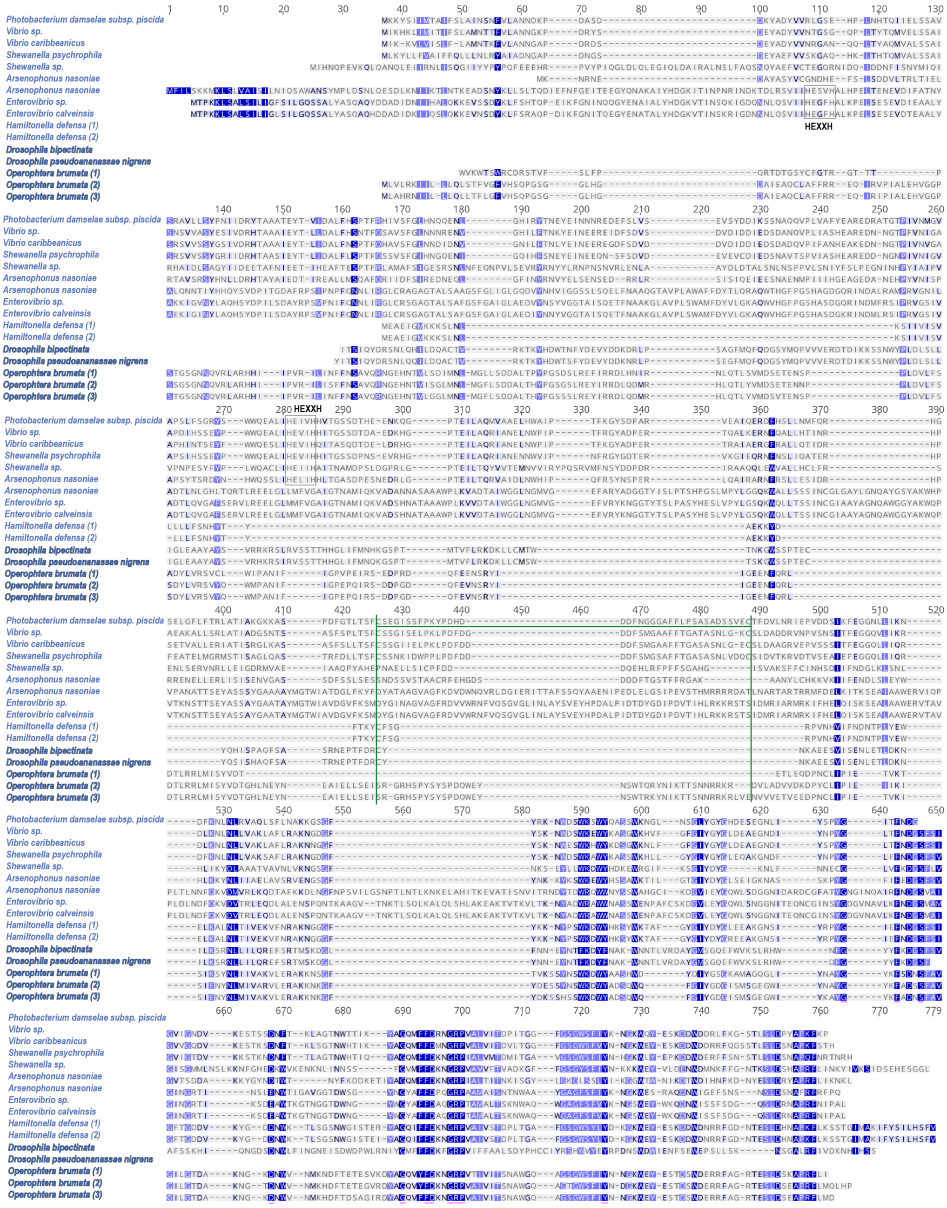


Fig. S8. MUSCLE alignment of aip56 from representative bacterial species and insect species. Bolded species re eukaryotic. Green borders indicate a disulfide bridge (green line) that separates the A (N-terminal) and B (C-terminal) domains. The HEXXH motif in bacterial species except *H. defensa* are boxed and labelled. Sequences were found from the following sources: *P. damselae* subsp. *piscida*: WP_094461508; *Vibrio* sp.: WP_089070319; *Vibrio caribbeanicus:* WP_009600485; *Shewanella psychrophila:* WP_077754668; *Shewanella* sp.: WP_012326868; *Arsenophonus nasoniae*: WP_051297127, WP_051296919; *Enterovibrio* sp.: WP_102315974; *Enterovibrio calviensis*: WP_017014894; *H. defensa* (1, 2): WP_015874047, WP_100096555, respectively; *Dr. bipectinata* 'tail': XP_017099943.1 residues 294-651; *Dr. pseudoananassae nigrens*: extracted and translated from (*11*); *O. brumata* (1, 2, 3): KOB68849, KOB69574, KOB68847, respectively.

Table S1. BLAST searches to the NCBI nr database were used to identify orthologs to *cdtB* from *S. flava*. For **Table S1a**, the TBLASTX search was run using default settings with *S. flava cdtB* as a query, which identified a *cdtB* ortholog in *Dr. bipectinata.* In **Table S1b**, BLASTP search was run using the *Dr. bipectinata* cdtB ortholog (NCBI ID: XP_017099970.1) as a query. Relevant matches are highlighted in purple with added description from authors in parentheses. **Table S1c** presents positive TBLASTN output results from searching *Myzus* genomes and transcriptomes using *M. persicae* cdtB (NCBI ID: XP_022163116) and *Ca.* H. defensa cdtB (UniProt ID: C4K6T7) as queries.

Table S1a.

Accession	Identity	E value	Total Score
gi|669352722|gb|KFD82708.1|	30.769	5.05E-15	84.3
gi|694140751|ref|WP_032479052.1|	30.256	5.38E-15	84.3
gi|506354327|ref|WP_015874046.1|;gi|212499713|ref|YP_002308521.1|;gi|75906026|gb|ABA29376.1|;gi|211731682|gb|ACJ10170.1|;gi|229466506|gb|ACQ68280.1| (*H. defensa)*	30.282	9.23E-14	81.6
gi|548579396|gb|AGX01517.1|	30.282	1.22E-13	81.3
gi|979475913|ref|WP_059330987.1|;gi|975193684|gb|KUS12230.1|	27.014	1.41E-13	80.1
gi|515335288|ref|WP_016857353.1|;gi|211731777|gb|ACJ10107.1|;gi|807061167|emb|CED78233.1|;gi|1239739342|gb|ASX26122.1|	29.897	2.72E-12	76.6
gi|1185521317|gb|OSJ40004.1|	26.804	9.67E-12	75.9
gi|1167649776|ref|WP_080070169.1|	28.571	1.24E-11	74.7
gi|487657737|ref|WP_001750060.1|;gi|459477644|gb|EMG72218.1|;gi|1185609746|gb|OSJ66146.1|;gi|1185636954|gb|OSJ92077.1|	31.731	1.72E-11	74.3
gi|981212232|ref|WP_059435158.1|;gi|962432511|emb|CUU80167.1|	28.571	2.44E-11	74.3
gi|1041901932|ref|WP_065236386.1|;gi|1041834151|gb|OBX05990.1|	29.665	3.32E-11	73.6
gi|1160998720|ref|WP_079825120.1|	32.161	7.37E-11	72.4
gi|981206578|ref|WP_059429813.1|;gi|962416553|emb|CUU68174.1|	25.091	8.06E-11	72.4
gi|974630655|ref|WP_059217482.1|;gi|1221820570|dbj|BBA13803.1|	29.665	1.19E-10	71.6
gi|446682590|ref|WP_000759936.1|;gi|1221813739|dbj|BBA13562.1|	27.5	1.38E-10	71.6
gi|260161890|dbj|BAI43479.1|	27.143	1.72E-10	71.2
gi|981209078|ref|WP_059432098.1|;gi|962410146|emb|CUU85271.1|;gi|962415329|emb|CUU73721.1|;gi|962420525|emb|CUU68881.1|;gi|1139937455|dbj|BAW94583.1|;gi|1139937461|dbj|BAW94587.1|	29.665	1.92E-10	71.2
gi|981209874|ref|WP_059432846.1|;gi|962428340|emb|CUU68669.1|	29.665	1.94E-10	71.2
gi|110591317|pdb|2F1N|A	29.187	2.56E-10	70.9
gi|974632782|ref|WP_059219457.1|;gi|1221813380|dbj|BBA13547.1|;gi|1221813388|dbj|BBA13550.1|;gi|1221984755|dbj|BBA13532.1|;gi|1221986235|dbj|BBA13538.1|	26.978	3.67E-10	70.5
gi|1067639108|dbj|BAV58431.1|	26.978	3.99E-10	70.5
gi|974639171|ref|WP_059225372.1|	26.978	4.22E-10	70.1
gi|1240311560|ref|WP_095574037.1|;gi|949422767|dbj|BAT35603.1|;gi|1221813773|dbj|BBA13574.1|	27.143	4.51E-10	70.1
gi|1221813830|dbj|BBA13594.1|	26.978	4.73E-10	70.1
gi|380503729|dbj|BAL72684.1|	26.978	4.78E-10	70.1

gi|239835498|dbj|BAH78179.1|	26.882	4.91E-10	70.1
gi|924626173|ref|WP_053530118.1|	29.843	5.34E-10	70.1
gi|754738927|ref|WP_042106220.1|	28.061	5.59E-10	70.1
gi|1254454909|ref|WP_097308716.1|;gi|1221820576|dbj|BBA13806.1|;gi|1261077766|gb|PFF93591.1|	29.843	5.82E-10	70.1
gi|57012651|sp|Q46669.1|CDTB_ECOLX;gi|436946|gb|AAA18786.1|	29.843	5.87E-10	69.7
gi|1240313368|ref|WP_095575842.1|;gi|949427335|dbj|BAT39876.1|	26.978	5.95E-10	69.7
gi|974666766|ref|WP_059251177.1|;gi|1221813396|dbj|BBA13553.1|;gi|1221813722|dbj|BBA13556.1|;gi|1221813732|dbj|BBA13559.1|;gi|1221820559|dbj|BBA13797.1|;gi|1221820565|dbj|BBA13800.1|	26.978	6.36E-10	69.7
gi|962418376|emb|CUU85691.1|	26.978	6.73E-10	69.7
gi|981206168|ref|WP_059429417.1|	29.665	6.83E-10	69.3
gi|740667214|ref|WP_038452504.1|;gi|669187494|gb|AII13947.1|;gi|971186009|gb|ALV23685.1|	29.665	7.45E-10	69.3
gi|633260061|dbj|BAO79465.1|	30.144	8.51E-10	69.3
gi|1186813941|ref|WP_085456383.1|;gi|1185798845|gb|OSL33282.1|	30.601	8.74E-10	68.2
gi|974672910|ref|WP_059257174.1|	26.619	1.06E-09	68.9
gi|446657247|ref|WP_000734593.1	29.843	1.08E-09	68.9
gi|446657246|ref|WP_000734592.1|;gi|1185798633|gb|OSL33070.1|	29.843	1.09E-09	68.9
gi|633260073|dbj|BAO79471.1|	29.843	1.09E-09	68.9
gi|1067639104|dbj|BAV58428.1|	29.508	1.09E-09	68.2
gi|633260069|dbj|BAO79469.1|	29.843	1.14E-09	68.9
gi|633260065|dbj|BAO79467.1|	30.601	1.16E-09	67.8
gi|935476351|ref|WP_054412172.1|;gi|921494658|emb|CTV99543.1|;gi|1221813371|dbj|BBA13544.1|;gi|1221813754|dbj|BBA13568.1|	30.601	1.24E-09	67.8
gi|446682591|ref|WP_000759937.1|;gi|52854784|gb|AAU88264.1|;gi|52854791|gb|AAU88269.1|;gi|170121349|gb|EDS90280.1|;gi|569539448|gb|AHE61755.1|;gi|689834862|dbj|GAL53245.1|;gi|1221813790|dbj|BBA13580.1|;gi|1221813799|dbj|BBA13583.1|;gi|1221813809|dbj|BBA13586.1|;gi|1221813818|dbj|BBA13589.1|;gi|1261081802|gb|PFF97598.1|	26.619	1.40E-09	68.6
gi|974650701|ref|WP_059236512.1|;gi|239793078|dbj|BAH72965.1|;gi|953766046|dbj|BAT44166.1|;gi|1221813747|dbj|BBA13565.1|;gi|1221813764|dbj|BBA13571.1|;gi|1221813781|dbj|BBA13577.1|;gi|1221986228|dbj|BBA13535.1|;gi|1221986241|dbj|BBA13541.1|	26.786	1.57E-09	68.6
gi|803451710|gb|KKB02544.1|	26.619	1.58E-09	68.6
gi|919162880|ref|WP_052718511.1|;gi|585571036|gb|AHJ58631.1|;gi|585571044|gb|AHJ58637.1|;gi|585571048|gb|AHJ58640.1|;gi|585571052|gb|AHJ58643.1|;gi|585571056|gb|AHJ58646.1|;gi|585571060|gb|AHJ58649.1|;gi|585571064|gb|AHJ58652.1|;gi|585571068|gb|AHJ58655.1|;gi|585571072|gb|AHJ58658.1|	28.934	2.13E-09	68.2
gi|974671417|ref|WP_059255682.1|	28.934	2.16E-09	68.2
gi|545596271|ref|WP_021724663.1|;gi|523672810|emb|CDG00158.1|;gi|585571040|gb|AHJ58634.1|	26.619	2.21E-09	68.2
gi|803448875|gb|KKA99832.1|	28.934	2.29E-09	68.2
gi|1172292388|ref|WP_080675667.1|;gi|380503743|dbj|BAL72697.1|;gi|573498779|gb|ETS99301.1|;gi|577062157|gb|EUC99176.1|	28.934	2.29E-09	68.2
gi|1196481905|ref|WP_086143604.1|	26.882	2.59E-09	67.8
gi|981211317|ref|WP_059434257.1|;gi|962409756|emb|CUU74482.1|;gi|962423522|emb|CUU82063.1|;gi|1139937448|dbj|BAW94578.1|	27.885	2.59E-09	67
gi|962426331|emb|CUU68708.1|	29.187	2.61E-09	67.8

gi|981207587|ref|WP_059430808.1|	29.187	2.67E-09	67.4
gi|504817615|ref|WP_015004717.1|;gi|407240331|gb|AFT90528.1|	29.187	2.87E-09	67.4
gi|1139937458|dbj|BAW94585.1|	28.899	3.41E-09	67.8
gi|633260051|dbj|BAO79460.1|;gi|633260053|dbj|BAO79461.1|;gi|633260057|dbj|BAO79463.1|;gi|633260059|dbj|BAO79464.1|;gi|633260063|dbj|BAO79466.1|	29.187	3.47E-09	67.4
gi|633260049|dbj|BAO79459.1|;gi|633260055|dbj|BAO79462.1|	30.055	3.81E-09	66.2
gi|633260047|dbj|BAO79458.1|	30.055	4.80E-09	66.2
gi|633260043|dbj|BAO79456.1|	30.055	4.94E-09	66.2
gi|922008764|ref|WP_053294693.1|	30.055	5.34E-09	65.9
gi|1172919822|ref|WP_080985521.1|	29.843	5.81E-09	67
gi|313128905|gb|EFR46522.1|	32.192	6.96E-09	67
gi|892368729|emb|CNB14664.1|	29.952	7.33E-09	66.6
gi|505266253|ref|WP_015453355.1|;gi|396078595|dbj|BAM31971.1|	32.192	7.39E-09	67
gi|1205002136|ref|WP_087737881.1|;gi|1132092518|emb|SIT48164.1|	29.952	7.96E-09	66.6
gi|1204938333|ref|WP_087687232.1|	28.444	8.09E-09	66.6
gi|343381794|gb|AEM17342.1|	28.856	1.06E-08	66.2
gi|504479811|ref|WP_014666913.1|;gi|385146935|dbj|BAM12443.1|	28.638	1.06E-08	66.2
gi|295291384|gb|ADF87419.1|	29.557	1.07E-08	66.2
gi|765033485|ref|WP_044599293.1|;gi|744807343|gb|AJC85289.1|	29.557	1.12E-08	66.2
gi|1036995336|ref|XP_017099970.1| *(D. bipectinata)*	29.851	1.14E-08	65.9

Table S1b.

Accession	Identity	E-value	Total Score
gi|1036995336|ref|XP_017099970.1| (D. *bipectinata*)	100	0	587
gi|964098914|ref|XP_014760894.1|;gi|939214777|gb|KPU72928.1| *(D. ananassae)*	64.894	4.88E-129	376
gi|1036994814|ref|XP_017099943.1| (*D. bipectinata)*	65.714	2.76E-118	362
gi|1036755562|ref|XP_016950904.1| (*D. biarmipes*)	46.014	8.04E-63	208
gi|1229885782|ref|XP_022163116.1| (*Myzus persicae)*	40.69	3.31E-56	191
gi|565846545|ref|WP_023929546.1|;gi|564727729|gb|ETD27707.1|	28.053	2.10E-17	89.7
gi|493855940|ref|WP_006802813.1|;gi|229376234|gb|EEO26325.1|	27.483	4.58E-17	89
gi|538019098|ref|WP_020995957.1|;gi|534480459|gb|EEO24760.2|	26.073	2.76E-15	84
gi|490188257|ref|WP_004086857.1|;gi|476632580|gb|EMZ39102.1|;gi|696178238|gb|KGL22057.1|;gi|696180431|gb|KGL24018.1|	26.733	7.29E-15	82.8

Table S1c.

Database	Query Mpersicae cdtB protein Hit name	Length of hit	e-value	Alignment length	% identity	Query APSE-2 cdtB protein Hit name	Length of hit	e-value	Alignment length	% identity
M. persicae genome	scaffold_179	502501	1e-101	80,42,181	100%,100%,96%	scaffold_179	502501	4e-08	112	34%
M. cerasi genome	Mc971	52308	4e-88	80,42,171	93%,90%,81%	Mc971	52308	2e-08	105	35%
Di. noxia genome	JOTR01000014	1276361	2e-21	75	67%	No hits found	N/A	N/A	N/A	N/A
M. persicae transcriptome	TRINITY_DN79496_c0_g1_i1	1950	0.0	292	100%	TRINITY_DN79496_c0_g1_i1	1950	1e-12	228	29%
	TRINITY_DN79496_c0_g1_i2	1545	2e-109	181	96%	TRINITY_DN79496_c0_g1_i2	1545	2e-08	112	34%
	TRINITY_DN79496_c1_g1_i1	420	1e-88	138	99%	TRINITY_DN79496_c1_g1_i1	420	4e-04	117	30%
	TRINITY_DN146321_c0_g1_i1	270	2e-41	68	100%	N/A				
M. cerasi transcriptome	TRINITY_DN79496_c0_g1_i1	1462	6e-174	293	86%	TRINITY_DN79496_c0_g1_i1	1462	2e-12	221	30%
	TRINITY_DN79496_c0_g2_i2	1530	8e-133	122,171	92%,81%	TRINITY_DN79496_c0_g2_i2	1530	5e-09	105	35%
Di. noxia transcriptome	No significant hits	N/A/	N/A	N/A	N/A	No significant hits	N/A	N/A	N/A	N/A

Table S2.
Details of genome assemblies and scaffolds from representative drosophilids or aphids in which *cdtB* was identified.

	Genome	Genome Scaffold N50	Genome Coverage	Scaffold Size	Exons
Dr. ananassae	GCA_000005115.1	4.59 Mb	8.9X	5.36 Mb (NW_001939298)	3 (GF26441)
Dr. bipectinata	GCA_000233415.2	664 kb	266.3X	732 kb (KB464254.1)	3 (LOC108127428)
Dr. bipectinata	GCA_000233415.2	664 kb	266.3X	732 kb (KB464254.1)	5 (LOC108127405)
Dr. biarmipes	GCF_000233415.1	3.38 Mb	186.9X	3.95 Mb (KB462579.1)	3 (LOC108025143)
S. flava	RKRM00000000.1	105 kb	90X	1.4 Mb	1
Dr. primaeva	Ellie Armstrong, unpublished	272 kb	71.5X	207 kb	no transcriptome data available
M. persicae	GCF_001856785.1	435 kb	200X	502 kb	3 (XM_022307424.1)

Table S3.

Primers used for PCR and RT-PCR. For associated gel images, please refer to **Figure S1;** for PCR conditions, please refer to **Materials and Methods.**

		Sequence 5'-3'				
Name	**Species**	**Forward**	**Reverse**	**Tm (Celsius)**	**Description**	**Gel image**
Sfla-Gen	*S. flava*	ACAACGCTGGCCTCAAGTAA	AGCCATTCATAGGAGCAGGC	55	Genomic amplification. Includes parts of 5'UTR and coding sequence.	A
Sfla-Coding	*S. flava*	ATGGCTATTATAACCCGTGA	CTGCTCCTATGAATGGCTAA	55	Amplifies *CdtB* from start to stop codon in *S. flava*. Used for cDNA amplification.	B, E, F
Sfla-Upstream	*S. flava*	GGTGAGGGCAACAGTGATCA	TGCCTCGCCATGTTACACAT	55	Amplifies region ~600 bp upstream of CdtB coding sequence in S. flava. Includes 5'UTR.	B, E, F
Sfla-Downstream	*S. flava*	AAGCCAGCATCAGACAGCTT	TGTTTGGATCTGAGTGGGCC	55	Amplifies region ~700 bp downstream of CdtB coding sequence in *S. flava*. Includes 3'UTR.	A, D
Scaptomyza-Intergenic	*S. flava*	CCGCATTGACAGTGCCTGCAAAA	AGGTGGACCGCCATAGACTCCAT	64.1	Amplifies ~6kb intergenic segment between *CdtB* and closest flanking gene (*dynamin-like protein 1*). Includes portions of these genes.	G
Dana-Gen	*D. ananassae*	ATCTTTCACTGCCGGTACCG	TCTGACGCCAGTTGTTTGGT	56	Spans ~50 bp upstream of first exon to beginning of third exon.	H
Dbia-Gen	*D. biarmipes*	ACCGACTACAGACTAACGACATG	AGCCTTGTGATCGGAACCTG	57	Spans middle of first exon to near end of third exon.	I
Dbip-Coding1	*D. bipectinata* 1	CTGATTCGCGCGGAACAATG	CTACTCTTCTTGGGGCAGGG	56	Spans the shorter *cdtB* copy. Used for RT-PCR.	J
Dbip-Coding2	*D. bipectinata* 2	CACAAGGTTATAGACTACAG	CTTGGTAGCCTATCTTTATC	46	Spans the longer *cdtB* copy from the first to third exon. Used for RT-PCR.	K
Spal-Coding	*S. pallida*	AGGAGGCCAGCGAGGTATACA	CACTGCTGCCGTCTCCAGTT	53	Amplifies coding sequence of *cdtB* in *S. pallida*.	L
Dbip-Coding2b	*D. bipectinata*	AGCTCCAATTGGTACCCGCT	AGCCAAGGATCCCAGTCGGA	55	Spans the longer *cdtB* copy from the third to fifth exon. Used for RT-PCR.	M

5

Table S4.

PacBio read alignment information for genes up and downstream of *cdtB* in *S. flava*. Gene homology was determined via highest-confidence BLASTX (*10*) hits (default settings) using genes predicted from the *S. flava* genome. 'Proximity to *cdtB'* is negative if 5' of *cdtB* and positive if 3' of *cdtB*. For example, a value of -4 indicates that the gene is four predicted genes upstream (5') of *cdtB.* For more information, please refer to the main text.

Name	**Proximity to *cdtB***	**Length**	**Pairwise % identity**	**Coverage of all bases**		**BLASTX hit**
				Mean	Stdev	
cdtB	0	1139	72.7	10.4	0.8	
maker-scaffold00004-augustus-gene-13.42	1	5120	74.8	5.9	1.4	PREDICTED: dynamin-1-like protein isoform X2 [*Drosophila arizonae*]
maker-scaffold00004-augustus-gene-13.46	2	573	92.4	5	0	PREDICTED: actin-related protein 2/3 complex subunit 5 [*Drosophila eugracilis*]
maker-scaffold00004-augustus-gene-13.43	3	823	87.4	4.5	0.5	PREDICTED: probably NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 12 [*Drosophila busckii*]
augustus_masked-scaffold00004-processed-gene-13.11	4	1049	89	3.3	0.7	PREDICTED: early boundary activity protien 2 [*Drosophila arizonae*]
augustus_masked-scaffold00004-processed-gene-13.6	5	1237	92.1	1.6	0.7	PREDICTED: protein insensitive isoform X2 [*Drosophila arizonae*]
maker-scaffold00004-augustus-gene-13.44	6	1976	81.5	3.9	0.3	PREDICTED: HCLS-1 associated protein X-1 [*Tribolium castaneum*]
maker-scaffold00004-augustus-gene-13.47	7	711	83.6	5.8	0.8	PREDICTED: NADH dehydrogenase [ubiquinone] iron-sulfur protein 6, mitochondrial [*Drosophila ficusphila*]
augustus_masked-scaffold00004-processed-gene-13.8-mRNA-1	-1	1264	76.4	13.9	0.6	PREDICTED: cilia- and flagella-associated protein 58 [*Drosophila arizonae*]
augustus_masked-scaffold00004-processed-gene-13.3-mRNA-1	-2	926	79.9	8.1	0.3	PREDICTED: GTP-binding nuclear protein Ran-like [*Drosophila busckii*]
augustus_masked-scaffold00004-processed-gene-13.0-mRNA-1	-3	1416	72.4	9.9	0.7	PREDICTED: casein kinase I [*Drosophila arizonae*]
maker-scaffold00004-augustus-gene-13.45-mRNA-1	-4	4397	75.1	10	1.8	PREDICTED: odorant receptore 23a-like [*Drosophila busckii*]
augustus_masked-scaffold00004-processed-gene-12.9	-5	2044	83.2	6.4	1.1	PREDICTED: TATA-box binding protein-like protein 1 [*Drosophila bipectinata*]

Table S5.

List of all drosophilid genomes (transcriptomes are included only if *cdtB* was found) screened for evidence of *cdtB*. Genomes or transcriptomes were searched using TBLASTX with default options using *S. flava, D. ananassae* and *Ca.* H. defensa *cdtB* as queries. A liberal E-value of 0.0001 was the cutoff to consider a positive identification of *cdtB*. Species in which *cdtB* was identified are highlighted in purple.

Accession numbers of genes used to construct the drosophilid species phylogeny are included. If no sequences were found on GenBank, the genomes (cited in the column labeled 'Reference') were searched using BLASTN (using the *Dr. melanogaster* gene sequence as a query) and the highest-confidence hit was used; in this case, the corresponding cells are marked 'genome'. If no sequences were found on GenBank or from searching the genome, the cell is highlighted yellow.

	16S	*Adh*	*COI*	*COII*	*nd1*	*nd2*	*Cytb*	*gpdh*	*marf*	Reference
Dr. albomicans	Genome	genome	AB488456.1	genome	genome	genome	genome	genome	genome	Vicoso and Bachtrog 2015
Dr. ananassae	JX896435.1	genome	AB032132.1	KM487042.1	HQ631796.1	BK006336.1:255-1278	EU601719.1	FJ795593.1	genome	Clark et al 2007
Dr. arizonae	AF185068.1:1-771	XM_018003622.1		DQ436072.1	XM_018016669.1	genome		XM_018013325.1	EU341636.1	ASM165402v1
Dr. biarmipes	Genome	DQ363229.1	AY098456.1	AF474094.1	HQ631799	genome	genome	AY098467.1	genome	Chen et al 2014
Dr. bipectinata		AB194421.1	AB032131.1	GQ376042.1	HQ631801.1		genome	DQ073911.1	genome	Chen et al 2014
Dr. busckii	KP730763.1	AB261141.1		KF601930.1	XM_017997480.1	XM_017997564.1		XM_017983641.1	XM_017994714.1	ASM127793v1
Dr. cyrtoloma	HQ170962.1	AY006418.1	HQ170761.1	AY006437.1	genome	HQ170874.1:32-520	genome	AY006457.1	genome	Ellie Armstrong, unpublished
Dr. differens	AY006397.1	AY006416.1	HQ170840.1	AY006435.1	genome	genome	genome	AY006455.1	genome	Ellie Armstrong, unpublished
Dr. elegans	AF164596.1:c1704-1187	DQ363230.1	AB032130.1	AF461307.1	AF164596.1:c1109-171			AB032146.2	genome	Chen et al 2014
Dr. ercepeae		AF459784.1	FJ795576.1	AF461306			transcriptome	FJ795602.1		Signor et al 2013
Dr. erecta	AF164585.1:c1716-1199	X54116.1	AF050744.1	GQ244453.1	AF164585.1:c1119-55	X58914.1:1-825	genome	DQ167751.1	genome	Clark et al 2007
Dr. eugracilis	AF164595.1:c1708-1191	DQ363231.1	AY098461.1	AF474079.1	AF164595.1:c1115-177	XM_017229863.1	AF164595.1:1-86	AY098472.1		Chen et al 2014
Dr. ficusphila	AF164594.1:c1704-1187	DQ363232.1	AY757285.1	AY757273.1	AF164594.1:c1109-171		genome	AB032149.2	genome	Chen et al 2014
Dr. grimshawi	genome	U48714.1	GU597459.1	GU597491.1	BK006341.1:c12623-11678	BK006341.1:233-1258	genome	JN815820.1	genome	Clark et al 2007
Dr. heteroneura	AY006396.1	M63287.1	HQ170843.1	AY006434.1	genome	genome	genome	AY006454.1	genome	Ellie Armstrong, unpublished
Dr. kikkawai	KP730792.1	AB669864.1	AF050746.1	AF461293.1	AF164583.1:c1109-171	XM_017180920.1	genome	HQ631711.1	genome	Chen et al 2014
Dr. limitata		genome		genome	genome	genome	genome	genome	genome	Ellie Armstrong, unpublished
Dr. melanogaster	KP730807.1	X60792.1	KJ767244.1	GQ222021.1	FJ158973.1	GQ229518.1	AM403328.1	NM_001273184.2	genome	Adams et al 2000
Dr. miranda	U07319.1	M60998.1	AF451104.1	M95148.1	U07318.1:c457-1	HQ110578.1	EF216276.1	genome	genome	Zhou and Bachtrog 2012
Dr. mojavensis	EU494341.1:c1130-511	XM_002002894.2	genome	AY437272.1	EU494341.1:c442-1	genome	EU494122.1		genome	Clark et al 2007
Dr. montana	AF508191.1	DQ471665.1	genome	DQ426799.1	EU494352.1:c442-1	DQ471461.1	EU494245.1:1-91	AB019546.1:1723-1766,1849-2026,5014-5220,5287-5659,5725-5878,5947-6039,7241-7253		ASM308661v1

Dr. murphyi	genome	genome	genome					genome		Ellie Armstrong, unpublished
Dr. nasuta	AF387335.1	AB261137.1	AB932738.1	AB932783.1	EU494216.1:c442-115	EU493589.1:1-271	EU494105.1	AB261149.1	KX863731.1	Mohanty and Khanna 2017
Dr. navojoa	EU494342.1:c1129-511	AY156524.1	genome	AY437285.1	XM_018109068.1	from genome	EU494123.1		EU341635.1	ASM165401v1
Dr. novamexicana	AF508183.1	AY165542.1	JF735929.1	JF735934.1		genome	AY646768.1	D50088.1		DnovRS1
Dr. obscura	U07303.1	JF735883.1	JF735919.1	AF081356.1	U07302.1:c457-1	EF216233.1	EF216277.1			Nozawa et al 2016
Dr. ochracea	EU494395.1	genome	genome	EU493797.1	genome	genome		genome	genome	Ellie Armstrong, unpublished
Dr. paucipuncta		genome				genome		genome	genome	Ellie Armstrong, unpublished
Dr. persimilis	U07329.1	AF006564.1	AF451101.1	M95143.1	EU189432.1:c893-147	genome	EF216278.1	XM_002014220.1	genome	Clark et al 2007
Dr. primaeva	HQ170993.1	AY006426.1	HQ170791.1	AY006445.1	genome		genome	AY006464.1	genome	Ellie Armstrong, unpublished
Dr.. prolaticilia	HQ171040.1	genome	HQ170837.1	HQ170734	genome			genome	genome	Ellie Armstrong, unpublished
Dr. pseudoananassae	transcriptome	GQ376034.1	AY757280.1	AY757268.1	HQ631820.1	transcriptome				Signor et al 2013
Dr. pseudoobscura	EU494363.1:c1131-514	X62214.1	AF451087.1	M95150.1	EU494363.1:c442-1	genome	EU494146		genome	Richards et al 2005
Dr. rhopaloa	genome	genome	genome		genome		genome	genome	genome	Drho_2.0
Dr. sechellia	AF164589.1:c1705-1188	X04672.1	KJ425948.1	GQ244458.1	AF164589.1:c1109-171	genome	NC_005780.1:10535-11671	genome	genome	Clark et al 2007
Dr. serrata	AF164581.1:C1710-1193	AB669879.1	AB669749.1	GQ376043.1	AF164581.1:c1109-171	XM_020954565.1	AF164581.1:1-86	HQ631730.1		Allen et al 2017
Dr. simulans	AF164588.1:C1705-1188	X57362.1	KX052973.1	GQ222022.1	AF164588.1:c1109-171	genome	JQ691661.1:10528-11664	AF085163.1	genome	Clark et al 2007
Dr. sproati	genome	genome	genome	JX455050.1	genome	genome	genome	JN815748.1	KR270027.1	Ellie Armstrong, unpublished
Dr. suzukii	KU588141.1:C14068-12739	XM_017082035.1	AB032128.1	LN867083.1	HQ631827.1	XM_017078168.1	KU588141.1:10498-11634	AB032144.2		Chiu et al 2013
Dr. takahashii	AF164592.1:C1713-1196	KX384731.1	KP863258.1	AF474089.1	AF164592.1:c1113-175	XM_017155339.1	AF164592.1:1-86	KR056774.1	genome	Chen et al 2014
Dr. villosipedis	HQ171042.1	genome	HQ170839.1	HQ170747.1	genome			JN815717.1	KR269997.1	Ellie Armstrong, unpublished
Dr. virilis	AF508180.1	DQ471668.1	DQ426807.1	HQ110559.1	EU494353.1:c442-1		AY646772.1	X59076.1	genome	Clark et al 2007
Dr. willistoni	EU494373.1:C1111-495	U95259.1	KT194321.1	HQ110560.1	EU494373.1:c428-1	EU493643.1:1-271	EU494155.1	genome	genome	Clark et al 2007
Dr. yakuba	genome	AY804555.1	X03240.1:1470-3009	X03240.1:3083-3766	X03240.1:c12680-11706		X03240.1:10515-11651	DQ167753.1		Clark et al 2007
Scaptodrosophila lebanonensis	EU494411.1	M97637.1	EU493686.1	HQ110572.1	EU494411.1:c442-1	HQ110598.1	EU494188.1			SlebRS1
Scaptomyza flava	KC609621.1	genome	JX160022.1	HQ170738.1		KC609644.1	genome	genome	JX160036.1	RKRM00000000.1
Scaptomyza nigrita	KC609624.1		JX160025.1	JX160029		KC609647.1			JX160039.1	this study
Scaptomyza pallida	LC061488.1	AB033645.1	KY847492.1	HQ110571.1		HQ110597.1		AB261157.1	JX160037.1	Gloss et al 2018

Table S6.

Primers used for RT-qPCR including concentrations used, size of the amplicon, and efficiencies. For description of reaction and cycling conditions please refer to **Methods**.

Species	Gene	F	R	Primer concentration (uM)	Size (bp)	Efficiency
Dr. ananassae	*rpl32*	AAGCCCAAGGGTATCGACAA	GAACCGTAACCGATGTTGGG	40	77	93
Dr. ananassae	*cdtB*	TCACGCTGAACCCACTAGAA	GAAATCAGCGCCTAGCATCC	60	109	111
S. flava	*rpl32*	CAAGTTGTCGCACAAATGGC	GTGCGCTTGTTGGAACCATA	40	91	98
S. flava	*cdtB*	GCGGTCCACCTAGTACAGC	CAATGATCACTGTTGCCCTCAC	40	69	92

Table S7.

Mass spectrometry-based identification of the components of purified *Dr. ananassae* cdtB. '# peptides' is equivalent to abundance of identified protein in purified *Dr. ananassae* cdtB solution.

Accession	Description	MW [kDa]	calc. pI	# Peptides
	Dr. ananassae cdtB	33.1	8.53	54
C6EIY5	ATP-dependent RNA helicase OS=Escherichia coli (strain B / BL21-DE3) GN=rhlE PE=3 SV=1 - [C6EIY5_ECOBD]	50	10.05	8
C6EGA6	CRP transcriptional dual regulator OS=Escherichia coli (strain B / BL21-DE3) GN=crp PE=4 SV=1 - [C6EGA6_ECOBD]	23.6	8.25	8
C6EC08	23S rRNA m1G745 methyltransferase OS=Escherichia coli (strain B / BL21-DE3) GN=rrmA PE=4 SV=1 - [C6EC08_ECOBD]	30.4	7.5	7
C6EGY4	UPF0227 protein ycfP OS=Escherichia coli (strain B / BL21-DE3) GN=ycfP PE=3 SV=1 - [C6EGY4_ECOBD]	21.2	6.61	6
C6EA14	Pseudouridine synthase OS=Escherichia coli (strain B / BL21-DE3) GN=rsuA PE=3 SV=1 - [C6EA14_ECOBD]	25.8	6.18	5
C6EE23	Regulator of sigma D OS=Escherichia coli (strain B / BL21-DE3) GN=rsd PE=3 SV=1 - [C6EE23_ECOBD]	18.2	6.02	4
C6EJQ1	Ferric uptake regulator OS=Escherichia coli (strain B / BL21-DE3) GN=fur PE=4 SV=1 - [C6EJQ1_ECOBD]	16.8	6.11	3
C6EH01	50S ribosomal protein L13 OS=Escherichia coli (strain B / BL21-DE3) GN=rplM PE=3 SV=1 - [C6EH01_ECOBD]	16	9.91	3
C6EHR2	Ribosomal RNA large subunit methyltransferase G OS=Escherichia coli (strain B / BL21-DE3) GN=ygjO PE=3 SV=1 - [C6EHR2_ECOBD]	42.3	6.8	2
C6EGB5	Peptidyl-prolyl cis-trans isomerase OS=Escherichia coli (strain B / BL21-DE3) GN=slyD PE=4 SV=1 - [C6EGB5_ECOBD]	20.8	5.05	2
C6EGI7	Conserved protein OS=Escherichia coli (strain B / BL21-DE3) GN=yrdA PE=4 SV=1 - [C6EGI7_ECOBD]	20.2	5.53	2
C6EIF0	tRNA (guanine-N(7)-)-methyltransferase OS=Escherichia coli (strain B / BL21-DE3) GN=yggH PE=3 SV=1 - [C6EIF0_ECOBD]	27.3	6.92	1
C6EKI9	Alpha-2-macroglobulin domain protein OS=Escherichia coli (strain B / BL21-DE3) GN=yfhM PE=4 SV=1 - [C6EKI9_ECOBD]	181.4	5.43	1

Table S8.

BLASTP results identifying the *Dr. bipectinata* cdtB + aip56 fusion. Results in eukaryotic species are indicated with an asterisk. **A.** BLASTP results from *Dr. bipectinata* non-alignable cdtB residues (residues 294-651 from XP_017099943.1) suggest homology to *apoptosis inducing protein 56 (aip56)*. **B.** BLASTP results from *Ca.* H. defensa ORF D.

Table S8a.

Description	**E-value**	**Identity**	**Accession #**
PREDICTED: uncharacterized protein LOC108127405 [*Drosophila bipectinata*]*	0	100	XP_017099943.1
hypothetical protein [*Candidatus* Hamiltonella defensa]	1.06E-08	32.075	WP_015874047.1
hypothetical protein [*Candidatus* Hamiltonella defensa]	1.65E-08	32.075	WP_100096555.1
Apoptosis inducing protein [*Operophtera brumata*]*	4.85E-05	28.497	KOB68847.1
Aip56 [*Danaus plexippus plexippus*] *	9.54E-05	29.012	OWR44524.1
hypothetical protein [*Vibrio* sp. 2521-89]	1.81E-04	26.396	WP_089070319.1
hypothetical protein [*Vibrio* sp. 2017V-1085]	3.05E-04	25.888	WP_113602841.1
hypothetical protein [*Vibrio* sp. 2015V-1076]	3.16E-04	25.888	WP_113597563.1
MULTISPECIES: hypothetical protein [*Vibrio*]	3.30E-04	25.888	WP_113604820.1
hypothetical protein [*Vibrio* sp. HI00D65]	3.42E-04	28.571	WP_063524616.1
hypothetical protein [*Vibrio* sp. 2017V-1070]	3.49E-04	25.888	WP_113592639.1
hypothetical protein [*Arsenophonus nasoniae*]	8.26E-04	31.544	WP_051296919.1

Table S8b.

Description	**E-value**	**Identity**	**Accession #**
hypothetical protein [*Candidatus* Hamiltonella defensa]	0	100	WP_015874047.1
hypothetical protein [*Candidatus* Hamiltonella defensa]	1.21E-180	98.819	WP_100096555.1
hypothetical protein [*Vibrio caribbeanicus*]	6.94E-68	61.78	WP_009600485.1
hypothetical protein [*Vibrio* sp.]	2.93E-67	59.259	WP_113602841.1
hypothetical protein [*Vibrio* sp.]	3.41E-67	59.259	WP_113597563.1
hypothetical protein [*Vibrio* sp.]	4.04E-67	59.259	WP_113604820.1
hypothetical protein [*Vibrio* sp.]	5.06E-67	59.259	WP_113592639.1
hypothetical protein [*Vibrio* sp.]	6.06E-67	59.788	WP_089070319.1
hypothetical protein [*Vibrio* sp.]	8.02E-66	60.847	WP_009841419.1
hypothetical protein [*Vibrio* sp.]	1.90E-65	59.474	WP_104037599.1
hypothetical protein [*Vibrio* sp.]	4.24E-64	55.208	WP_021710670.1
hypothetical protein [*Vibrio* sp.]	3.45E-63	54.45	WP_063524616.1
hypothetical protein [*Vibrio* sp.]	1.90E-62	54.45	WP_102424773.1
hypothetical protein [*Vibrio* sp.]	2.32E-62	54.45	WP_105025149.1
hypothetical protein [*Vibrio* sp.]	2.64E-62	54.45	WP_102350207.1
hypothetical protein [*Vibrio* sp.]	2.69E-62	54.45	WP_102413802.1
hypothetical protein [*Vibrio* sp.]	2.69E-62	55.866	WP_039981518.1
hypothetical protein [*Vibrio* sp.]	2.72E-62	54.45	WP_017104811.1
hypothetical protein [*Vibrio* sp.]	2.84E-62	54.45	WP_102459559.1

hypothetical protein [*Vibrio* sp.]	3.06E-62	54.45	WP_102276947.1
hypothetical protein [*Vibrio* sp.]	3.41E-62	54.45	WP_032554400.1
non-LEE encoded type III effector C [*Arsenophonus nasoniae*]	1.12E-61	50.698	CBA76058.1
hypothetical protein [*Arsenophonus nasoniae*]	2.40E-61	56.452	WP_051297188.1
hypothetical protein [*Shewanella psychrophila*]	7.11E-56	58.659	WP_077754668.1
Aip56 [*Photobacterium damselae*]	3.86E-55	52.577	WP_012954632.1
apoptosis inducing protein [*Photobacterium damselae* subsp. *piscicida*]	4.30E-55	52.577	BAF99004.1
hypothetical protein [*Arsenophonus nasoniae*]	1.45E-48	50.811	WP_051296919.1
Aip56 [*Danaus plexippus plexippus*]*	1.27E-44	50	OWR44524.1
Apoptosis inducing protein [*Operophtera brumata*]*	1.02E-43	51.163	KOB68849.1
Apoptosis inducing protein [*Operophtera brumata*]*	4.59E-43	52.381	KOB68847.1
Apoptosis inducing protein [*Operophtera brumata*]*	1.41E-41	50.595	KOB69574.1
Aip56 [*Danaus plexippus plexippus*]*	4.63E-38	47.191	OWR45007.1
hypothetical protein [*Arsenophonus nasoniae*]	3.53E-32	40.201	WP_051297127.1
Apoptosis inducing protein [*Operophtera brumata*]*	8.71E-25	42.949	KOB51764.1
hypothetical protein [*Enterovibrio calviensis*]	2.68E-16	36.216	WP_017014894.1
hypothetical protein [*Enterovibrio calviensis*]	2.79E-16	36.216	WP_017015789.1
MULTISPECIES: hypothetical protein [*Enterovibrio*]	2.87E-16	36.216	WP_017009152.1
hypothetical protein [*Enterovibrio norvegicus*]	6.29E-16	35.676	WP_102315974.1
hypothetical protein [*Enterovibrio norvegicus*]	8.76E-16	35.676	WP_102390145.1
hypothetical protein [*Enterovibrio norvegicus*]	9.18E-16	35.676	WP_102395244.1
hypothetical protein [*Photobacterium damselae*]	5.85E-15	47.778	WP_094461508.1
hypothetical protein [*Enterovibrio norvegicus*]	1.22E-14	34.054	WP_016961832.1
hypothetical protein [*Enterovibrio norvegicus*]	1.28E-14	34.054	WP_017006435.1
MULTISPECIES: hypothetical protein [*Shewanella*]	8.30E-12	32.515	WP_012326868.1
PREDICTED: uncharacterized protein LOC108127405 [*Drosophila bipectinata*]*	1.02E-08	31.447	XP_017099943.1
hypothetical protein VIBC2010_06069 [*Vibrio caribbeanicus* ATCC BAA-2122]	6.92E-06	48.889	EFP97680.1

Table S9.

Microsyntenic analysis of genes immediately up and downstream of *cdtB* in representative drosophilid species. Genes were determined in FlyBase (*Dr. ananassae*), from Chen et. al. 2014 (*85*) (*Dr. bipectinata* and *Dr. biarmipes*), the unpublished *S. flava* genome, or Mathers et. al. 2017 (*41*). Homology of up and downstream genes was predicted via default BLASTX searches using genes of interest as queries. 'Proximity to *cdtB'* is described in **Table S4.**

Synteny between the aphid species is supported by macrosyntenic analysis (see **Supplementary Text**).

	D. ananassae		***D. bipectinata***		***D. biarmipes***	
Proximity to *cdtB*	FlyBase ID	Predicted Homolog	Accession #	Predicted Homolog	Accession #	Predicted Homolog
-5	Dana/GF22876	**Mitotic checkpoint protein BUB3.** Has a dual function in spindle-assembly checkpoint signaling and in promoting the establishment of correct kinetochore-microtubule (K-MT) attachments. Promotes the formation of stable end-on bipolar attachments.	LOC108127480/XP_017100063.1	**Mitotic checkpoint protein BUB3.** Has a dual function in spindle-assembly checkpoint signaling and in promoting the establishment of correct kinetochore-microtubule (K-MT) attachments. Promotes the formation of stable end-on bipolar attachments.	LOC108025083/XP_016950831.1	**tRNA (adenine(58)-N(1))-methyltransferase catalytic subunit TRMT61A.** Belongs to family of transferases transferring one-carbon group methyltransferases.
-4	Dana/GF22875	**Lysoplasmalogenase-like protein TMEM86A.** Enzyme catalyzing the degradation of lysoplasmalogen, which are formed by the hydrolysis of the abundant membrane glycerophospholipids plasmalogens. May control respective level of plasmalogens and lysoplasmalogens in cells and modulate cell membrane properties.	LOC108127421/XP_017099965.1	**Lysoplasmalogenase-like protein TMEM86A.** Enzyme catalyzing the degradation of lysoplasmalogen, which are formed by the hydrolysis of the abundant membrane glycerophospholipids plasmalogens. May control respective level of plasmalogens and lysoplasmalogens in cells and modulate cell membrane properties.	LOC108025084/XP_016950832.1	**Uncharacterized protein.** Has predicted KIAA1430 superfamily domain.
-3	Dana/GF23359	**Head-specific guanylate cyclase.** May have a role in phototransduction. Molecular function: GTP binding, guanylate cyclase activity, heme binding. cGMP biosynthetic process, intracellular signal transduction, positive phototaxis, rhodopsin mediated signaling pathway, visual perception.	LOC108127469/XP_017100047.1	**Head-specific guanylate cyclase.** May have a role in phototransduction. Molecular function: GTP binding, guanylate cyclase activity, heme binding. cGMP biosynthetic process, intracellular signal transduction, positive phototaxis, rhodopsin mediated signaling pathway, visual perception.	LOC108025050/XP_016950796.1	**Neuronal PAS domain-containing protein 4.** Required for contextual memory in the hippocampus.
-2	Dana/Obp99c	**Odorant binding protein.**	LOC10812740/XP_017100048.1	**General odorant-binding protein 99a**	LOC108025437/XP_016951417.1	**Suppressor protein SRP40.** Function not known.

-1	Dana/GF23361	**Doublesex- and mab-3 related transcription factor A2.** Protein features are: DM DNA-binding domain. Molecular function is described by: sequence-specific DNA binding; transcription factor activity; sequence specific DNA binding.	LOC108127418/ XP_017099960.1	**Doublesex- and mab-3 related transcription factor A2.** Protein features are: DM DNA-binding domain. Molecular function is described by: sequence-specific DNA binding; transcription factor activity; sequence specific DNA binding.	LOC108025555/XP_016951578.1	**Uncharacterized protein.** Protein features include: FAD/NAD(P)-binding domain; glucose-methanol-choline oxidoreductase.
0	GF26441	***cdtB***	LOC108127428 / LOC108127405	***cdtB***	LOC108025143	***cdtB***
1	Dana/Obp99d	**Odorant binding protein.**	LOC108127413/XP_017099952.1	**Odorant-binding protein 99d**	LOC108025356/XP_016951305.1	**Uncharacterized protein.**
2	Dana/Obp99b-3	**Odorant binding protein.**	LOC108127412/XP_017099951.1	**General odorant-binding protein 99b-like**	LOC108025355/XP_016951303.1	**Uncharacterized protein.** Contains conserved protein domain DUF3421, which is found in the fish toxin natterin and other uncharacterized proteins.
3	Dana/Obp99b-1	**Odorant binding protein.**	LOC108127419/XP_017099961.1	**General odorant-binding protein 99b-like**	LOC108025354/XP_016951302.1	**HGH1.** Predicted to be involved in ribosome biogenesis.
4	Dana/GF23363	**Calcium-binding protein P isoform X1.**	LOC108127415/XP017099958.1	**Calcium-binding protein P isoform X2**: proteins that participate in calcium cell signaling pathways by binding to Ca++	LOC108025150/XP_016950914.1	**Trypsin-1.** Member of the trypsin family of serine proteases.
5	Dana/GF23364	**Uncharacterized protein.**	LOC108127417/XP_017099959.1	**Uncharacterized protein.**	LOC108025146/XP_016950911.1	**Trypsin alpha.**

	S. flava	
Genes up or downstream of *cdtB*	Annotation ID	Predicted Homolog
-5	augustus_masked-scaffold00004-processed-gene-12.9	**TATA-box binding protein-like protein 1.** Part of a specialized transcription system that mediates the transcription of most ribosomal proteins through 5'-TCT-3' motif.
-4	maker-scaffold00004-augustus-gene-13.45	**Odorant receptor 23a-like.**
-3	augustus_masked-scaffold00004-processed-gene-13.0	**Casein kinase I.** Serine/threonine selective enzymes that function as regulators of signal transduction pathways in most eukaryotes.
-2	augustus_masked-scaffold00004-processed-gene-13.3	**GTP-binding nuclear protein Ran-like.** GTP-binding protein involved in nucleocytoplasmic transport. Required for the import of protein into the nucleus and also for RNA export.
-1	augustus_masked-scaffold00004-processed-gene-13.8	**Cilia- and flagella-associated protein 58.**
0	No annotation ID	***cdtB***
1	maker-scaffold00004-augustus-gene-13.42	**Dynamin-1-like protein.** Functions in mitochondrial and peroxisomal division. Required for normal rate of cytochrome c release and caspase activation during apoptosis. Required for formation of endocytic vesicles.

2	maker-scaffold00004-augustus-gene-13.46	**Actin-related protein 2/3 complex subunit 5.** Component of Arp2/3 complex which is involved in regulation of actin polymerization and together with activating nucleation-promoting factor mediates formation of branched actin networks.
3	maker-scaffold00004-augustus-gene-13.43-mRNA-1	**Probably NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 12.** Accessory subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase.
4	Gloss et al 2018. / augustus_masked-scaffold00004-processed-gene-13.11 / [1387264-1388214]	**Early boundary activity protein 2.** Required for chromatin domain boundary function during early embryogenesis.
5	Gloss et al 2018./ augustus_masked-scaffold00004-processed-gene-13.6-mRNA-1 / [1389880,1391098]	**Protein insensitive isoform X2.** Can act as a transcriptional repressor and corepressor.

Supplementary References

33. S. R. Eddy, A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
34. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
35. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
36. M. Nguyen, A. Ekstrom, X. Li, Y. Yin, HGT-finder: A new tool for horizontal gene transfer finding and application to *Aspergillus* genomes. *Toxins* **7**, 4035–4053 (2015).
37. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289-290 (2004).
38. K. P. Schliep, phangorn: phylogenetic analysis in R. *Bioinformatics*. **27**, 592–593 (2011).
39. I. Letunic, P. Bork, Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
40. E. Nováková, V. Hypsa, J. Klein, R. G. Foottit, C. D. von Dohlen, N. A. Moran, Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the obligate symbiont *Buchnera aphidicola*. *Mol. Phylogenet. Evol.* **68**, 42–54 (2013).
41. T. C. Mathers, Y. Chen, G. Kaithakottil, F. Legeai, S. T. Mugford, P. Baa-Puyoulet, A. Bretaudeau, B. Clavijo, S. Colella, O. Collin, T. Dalmay, Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. *Genome Biol.* **18**, 1–20 (2017).
42. International Aphid Genomics Consortium, Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol.* **8** (2010).
43. S. J. Nicholson, M. L. Nickerson, M. Dean, Y. Song, P. R. Hoyt, H. Rhee, C. Kim, G. J. Puterka, The genome of *Diuraphis noxia*, a global aphid pest of small grains. *BMC Genomics* **16**, 1–16 (2015).
44. D. G. Miller, S. P. Lawson, D. C. Rinker, H. Estby, P. Abbot, The origin and genetic differentiation of the socially parasitic aphid *Tamalia inquilinus*. *Mol. Ecol.* **24**, 5751–5766 (2015).
45. R. P. Duncan, H. Feng, D. M. Nguyen, A. C. C. Wilson, Gene family expansions in aphids maintained by endosymbiotic and nonsymbiotic traits. *Genome Biol. Evol.* **8**, 753–764 (2016).
46. S. S. L. Birnbaum, D. C. Rinker, N. M. Gerardo, P. Abbot, Transcriptional profile and differential fitness in a specialist milkweed insect across host plants varying in toxicity. *Mol. Ecol.* **26**, 6742–6761 (2017).
47. S. P. Lawson, L. T. Sigle, A. L. Lind, A. W. Legan, J. M. Mezzanotte, H. Honegger, P. Abbot, An alternative pathway to eusociality: Exploring the molecular and functional basis of fortress defense. *Evolution* **71**, 1986–1998 (2017).
48. P. H. Degnan, Y. Yu, N. Sisneros, R. A. Wing, N. A. Moran, *Hamiltonella defensa*, genome evolution of protective bacterial endosymbiont from pathogenic ancestors. *Proc. Natl. Acad. Sci.* **106**, 9063–9068 (2009).
49. H. Feng, L. Wang, S. Wuchty, A. C. C. Wilson, microRNA regulation in an ancient obligate endosymbiosis. *Mol. Ecol.* **27**, 1777–1793 (2018).
50. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X.

Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2013).

51. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
52. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30** 1312-1313 (2014).
53. X. Zhou, X. X. Shen, C. T. Hittinger, A. Rokas, Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Mol. Biol. Evol.* **35**, 486–503 (2018).
54. H. Shimodaira, M. Hasegawa, CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247 (2001).
55. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
56. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
57. M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, A. Drummond, Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
58. K. Tamura, M. Nei, Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. **10**, 512–526 (1993).
59. L. J. Revell, phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
60. R. Development Core Team, R: A Language and Environment for Statistical Computing. **2.11.1** (2010), (available at http://www.r-project.org).
61. S. P. Bainbridge, M. Bownes, Staging the metamorphosis of *Drosophila melanogaster*. *J. Embryol. Exp. Morphol.* **66**, 57–80 (1981).
62. M. W. Pfaffl, A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **29**, 2002–2007 (2001).
63. M. Z. Li, S. J. Elledge, SLIC: A Method for Sequence- and Ligation-independent cloning. *Methods Mol. Biol.* **852**, 51–59 (2012).
64. L. Käll, A. Krogh, E. L. L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).
65. C. Q. Pan, R. A. Lazarus, Hyperactivity of Human DNase I Variants. *Biochemistry*. **273**, 11701–11708 (1998).
66. D. Obradović, R. Gaspersic, S. Caserman, A. Leonardi, M. Jamnik, Z. Podlesek, K. Seme, G. Anderluh, I. Krizaj, P. Macek, M. Butala, A cytolethal distending toxin variant from *Aggregatibacter actinomycetemcomitans* with an aberrant CdtB that lacks the conserved catalytic histidine 160. *PLoS One*. **11**, 1–16 (2016).
67. M. C. Thomas, C.-M. Chiang, The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41**, 105–178 (2006).
68. M. Kozak, An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15**, 8125–8148 (1987).

69. M. Kozak, Pushing the limits of the scanning mechanism for initiation of translation. *Gene* **299**, 1–34 (2002).
70. D. R. Cavener, Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.* **15**, 1353–1361 (1987).
71. M. C. Blake, R. C. Jambou, A. G. Swick, J. W. Kahn, J. C. Azizkhan, Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. *Mol. Cell. Biol.* **10**, 6632–6641 (1990).
72. B. J. Graves, P. F. Johnson, S. L. McKnight, Homologous recognition of a promoter domain common to the MSV LTR and the HSV tk gene. *Cell* **44**, 565–576 (1986).
73. M. Raymondjean, S. Cereghini, M. Yaniv, Several distinct "CCAAT" box binding proteins coexist in eukaryotic cells. *Proc. Natl. Acad. Sci.* **85**, 757–61 (1988).
74. J. Shine, L. Dalgarno, The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci.* **71**, 1342–1346 (1974).
75. N. J. Proudfoot, Ending the message: poly(A) signals then and now. *Genes Dev.* **25**, 1770–82 (2011).
76. H. T. McMahon, I. G. Mills, COP and clathrin-coated vesicle budding: different pathways, common approaches. *Curr. Opin. Cell Biol.* **16**, 379–391 (2004).
77. J. Sprenger, J. L. Fink, S. Karunaratne, K. Hnason, N. A. Hamilton, R. D. Teasdale, LOCATE: A mammalian protein subcellular localization database. *Nucleic Acids Res.* **36**, 230–233 (2008).
78. J. Perez-Vilar, S. H. Randell, R. C. Boucher, C-mannosylation of MUC5AC and MUC5B Cys subdomains. *Glycobiology*. **14**, 325–337 (2004).
79. J. S. Chauhan, A. H. Bhat, G. P. S. Raghava, A. Rao, GlycoPP: A webserver for prediction of N- and O-glycosites in prokaryotic protein sequences. *PLoS One*. **7**, e40155 (2012).
80. J. M. Capasso, C. B. Hirschberg, Mechanisms of glycosylation and sulfation in the Golgi apparatus: evidence for nucleotide sugar/nucleoside monophosphate and nucleotide sulfate/nucleoside monophosphate antiports in the Golgi apparatus membrane. *Proc. Natl. Acad. Sci.* **81**, 7051–7055 (1984).
81. S. Kosugi, M. Hasebe, N. Matsumura, H. Takashima, E Miyamoto-Sato, M. Tomita, H. Yanagawa, Six classes of nuclear localization signals specific to different binding grooves of importin α. *J. Biol. Chem.* **284**, 478–485 (2009).
82. D. Chelsky, R. Ralph, G. Jonak, Sequence requirements for synthetic peptide-mediated translocation to the nucleus. *Mol Cell Biol*. **9**, 2487–2492 (1989).
83. J. D. Bendtsen, H. Nielsen, G. Von Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
84. A. K. Kutach, J. T. Kadonaga, The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol. Cell. Biol.* **20**, 4754–4764 (2000).
85. Z. Chen, D. Sturgill, J. Qu, H. Jiang, Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. *Genome Res.* **24**, 1209–1223 (2014).
86. N. K. Whiteman, A. D. Gloss, T. B. Sackton, S. C. Groen, P. T. Humphrey, R. T. Lapoint, I. E. Sonderby, B. A. Halkier, C. Kocks, F.M. Ausubel, N. E. Pierce. Genes involved in the evolution of herbivory by a leaf-mining, Drosophilid fly. *Genome Biol. Evol.* **4**, 900–16 (2012).

87. E. Lerat, Identifying repeats and transposable elements in sequenced genomes : how to find your way through the dense forest of programs. *Heredity* **104**, 520–533 (2010).
88. P. Degnan, N. Moran, Evolutionary genetics of a defensive facultative symbiont of insects : exchange of toxin-encoding bacteriophage. *Mol. Ecol.* 916–929 (2008).
89. M. T. Silva, N. M. S. Santos, A. Vale, AIP56: A Novel Bacterial Apoptogenic Toxin. *Toxins* **2**, 905–918 (2010).
90. A. do Vale, C. Pereira, C. R. Osorio, N. M. S. dos Santos, The apoptogenic toxin aip56 is secreted by the type II secretion system of *Photobacterium damselae* subsp. *piscicida*. *Toxins* **9** (2017).
91. L. Pereira, R. D. Pinto, D. S. Silva, A. R. Moreira, C. Beitzinger, P. Oliveira, P. Sampaio, R. Benz, J. E. Azevedo, N. dos Santos, A. do Vale. Intracellular trafficking of AIP56, an NF-KB cleaving toxin from *Photobacterium damselae piscicida*. *Infect. Immun.* (2014).
92. A. Stolle, S. Norkowski, B. Korner, J. Schmitz, L. Luken, M. Frankenberg, C. Ruter, M. A. Schmidt, T3SS-Independent Uptake of the Recombinant NleC Effector of Enteropathogenic *Escherichia coli* Leads to NF- κ B p65 Cleavage. *Front. Cell. Infect. Microbiol.* **7**, 1–23 (2017).
93. M. Reis, A. do Vale, C. Pinto, D. S. Nascimento, C. Costa-Ramos, D. S. P. Silva, M. T. Silva, N. M. S. dos Santos, First molecular cloning and characterization of caspase-9 gene in fish and its involvement in a gram negative septicaemia. *Mol. Immunol.* **44**, 1754–1764 (2007).
94. A. do Vale, M. T. Silva, N. M. S. dos Santos, D. S. Nascimento, P. Reis-Rodrigues, C. Costa-Ramos, A. E. Ellis, J. E. Azevedo, AIP56, a novel plasmid-encoded virulence factor of *Photobacterium damselae* subsp. *piscicida* with apoptogenic activity against sea bass macrophages and neutrophils. *Mol. Microbiol.* **58**, 1025–1038 (2005).
95. D. S. Silva, L. M. G. Pereira, A. R. Moreira, F. Ferreira-da-Silva, R. M. Brito, T. Q. Faria, I. Zornetta, C. Montecucco, P. Oliveira, J. E. Azevedo, P. J. B. Pereira, S. Macedo-Ribeiro, A. do Vale, N. M. S. dos Santos, The apoptogenic toxin AIP56 is a metalloprotease A-B toxin that cleaves NF-KB P65, Is a Metalloprotease A-B Toxin that Cleaves NF-κb P65. *PLoS Pathog.* **9** (2013).
96. L. Guerra, K. Teter, B. N. Lilley, B. Stenerlow, R. K. Holmes, H. L. Ploegh, K. Sandvig, M. Thelestam, T. Frisan, Cellular internalization of cytolethal distending toxin: a new end to a known pathway. *Cell. Microbiol.* **7**, 921–934 (2005).
97. X. Cortes-bratti, T. Frisan, M. Thelestam, The cytolethal distending toxins induce DNA damage and cell cycle arrest. *Toxicon* **39**, 1729–1736 (2001).
98. J. R. Zaneveld, D. R. Nemergut, R. Knight, Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. *Microbiology* **154**, 1–15 (2008).
99. P. J. Keeling, J. D. Palmer, Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* **9**, 605–618 (2008).
100. C. Canchaya, G. Fournous, S. Chibani-Chennoufi, M. L. Dillmann, H. Brüssow, Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* **6**, 417–424 (2003).
101. A. J. Michael, Evolution of biosynthetic diversity. *Biochem. J.* **474**, 2277–2299 (2017).
102. B. Lacroix, V. Citovsky, Transfer of DNA from bacteria to eukaryotes. *MBio*. **7**, 1–9 (2016).
103. T. Rolland, C. Neuvéglise, C. Sacerdot, B. Dujon, Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. *PLoS One*. **4** (2009).

104. D. Wheeler, A. J. Redding, J. H. Werren, Characterization of an Ancient Lepidopteran Lateral Gene Transfer. *PLoS One*. **8** (2013).
105. D. A. Fitzpatrick, M. E. Logue, G. Butler, Evidence of recent interkingdom horizontal gene transfer between bacteria and *Candida parapsilosis*. *BMC Evol. Biol.* **8**, 1–15 (2008).
106. E. Lyons, B. Pedersen, J. Kane, M. Freeling, The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. *Trop. Plant Biol.* **1**, 181–190 (2008).
107. *Drosophila* 12 Genomes Consortium, Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203–18 (2007).