Title: Horizontal transfer of microbial toxin genes to gall midge genomes

Authors and affiliations: Kirsten I. Verster*, Rebecca L. Tarnopol*, Saron M. Akalu, and Noah K. Whiteman†
Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720
*equal contribution

†Corresponding author: Dr. Noah K. Whiteman, Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720
Telephone: (510)-859-7749
Email: whiteman@berkeley.edu

Abstract:

A growing body of evidence points to a role for horizontal gene transfer (HGT) in the evolution of animal novelties. Previously, we discovered the horizontal transfer of the gene encoding the eukaryotic genotoxin cytolethal distending toxin B (CdtB) from the *Acyrthosiphon pisum* Secondary Endosymbiont (APSE) bacteriophage to drosophilid and aphid genomes. Here, we report that *cdtB* is also found in the nuclear genome of the gall-forming ‘swede midge’ *Contarinia nasturtii* (Diptera: Cecidomyiidae). We subsequently searched genome sequences of all available cecidomyiid species for evidence of microbe-to-insect HGT events. We found evidence of pervasive transfer of APSE-like toxin genes to cecidomyiid nuclear genomes. Many of the toxins encoded by these horizontally transferred genes target eukaryotic cells, rather than prokaryotes. In insects, catalytic residues important for toxin function are conserved. Phylogenetic analyses of HGT candidates indicated APSE phages were often not the ancestral donor of the toxin gene to cecidomyiid genomes, suggesting a broader pool of microbial donor lineages. We used a phylogenetic signal statistic to test a transfer-by-proximity hypothesis for HGT, which showed, that prokaryotic-to-insect HGT was more likely to occur between taxa in common environments. Our study highlights the horizontal transfer of genes encoding a new functional class of proteins in insects, toxins that target eukaryotic cells, which is potentially important in mediating interactions with eukaryotic pathogens and parasites.

Key Words: horizontal gene transfer; Diptera; toxins; cdtB; shiga toxin; lysozyme

Significance Statement: The diversity of genes encoded by phages infecting bacterial symbionts of eukaryotes represents an enormous, relatively unexplored pool of new eukaryotic genes through horizontal gene transfer (HGT). In this study, we discovered pervasive HGT of toxin genes encoded by *Acyrthosiphon pisum* secondary endosymbiont (APSE) bacteriophages and other microbes to the nuclear genomes of gall midges (Diptera: Cecidomyiidae). We found five toxin genes were transferred horizontally from phage, bacteria, or fungi into genomes of several cecidomyiid species. These genes were *aip56, cdtB, lysozyme, rhs*, and *sltxB*. Most of the toxins encoded by these genes antagonize eukaryotic cells, and we posit that they may play a protective role in the insect immune system.

Main Text:

There is growing evidence that horizontal gene transfer (HGT) from microbes to animals has played an important role in animal evolution (Husnik & McCutcheon 2018). Toxin-encoding genes have been horizontally transferred into arthropod genomes and even integrated into their immune systems (Di Lelio et al. 2019; Hayes et al. 2020; Li et al. 2021). However, many of these events involve transfer of genes encoding toxins that target prokaryotes, and few characterized HGTs in animals are of genes encoding toxins that target eukaryotes.

We previously discovered HGT of a eukaryote-targeting toxin gene, *cytolethal distending toxin B* (*cdtB*), into the nuclear genomes of four insect lineages within two orders, Diptera and Hemiptera (Verster et al. 2019). The closest relatives of these insect *cdtB* copies were copies isolated from *Acyrthosiphon pisum* secondary endosymbiont (APSE) bacteriophage (Verster et al. 2019), which infect the secondary bacterial endosymbiont *Hamiltonella defensa* (Degnan & Moran 2008; Oliver et al. 2009, 2010) of hemipterans and other cosmopolitan symbiotic bacteria such as *Arsenophonus* spp. (Duron 2014). APSE phages encode diverse toxins within a highly variable "toxin cassette" region of their genomes (Rouïl et al. 2020). We found another APSE toxin, *apoptosis inducing protein 56* (*aip56),* fused to an additional full-length *cdtB* copy in nuclear genome sequences of all *Drosophila ananassae* subgroup species examined (Verster et al. 2019). The synteny of *aip56* and *cdtB* genes is the same in APSE genomes, which further supports HGT of toxin genes from APSE phages to insect nuclear genomes.

Subsequently, we serendipitously discovered a full-length *cdtB* sequence in the nuclear genome sequence of the gall midge *Contarinia nasturtii* (Diptera: Cecidomyiidae) (**Table S1**). The Cecidomyiidae (Diptera: Nematocera) contains over 6,600 fly species with diverse life histories, behaviors and host use patterns (Yukawa & Rohfritsch 2005; Dorchin et al. 2019; O'Connor et al. 2019). Many cecidomyiids create destructive galls on crops (Hall et al. 2012). Previously, an APSE-3-like rearrangement hotspot (RHS) toxin gene was found in the genome of the wheat pest *Mayetiola destructor* (Zhao et al. 2015). To further investigate the extent of HGT from APSE to cecidomyiid genomes, we conducted tblastn searches using proteins encoded by APSE genomes as queries against all publicly available cecidomyiid whole genome sequences: *C. nasturtii, M. destructor, Sitodiplosis mosellana,* and *Catotricha subobsoleta* (**Table S2**). Each of these species had genomic reads and assembled contigs, and all but the latter had transcriptomic data available. We generated a short list of HGT candidates by excluding top matches to *bona fide* insect genes, hits <50 AA long, and hits on short scaffolds (for more details see **Supplementary Methods**).

We considered there to be strong evidence of HGT if the candidate gene of interest (GOI) met at least 2/5 of the following criteria:

1. Non-anomalous read depth via BWA analysis
2. The GOI is on scaffolds with other *bona fide* eukaryotic genes
3. The GOI is syntenic in two or more species
4. The GOI is transcribed in dT-enriched transcriptomes
5. The GOI is predicted to have introns

Many of these genes show some signatures of eukaryotic domestication (**Supplementary Text**). For *C. nasturtii*, we validated HGTs with PCR and bi-directional Sanger sequencing (see **Supplementary Methods** and **Table S4**) of genomic DNA from larvae and adults of this midge species. A summary of our QC methods for each species is shown in **Supplementary File 1** and our list of HGT candidates is shown in **Table 1**.

The short list of HGT candidates almost exclusively includes toxin genes. They are *aip56, cdtB, lysozyme, rhs toxin*, and *Shiga-like toxin B* (*sltxB*). Additionally, we found multiple copies of an APSE-4 hypothetical protein in *S. mosellana*. This gene is found in the “toxin cassette” of APSE genomes (Rouïl et al. 2020). We excluded it from further analyses as it is poorly characterized form a functional perspective. To discern the timing and provenance of these HGTs, we incorporated phylogenetic information and, where applicable, synteny information (see **Figure 1, Figure S1,** and **Table S3**). We then used structural analysis with Phyre2 (Kelley et al. 2015) and MAFFT (Katoh et al. 2019) to determine if they have retained their function following transfer into insect genomes. Below we summarize our findings for each of the HGT candidates.

Aip56. Aip56 is a secreted toxin of *Photobacterium damselae* subsp. *piscicida,* a fish pathogen that induces apoptosis of blood cells (do Vale et al. 2017). We previously found the Aip56 B domain encoded in a fusion gene comprised of a full-length *cdtB* copy and a partial *aip56* copy. The Aip56 B domain facilitates internalization to target cells (Pereira et al. 2014), and was horizontally transferred to the *Drosophila ananassae* species complex from an APSE-like phage (Verster et al. 2019).

Insect Aip56 protein sequences form a paraphyletic clade consisting largely of insects or insect symbionts (**Figure 2, Figure S1**). Cecidomyiidae Aip56 is closely related to sequences that include Lepidoptera-associated viruses and several other insect taxa (**Figure 2**), indicating *aip56* has been transferred multiple times in insects from the APSE phage lineage. As in previous studies (Silva et al. 2013; Verster et al. 2019), we did not find conservation of the zinc-binding motif HEXXH in insect or insect-associated sequences, so catalytic activity is likely absent in insect Aip56 (**Figure 3**). Short domains appear to be conserved in the Aip56 B domain (Verster et. al. 2019), which is necessary for cellular uptake of the toxin (Silva et al. 2013; Pereira et al. 2014). However, given the dearth of information about Aip56, it is difficult to assess their biological significance. Still, domains in insect Aip56 show homology to immunity proteins and lectin-binding motifs (**Table S5**), suggesting an immune function.

CdtB. CdtB is a DNase I encoded within the genomes of diverse Actinobacteria and Proteobacteria (Jinadasa et al. 2011). CdtB complexes with CDT subunits A and C to enter eukaryotic cells, after which CdtB nicks the DNA, triggering mitotic arrest and apoptosis (Jinadasa et al. 2011). In aphids, CdtB plays a role in resistance of aphids to parasitoid wasps (Oliver et al. 2009), and we suspect it retains this function in drosophilids (Verster et al. 2019).

Since we only found *cdtB* in the genome of *C. nasturtii* among the gall midge genome sequences we searched, we infer that it was introduced into the genome after the split with *S. mosellana* ancestors ca. 70 mya (Dorchin et al. 2019). The CdtB phylogeny shows *C. nasturtii* CdtB is monophyletic with other insects, endosymbiotic bacteria, and phage (**Figure 2,** see also **Figure S1**), consistent with our previous study (Verster et al. 2019). We hypothesize that CdtB was donated from the genome of an endosymbiotic bacterium or the APSE phage lineage to a *C. nasturtii* ancestor. Consistent with previous analyses (Verster et al. 2019), amino acid residues important for CdtB metal binding, DNA binding, and enzyme activity (Jinadasa et al. 2011) were conserved in *C. nasturtii* (**Figure 3, Figure S5),** indicating the ancestral function is maintained.

Lysozymes. Lysozymes hydrolyze glycosidic bonds in peptidoglycan, a component of bacterial cell walls. Lysozymes play diverse roles including in immune defense, bacterial digestion, bacterial cell wall synthesis, and release of mature phages from infected bacterial cells (Van Herreweghe & Michiels 2012).

Interestingly, our phylogeny shows that Cecidomyiidae lysozyme genes were transferred from fungi, rather than from the APSE phage lineage. The cecidomyiid lysozyme sequences (*M. destructor* + *C. nasturti*i) are nested in a highly supported monophyletic clade sister to the fungal phyla Ascomycota and Basidiomycota, distant from APSE lysozyme sequences (**Figure 2, Figure S1**). This fungal lysozyme clade is sister to a large clade of lysozymes from Proteobacteria, consistent with the fact that GH25 lysozymes were transferred across the tree of life from Proteobacterial donors (Metcalf et al. 2014). The cecidomyiid and fungal lysozyme sequences share high structural similarity with phage lysozyme GH24 (**Table S5**). Previous studies show fungal GH24 lysozyme may play a role in nematode defense (Plaza et al. 2015; Kombrink et al. 2019). Many residues vital for binding and catalysis (Shoichet et al. 1995) are the same between insect, fungal, and phage lysozyme sequences, indicating a conserved proximal function in insects (**Figure 3**).

RHS toxins. Rearrangement hotspot (RHS) toxins, or YD-repeat toxins, are found widely among bacteria and archaea (Jamet & Nassif 2015). RHS toxins are large and highly polymorphic, consisting of several tyrosine/aspartate (YD) repeats that are involved in trafficking and delivery of the toxin and a variable C-terminal domain that catalyzes the enzyme's toxic activity (Zhang et al. 2012).

The cecidomyiid RHS proteins form a single clade sister to *Xenorhabdus vietnamensis*, a symbiotic bacteria of the entomopathogenic nematode *Steinernema sangi* (Lalramnghaki et al. 2017) (**Figure 3**). This clade, in turn, is sister to a group that includes *Xenorhabdus* and *Photorhabdus* species, which are associated with nematode parasitism of insects (Boemare 2002; Busby et al. 2013). The most inclusive clade includes APSE phage and associated endosymbionts. The APSE-3 RHS toxin has been implicated in parasitoid defense, as pea aphids harboring *H. defensa* infected with APSE-3 had near-complete protection against parasitism by the parasitoid braconid wasp *Aphidius ervi* (Martinez et al. 2018). The cecidomyiid RHS sequences retain residues important for toxin function. Insect RHS sequences maintain the YDXXGR core repeat motif shared among bacterial RHS toxins (Wang et al. 1998). Additionally, three residues involved in C-terminal autoproteolysis, R650, D663, and D686 (Busby et al. 2013), are conserved in insect RHS toxin copies (**Figure 3**). Cecidomyiid RHS toxins show structural similarity to the insecticidal *P. luminescens* Tc toxin complex (**Figure S8**), consistent with a toxic functional role.

SltxB. Shiga-like toxins (Sltxs) are ribosome-inactivating toxins (Chan & Ng 2016). Sltxs are AB_5 toxins, where the B pentamer binds to globotriaosylceramide (Gb3) binding sites to retrograde traffic the active A subunit into the cell (Malyukova et al. 2009).

Most cecidomyiid SltxB sequences form a monophyletic clade sister to an unidentified bacteria isolated from *Populus* trees (Crombie et al. 2018). This clade is sister to APSE SltxB sequences, so the gene was likely originally transferred from an APSE-like ancestor (see **Figure 2** and **Figure S1**). Synteny (**Table S3**) between *C. nasturtii* and *S. mosellana sltxB* sequences show that *sltxB* was transferred to a common ancestor prior to their divergence ca. 70 mya (Dorchin et al. 2019). The gene was duplicated in *C. nasturtii* and appears in tandem copies on several scaffolds in the *S. mosellana* genome (**Table S3**). Most insect SltxB sequences form a large polytomy, consistent with a recent expansion (Whitfield & Lockhart 2007). We found several motifs involved in Gb3 binding and cytotoxicity (Bast et al. 1999) were conserved in insect and bacterial SltxB copies. Residues that contribute to cytotoxicity, including F50, A63 and G82 (Clark et al. 1996), were highly conserved between bacterial and cecidomyiid species

(**Figure 3**). Phyre2 analyses show several insect SltxB sequences have retained a typical oligomer-binding fold, a SltxB structure (**Table S5**) consistent with our expectations.

Across all of the toxin genes we discovered in galling midge genomes, our phylogenetic analyses show reticulated patterns of ancestry between species. For example, in one case, APSE-like phages may not have been the ancestral donors (e.g. *lysozyme*), and, in another, HGT occurred in several insect taxa (e.g. *aip56*). Having several possible donors for HGT is not unexpected in cecidomyiids, since HGT has been reported from APSE phages (Zhao et al. 2015), fungi (Cobbs et al. 2013), and even other cecidomyiid species (Ben Amara et al. 2017). We observed that, in most cases, the phylogenies show that HGT occurred between insects and microbes associated with insects (**Figure 2, Figure S1**). The intimate associations between plant- and microbe-feeding insects, bacterial symbionts, and their phages may lead to HGT events because their DNA is in close proximity to the insect germline, which could facilitate transfer of DNA into the germline nucleus. While a simplistic model, we could find no other evidence in the literature of studies that have formally addressed if associations between insects and their symbionts facilitate HGT. We tested this ‘transfer-by-proximity’ hypothesis (that sharing similar habitats facilitates HGT) using the δ value (Borges et al. 2019), a metric for measuring the degree of phylogenetic signal for categorical traits. Phylogenetic signal is the tendency of traits from related species to resemble each other more than species drawn at random from the same tree for a phenotype (Blomberg & Garland 2002), and higher δ values correspond to higher phylogenetic signal (Borges et al. 2019). Every species on our protein phylogenies was assigned a “Lifestyle” by searching the existing literature, which included living on plants, plant roots, soil, arthropods (e.g. endosymbionts), or other habitats. We compared the δ value for the observed (“true”) trait distribution across all phylogenies with those for which the traits were shuffled randomly across the tips. We found that the δ values for the “true” trait distribution were consistently higher than the distribution of δ values when the trait was shuffled along the phylogeny (**Table S6**). Thus, there is some non-random association between habitat and transfer of genes between species. However, not all sampled tips represent HGT events, and the overabundance of vertical transfer events limited our ability to use this metric to test the hypothesis. To address this issue, we also calculated δ after removing vertically inherited tips from the phylogeny (see **Supplementary Methods**) and obtained similar results (**Table S6**). Our analysis of phylogenetic signal suggests that physical proximity of genetic material (e.g. that between eukaryotic hosts and their endosymbionts) may facilitate gene transfer. Many of the species closely related to our candidate insect genes are associated with insects, such as APSE, *Trichoplusia ni* ascovirus, or *Xenorhabdus vietnamensis*. This analysis suggests, but does not prove, that the intimate associations between these species and their insect hosts may have facilitated ancient opportunities for HGT.

Our previous knowledge of prokaryote-to-insect HGT events centered on genes involved in conferring new metabolic capabilities, particularly those that allow insects to colonize new plant hosts (Wybouw et al. 2016). HGTs of toxin-encoding genes are largely antibacterial in function, when function can be acsertained (Di Lelio et al. 2019; Hayes et al. 2020; Li et al. 2021). The galling midge HGT dataset as a whole highlights that a new functional class of proteins, toxins that antagonize eukaryotic cells, may be common among insects. Given that many of these horizontally transferred toxins (with the exception of lysozyme) target eukaryotic cellular components, the genes encoding them may have become integrated into existing immunological networks to protect cecidomyiids from attack by parasitoid wasps or other eukaryotic enemies. Notably, the cecidomyiid species sampled in our study face strong selective pressure from a wide number of parasitoid taxa (Chen et al. 1991; Abram et al. 2012; Chavalle et al. 2018). We hypothesize that *cdtB, rhs,* and *sltxB* in particular may protect developing cecidomyiid larvae and pupae from parasitoid wasps, since these three genes confer this protective function in other insects (Oliver et al. 2009; Martinez et al. 2018; McLean et al. 2018).

Our work contributes to the growing body of literature on HGT in eukaryotes, particularly of eukaryotic-targeting toxin genes. Here we also took a new approach for assessment of phylogenetic signal and found results consistent with the transfer-by-proximity hypothesis of animal HGT from microbes living in similar environments as the insects. Further sampling of genomes across Cecidomyiidae can corroborate the timing of these HGT events, in addition to revealing more about the evolution and biology of this agriculturally important insect family.

Acknowledgments:

We thank Dr. Yolanda Chen and Andrea Swan at University of Vermont for providing *Contarinia nasturtii* samples. Discussions with Dr. Iswar Hariharan and Dr. Michael Shapira motivated the phylogenetic signal analysis. Nicolas Alexandre, Dr. Carrie Malina, Dr. Jenna Ekwealor, and Dr. Jennifer Wisecaver provided valuable feedback for bioinformatic analysis and phylogenetic tree estimation. *Funding:* KIV was funded by the National Science Foundation Graduate Research Fellowship as well as grants from the Golden Gate Science into Action Fund at Golden Gate National Recreation Area and University of California - Berkeley Integrative Biology Summer Research Award. RLT was funded by the University of California - Berkeley, Berkeley Fellowship and the National Science Foundation Graduate Research Fellowship. SMA was funded by the National Institute of Health - Bridges to the Baccalaureate program, Berkeley Transfer Scholarship, and the Cal Alumni Scholarship. This work was also supported by the National Institute of General Medical Science of the National Institutes of Health award number R35GM119816 to NKW.

Data Availability:

Genomic and transcriptomic resources utilized in this text are shown in **Table S2**.

References:

1. Abram PK et al. 2012. Identity, distribution, and seasonal phenology of parasitoids of the swede midge, *Contarinia nasturtii* (Kieffer) (Diptera: Cecidomyiidae) in Europe. Biol. Control. 62:197–205.
2. Alsmark C et al. 2013. Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. Genome Biology. 14:R19.
3. Bast DJ, Banerjee L, Clark C, Read RJ, Brunton JL. 1999. The identification of three biologically relevant globotriaosyl ceramide receptor binding sites on the Verotoxin 1 B subunit. Mol. Microbiol. 32:953–960.
4. Ben Amara W et al. 2017. An overview of irritans-mariner transposons in two *Mayetiola* species (Diptera: Cecidomyiidae). Eur. J. Entomol. 114.
5. Blomberg SP, Garland T Jr. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods: Phylogenetic inertia. J. Evol. Biol. 15:899–910.
6. Boemare N. 2002. Systematics of *Photorhabdus* and *Xenorhabdus*. Entomopathogenic nematology. 35–56.
7. Borges R, Machado JP, Gomes C, Rocha AP, Antunes A. 2019. Measuring phylogenetic signal between categorical traits and phylogenies. Bioinformatics. 35:1862–1869.
8. Busby JN, Panjikar S, Landsberg MJ, Hurst MRH, Lott JS. 2013. The BC component of ABC toxins is an RHS-repeat-containing protein encapsulation device. Nature. 501:547–550.
9. Chan YS, Ng TB. 2016. Shiga toxins: from structure and mechanism to applications. Appl. Microbiol. Biotechnol. 100:1597–1610.
10. Chavalle S, Buhl PN, San Martin y Gomez G, De Proft M. 2018. Parasitism rates and parasitoid complexes of the wheat midges, *Sitodiplosis mosellana*, *Contarinia tritici* and *Haplodiplosis marginata*. Biocontrol. 63:641–653.
11. Chen BH, Foster JE, Araya JE, Taylor PL. 1991. Parasitism of *Mayetiola destructor* (Diptera: Cecidomyiidae) by *Platygaster hiemalis* (Hymenoptera: Platygasteridae) on Hessian Fly-Resistant Wheats. J. Entomol. Sci. 26:237–243.
12. Clark C et al. 1996. Phenylalanine 30 plays an important role in receptor binding of verotoxin-1. Mol. Microbiol. 19:891–899.
13. Cobbs C, Heath J, Stireman JO 3rd, Abbot P. 2013. Carotenoids in unexpected places: gall midges, lateral gene transfer, and carotenoid biosynthesis in animals. Mol. Phylogenet. Evol. 68:221–228.
14. Crombie AT et al. 2018. Poplar phyllosphere harbors disparate isoprene-degrading bacteria. Proc. Natl. Acad. Sci. U. S. A. 115:13081–13086.
15. Degnan PH, Moran NA. 2008. Diverse phage-encoded toxins in a protective insect endosymbiont. Appl. Environ. Microbiol. 74:6782–6791.
16. Di Lelio I et al. 2019. Evolution of an insect immune barrier through horizontal gene transfer mediated by a parasitic wasp. PLoS Genet. 15:e1007998.
17. Dorchin N, Harris KM, Stireman JO 3rd. 2019. Phylogeny of the gall midges (Diptera, Cecidomyiidae, Cecidomyiinae): Systematics, evolution of feeding modes and diversification rates. Mol. Phylogenet. Evol. 140:106602.
18. Duron O. 2014. Arsenophonus insect symbionts are commonly infected with APSE, a bacteriophage involved in protective symbiosis. FEMS Microbiol. Ecol. 90:184–194.

19. Hall DR et al. 2012. The chemical ecology of cecidomyiid midges (Diptera: Cecidomyiidae). J. Chem. Ecol. 38:2–22.
20. Hayes BM et al. 2020. Ticks Resist Skin Commensals with Immune Factor of Bacterial Origin. Cell. 183:1562–1571.e12.
21. Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. Nat. Rev. Microbiol. 16:67–79.
22. Jamet A, Nassif X. 2015. New players in the toxin field: polymorphic toxin systems in bacteria. MBio. 6:e00285–15.
23. Jinadasa RN, Bloom SE, Weiss RS, Duhamel GE. 2011. 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.
24. Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20:1160–1166.
25. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 10:845–858.
26. Kombrink A et al. 2019. Induction of antibacterial proteins and peptides in the coprophilous mushroom *Coprinopsis cinerea* in response to bacteria. ISME J. 13:588–602.
27. Lalramnghaki HC, Vanlalhlimpuia, Vanramliana, Lalramliana. 2017. Characterization of a new isolate of entomopathogenic nematode, *Steinernema sangi* (Rhabditida, Steinernematidae), and its symbiotic bacteria *Xenorhabdus vietnamensis* (γ-Proteobacteria) from Mizoram, northeastern India. J. Parasit. Dis. 41:1123–1131.
28. Li H-S et al. 2021. Horizontally acquired antibacterial genes associated with adaptive radiation of ladybird beetles. BMC Biol. 19:7.
29. Malyukova I et al. 2009. Macropinocytosis in Shiga toxin 1 uptake by human intestinal epithelial cells and transcellular transcytosis. Am. J. Physiol. Gastrointest. Liver Physiol. 296:G78–92.
30. Martinez AJ, Doremus MR, Kraft LJ, Kim KL, Oliver KM. 2018. Multi-modal defences in aphids offer redundant protection and increased costs likely impeding a protective mutualism. J. Anim. Ecol. 87:464–477.
31. McLean AHC et al. 2018. Consequences of symbiont co-infections for insect host phenotypes. J. Anim. Ecol. 87:478–488.
32. Metcalf JA, Funkhouser-Jones LJ, Brileya K, Reysenbach A-L, Bordenstein SR. 2014. Antibacterial gene transfer across the tree of life. Elife. 3. doi: 10.7554/eLife.04266.
33. O’Connor TK, Laport RG, Whiteman NK. 2019. Polyploidy in creosote bush (*Larrea tridentata*) shapes the biogeography of specialist herbivores. J. Biogeogr. 46:597–610.
34. Oliver KM, Degnan PH, Burke GR, Moran NA. 2010. Facultative Symbionts in Aphids and the Horizontal Transfer of Ecologically Important Traits. Annu. Rev. Entomol. 55:247–266.
35. Oliver KM, Degnan PH, Hunter MS, Moran NA. 2009. Bacteriophages encode factors required for protection in a symbiotic mutualism. Science. 325:992–994.
36. Pereira LMG et al. 2014. Intracellular trafficking of AIP56, an NF-κB-cleaving toxin from Photobacterium damselae subsp. piscicida. Infect. Immun. 82:5270–5285.

37. Plaza DF, Schmieder SS, Lipzen A, Lindquist E, Künzler M. 2015. Identification of a novel nematotoxic protein by challenging the model mushroom *Coprinopsis cinerea* with a fungivorous nematode. G3 . 6:87–98.
38. Rouïl J, Jousselin E, Coeur d'acier A, Cruaud C, Manzano-Marín A. 2020. The protector within: Comparative genomics of APSE phages across aphids reveals rampant recombination and diverse toxin arsenals. Genome Biol. Evol. 12:878–889.
39. Shoichet BK, Baase WA, Kuroki R, Matthews BW. 1995. A relationship between protein stability and protein function. Proc. Natl. Acad. Sci. U. S. A. 92:452–456.
40. Silva DS et al. 2013. The apoptogenic toxin AIP56 is a metalloprotease A-B toxin that cleaves NF-κb P65. PLoS Pathogens. 9:e1003128.
41. do Vale A, Pereira C, Osorio CR, dos Santos NMS. 2017. The apoptogenic toxin AIP56 is secreted by the type II secretion system of *Photobacterium damselae* subsp. *piscicida.* Toxins . 9. doi: 10.3390/toxins9110368.
42. Van Herreweghe JM, Michiels CW. 2012. Invertebrate lysozymes: diversity and distribution, molecular mechanism and in vivo function. J. Biosci. 37:327–348.
43. Verster KI et al. 2019. Horizontal transfer of bacterial cytolethal distending toxin B genes to insects. Mol. Biol. Evol. 36:2105–2110.
44. Wang YD, Zhao S, Hill CW. 1998. Rhs elements comprise three subfamilies which diverged prior to acquisition by *Escherichia coli*. J. Bacteriol. 180:4102–4110.
45. Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–265.
46. Wybouw N, Pauchet Y, Heckel DG, Van Leeuwen T. 2016. Horizontal gene transfer contributes to the evolution of arthropod herbivory. Genome Biol. Evol. 8:1785–1801.
47. Yukawa J. 2005. Biology and ecology of gall-inducing Cecidomyiidae (Diptera). Biology, ecology, and evolution of gall-inducing arthropods. 1: 273-304.
48. Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L. 2012. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biol. Direct. 7:18.
49. Zhao C et al. 2015. A massive expansion of effector genes underlies gall-formation in the wheat pest *Mayetiola destructor.* Curr. Biol. 25:613–620.

Table 1. Final list of HGT candidate genes from sequenced cecidomyiid genomes.

Species	**Protein Name**	**Scaffold(s)**	**GenBank ID or scaffold coordinates**
C. nasturtii	Aip56	NW_022197544.1 NW_022200251.1	LOC116352557 LOC116349575
	CdtB	NW_022197544.1 NW_022203704.1	LOC116352617 LOC116351853
	Lysozyme	NW_022201606.1	LOC116350899 LOC116350797 NW_022201606.1: 2605394-2605828
	SltxB	NW_022197768.1	LOC116338454 LOC116338453
S. mosellana	Aip56	VUAH01001532.1	VUAH01001532.1: 2829-2410
	Hypothetical Protein	VUAH01004499.1 VUAH01000029.1	VUAH01004499.1: 1380-1123 VUAH01000029.1: 4532498-4532776 4520662-4520964
	RHS	VUAH01000029.1	VUAH01000029.1: 8-4950
	SltxB	VUAH01000060.1 VUAH01000016.1 VUAH01006166.1	VUAH01000060.1: 423458-423673 425034-425252 VUAH01000016.1: 900502-900260 896597-896355 901304-901146 899522-899256 VUAH01006166.1: 4792842-4792618
M. destructor	Aip56	GL501523	GL501523: 1201296-1200694
	Lysozyme	AEGA01007297.1 AEGA01024319.1 GL501497	Mdes015794 Mdes01448 GL501497: 230410-230841
	RHS protein	GL501425	Mdes011034

Fig 1. Maximum likelihood Cecidomyiidae species phylogeny (log likelihood = -10311.662040) shows the approximate timing of each HGT event. The tree was built in RAxML using concatenated sequences of the *co1* (542 nt), *CAD* (1439 nt), *ef1a* (725 nt), and *28S* (429 nt) genes (see **Table S7** for accession IDs for genes included in the species phylogeny**)**. *Bibio marci* (Diptera: Bibionidae) is included as an outgroup. Filled boxes indicate presence of the toxin, and the numbers indicate copy number if >1. Bootstrap values are reported out of *n* = 1000 bootstraps, and scale bar is substitutions per site. Tick marks on the phylogeny indicate approximate timing of the HGT event based on a parsimony approach incorporating presence/absence of the HGT candidate, individual gene phylogenies, and synteny data. Dashed ticks indicate HGT events for which synteny data were inconclusive.

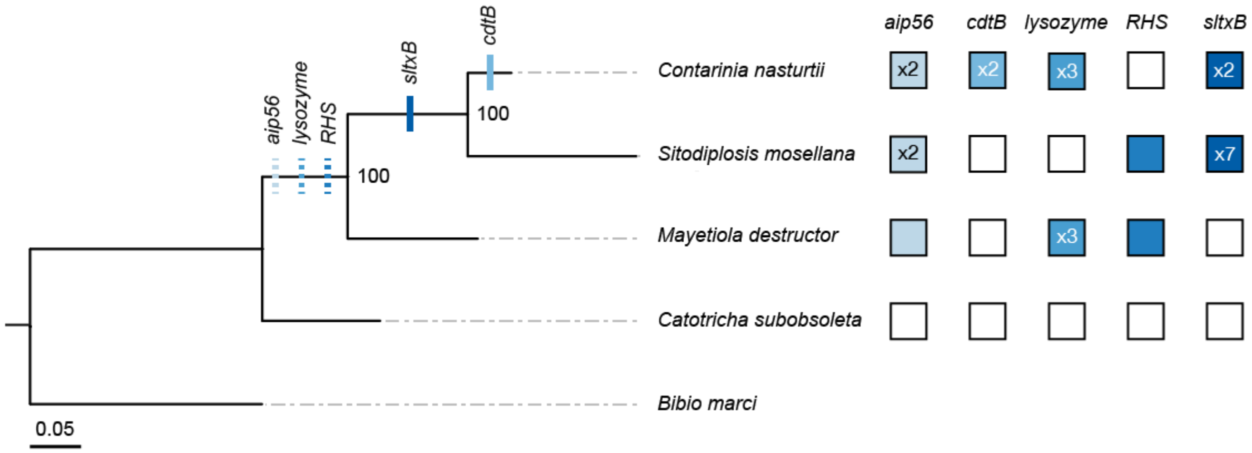


Fig 2. Simplified phylogenies of horizontally transferred genes show they are derived from several sources, including endosymbiont bacteriophage, Proteobacteria, and fungi. Indicated co-associated species suggest opportunities facilitating HGT events. Black organisms are species, while grey organisms are co-associated species. For full phylogenies, see **Figure S1**.

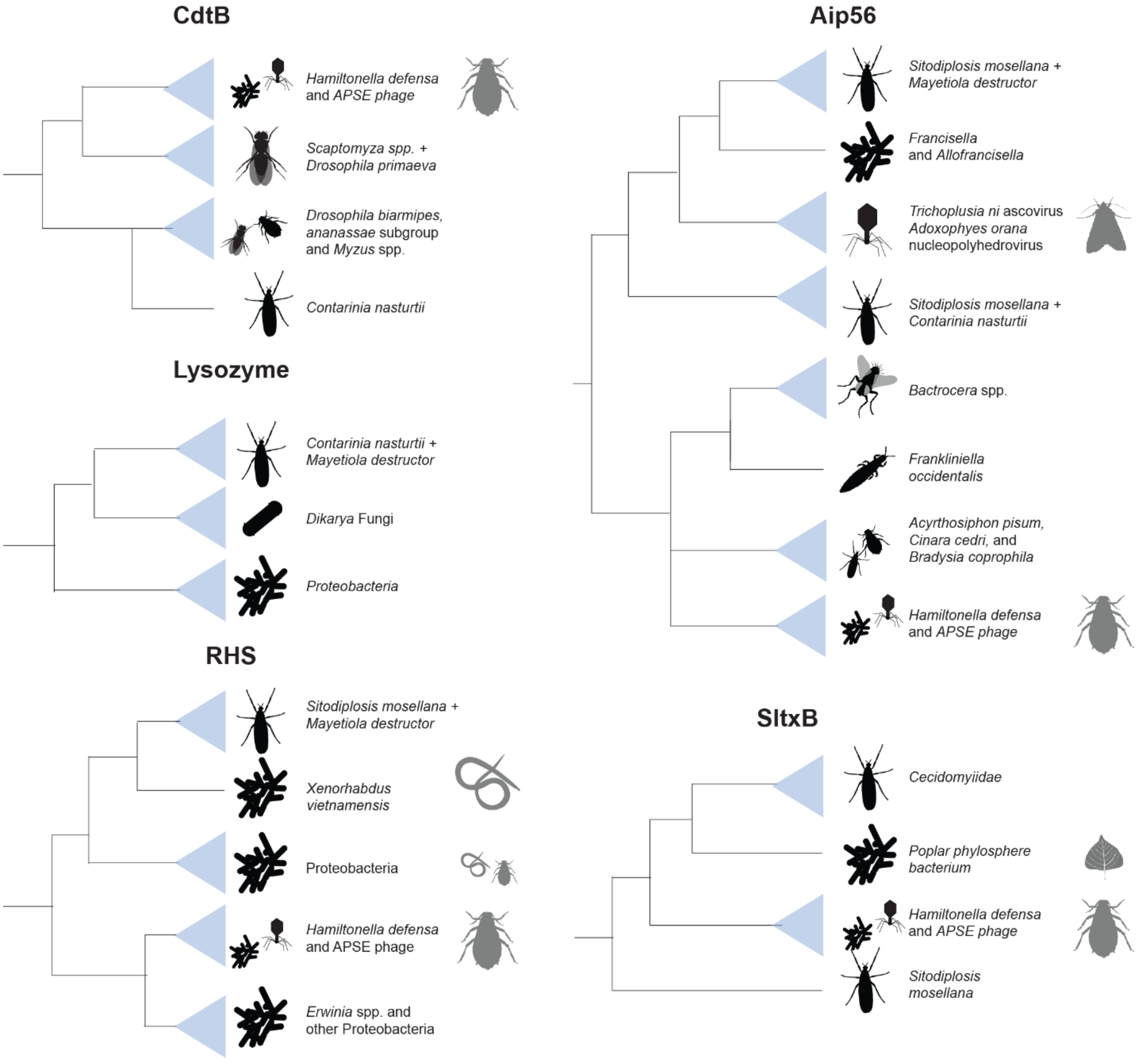
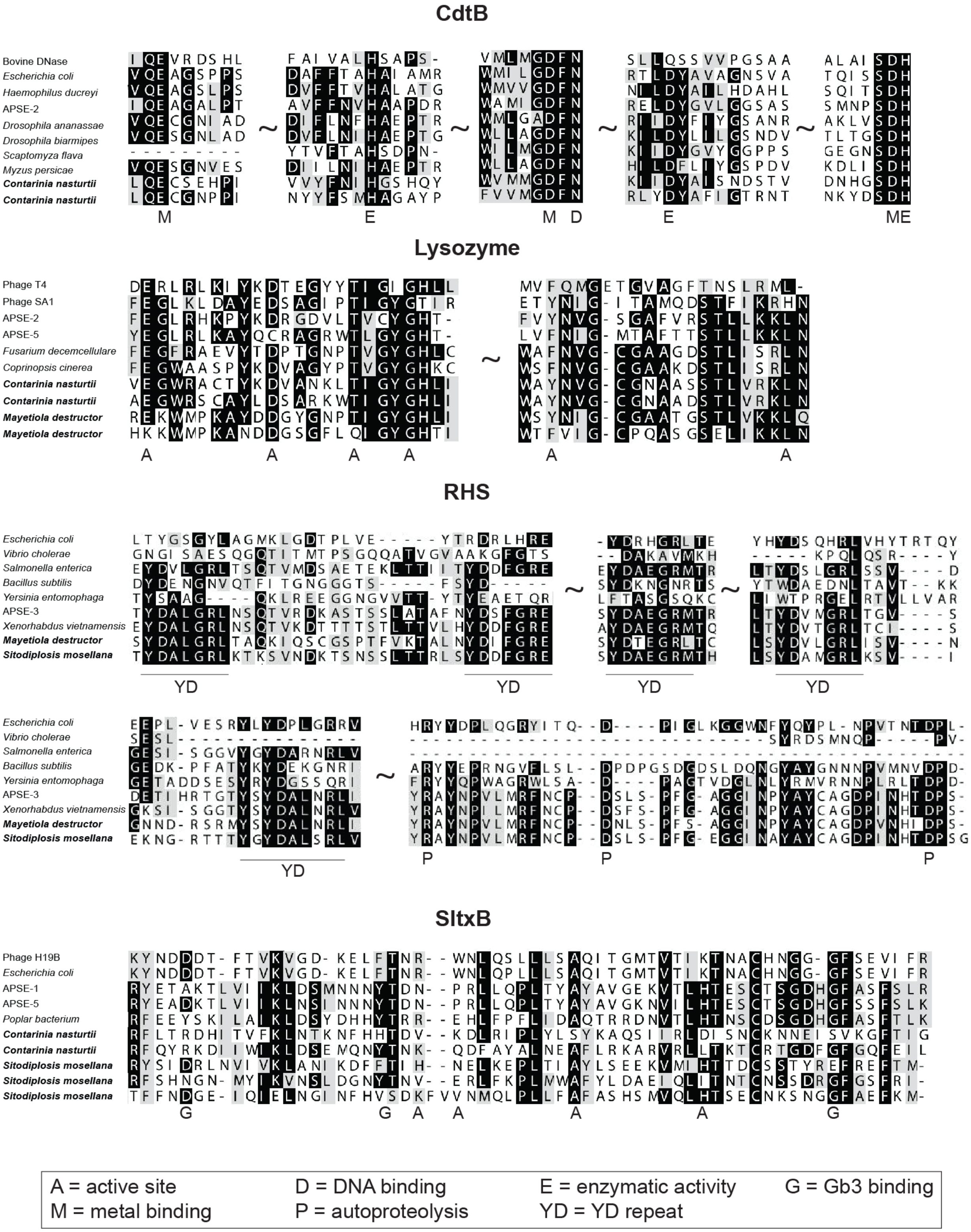

CdtB
Hamiltonella defensa and *APSE phage*
Scaptomyza spp. + *Drosophila primaeva*
Drosophila biarmipes, *ananassae* subgroup and *Myzus* spp.
Contarinia nasturtii
Lysozyme
Contarinia nasturtii + *Mayetiola destructor*
Dikarya Fungi
Proteobacteria
RHS
Sitodiplosis mosellana + *Mayetiola destructor*
Xenorhabdus vietnamensis
Proteobacteria
Hamiltonella defensa and APSE phage
Erwinia spp. and other Proteobacteria
Aip56
Sitodiplosis mosellana + *Mayetiola destructor*
Francisella and *Allofrancisella*
Trichoplusia ni ascovirus *Adoxophyes orana* nucleopolyhedrovirus
Sitodiplosis mosellana + *Contarinia nasturtii*
Bactrocera spp.
Frankliniella occidentalis
Acyrthosiphon pisum, *Cinara cedri*, and *Bradysia coprophila*
Hamiltonella defensa and *APSE phage*
SltxB
Cecidomyiidae
Poplar phylosphere bacterium
Hamiltonella defensa and *APSE phage*
Sitodiplosis mosellana

Fig. 3. Protein alignments of toxin proteins transferred to midges reveal that many critical residues are conserved (see **Supplementary Methods** for accession IDs of representative sequences).



Supplementary Methods

Identifying horizontal gene transfer (HGT) candidates in cecidomyiids.

We used HGT screening methods described in Nikoh et al. (2010), but adjusted to the scope of this study and the bioinformatic resources available. A summary of our methods is shown in **Figure S2**. To identify possible HGT candidates in the Cecidomyiidae, we ran TBLASTN on APSE proteomes against existing genomic and/or transcriptomic resources for Cecidomyiidae species (see **Table S2** for proteomic queries and Cecidomyiidae databases). These searches were conducted throughout June-July 2020. We initially retained all hits for consideration as HGT candidate Genes of Interest (or HGT-GOIs). Sequences were eliminated as HGT candidates if BLASTP searches of the predicted subject amino acid sequence (either the High-scoring Segment Pair, predicted ORF, or whole length predicted annotation) to the NCBI nr database showed the top 2+ hits were to canonical insect genes. If HGT candidates were < 50 continuous amino acids long, they were removed from consideration. Redundant hits, defined as hits where the same HGT-GOIs from different APSE strains mapped back to the same genomic coordinate, were then removed. We also removed hits encoded on scaffolds <1 kb long, as these are highly likely to be bacterial contaminants or mis-assembled regions. Additionally, we removed hits with an E-value > 0.01.

Quality Control (QC) for HGT candidates.

We considered strong evidence of HGT if the gene of interest (henceforth referred to as ‘GOI’) met 2/5 of the following criteria:

i. Non-anomalous read depth via BWA analysis (**Supplementary File 1**)
ii. The GOI was on scaffolds with other *bona fide* eukaryotic genes (**Supplementary File 1**)
iii. The GOI is syntenic in two or more species (**Table S3**)
iv. The GOI is transcribed in dT-enriched transcriptomes (**Table S2**)
v. The GOI is predicted to have introns (**Supplementary File 1**)

In the case of *Contarinia nasturtii*, we validated the HGTs with PCR and bi-directional Sanger sequencing (see **Supplementary Methods** and **Table S4**) of genomic DNA isolated from larvae and adults of this midge species. In cases where the distance between the GOI and a proximal gene was <2000 bp, we amplified regions that included other *bona fide* eukaryotic genes. A summary of our QC methods for each species is shown in **Supplementary File 1.** Our final list of HGT candidates is shown in **Table 1.**

To determine if multiple HGT-GOIs were actually duplicates or a consequence of mis-assembly, we compared the scaffolds of gene duplicates using progressiveMauve (Darling et al. 2010). If there was >90% nucleotide identity between scaffolds, we considered those mis-assembly artifacts. For tandem duplicates, we further used BWA to analyze read depth at each paralog to determine if one or all paralogs had anomalous read depth that could signal mis-assembly. Additionally, if encoded genes were <10% of the size of the canonical, functional protein, they were discarded as candidates.

A simplified workflow to identify functional HGT candidates is shown in **Fig S2**.

Burrows Wheeler Alignment (BWA) analysis.

We aligned Illumina reads (see **Table S2** for SRA accessions) to the genome via BWA (Li & Durbin 2009) to search for unusual coverage depth relative to neighboring genes, which can be due to contamination (Koutsovoulos et al. 2016). Read quality and trimming were assessed with FastQC (Andrews 2010), which showed high per base sequence quality, low per base N content, and low adapter content. The read alignment was visualized and assessed in the software package Geneious v. 11.1.5 (https://www.geneious.com). Since the majority of the genes were encoded on scaffolds encoding other *bona fide* eukaryotic genes, we included the read depth of all candidate scaffolds, per species, in a Grubbs' test and removed scaffolds with read depth outliers. Following this, we did the same with the loci containing the horizontally transferred genes (HTGs). The results show there are no coverage abnormalities, indicating these candidate HTGs are not due to assembly artifacts or microbial contamination.

Transcription analysis.

We submitted the GOI (+/- up to 20 kb up and downstream) as a blastn query to representative polyA-enriched transcriptomes. These representative transcriptomes are shown in **Table S2**. The top hits (≤ 5000) were extracted and mapped back to the region using Geneious RNA Mapper (Sensitivity: Highest Sensitivity / Slow; Span annotated introns). We report the mean read depth and standard deviation across the GOI. Limitations of the transcriptional analysis are that characterizations of any given transcriptome are contingent on tissue type and life stage, and as such, the expression patterns of one transcriptome may not reflect the importance of the gene in a species.

Synteny analyses.

Possibly due to the high divergence between sequenced species (e.g. *C. nasturtii* and *S. mosellana* are estimated to have split ~70 mya [Dorchin et al. 2019]), macro-syntenic analyses using progressiveMauve (Darling et al. 2010) and CoGe SynMap (Lyons et al. 2008) were not fruitful. Instead, we employed a qualitative micro-syntenic approach. In annotated genomes, we extracted the protein sequences of genes up and downstream of the GOI and indicated their position with -*n* or +*n* (for example, a positionality of -3 indicates the gene is located three genes upstream of the GOI). These sequences were then submitted as TBLASTN queries (Altschul et al. 1997) to the representative genomes. The scaffolds of top hits were then extracted; if there were no hits, we indicate "NA" in the cell. We considered there to be some evidence of synteny if one or more genes proximal to the GOI were located on the same scaffold within a species. Results are shown in **Table S3.**

gDNA extraction and PCR conditions.

Ethanol-preserved samples of *C. nasturtii* larvae and adult males and females from a lab-reared colony were provided to us by Andrea Swan and Dr. Yolanda Chen at University of Vermont. Ethanol-preserved specimens were rehydrated in sterile water and allowed to dry on a Kimwipe prior to DNA extraction. Rehydrated specimens were homogenized with a bead-beater for 2 minutes at 30Hz. DNA was extracted from the homogenized samples using a DNEasy Kit

(Qiagen) with an overnight Proteinase K digestion at 55˚C. gDNA samples contained 1-5 larvae or a single adult midge.

We designed PCR primers to capture the HGT-GOI and, if the nearest predicted gene was <2kb distant, a neighboring *bona fide* eukaryotic gene. PCR primers were designed using Geneious. PCR reaction mixes were composed of: 7.5µl Failsafe Premix E (Epicentre), 4.2µl nuclease-free water, 1.2µl each of F and R primers (IDT), 1µl template DNA, and 0.12µl *Taq* polymerase (New England Biolabs). Thermocycler settings were: 5 m at 95°C and 35 cycles of 95°C for 30 s, Ta for 30 s, and 68°C for 1-2.5 m depending on amplicon size (see **Table S4**), followed by a final 5 m extension at 68°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 of PCR product was mixed with 1 µL ThermoScientific 6X Loading Dye. 1Kb Plus DNA ladder (Invitrogen) was included as a molecular marker. PCR product was run on gels using the Owl™ EasyCast™ B1 Mini Gel Electrophoresis System rigs for 30 minutes at 120V. Gels were visualized using AlphaImager™ Gel Imaging System (Alpha Innotech). Primer sequences, the region captured by the amplicon, melting temperature, extension times, and the expected amplicon length are detailed in **Table S4** along with gel images. All PCR products were Sanger sequenced in both directions at the UC Berkeley DNA Sequencing facility.

Species phylogeny and ancestral state reconstruction.

Nucleotide sequences for *co1*, *cad*, *ef1a*, and *28S* were retrieved from GenBank for each of the five species included in the species phylogeny (**Table S7**). *Bibio marci* (Diptera: Bibionidae) was included as an outgroup to the Cecidomyiidae family, consistent with phylogenies previously generated for the family (Sikora et al. 2019). Each gene was aligned individually using the default settings on the MAFFT v. 7 web server (Katoh et al. 2019). Individual gene alignments were inspected and manually trimmed before concatenation. The final alignment consisted of five species and a total of 3135 nucleotide sites. Total sequence lengths for each gene were as follows: *co1* (542 nt), *cad* (1439 nt), *ef1a* (725 nt), *28S* (429 nt). The concatenated alignment was uploaded to CIPRES web portal for maximum likelihood (ML) tree construction. A ML tree was generated using RAxML-HPC2 on XSEDE using default settings (Miller et al. 2010; Stamatakis 2014). The ML species tree is shown (log likelihood = -10311.662040) with bootstrap values at each node (*n* = 1000 bootstraps) (**Figure 1**).

Due to the low number of taxa on our tree, maximum likelihood approaches to timing HGT events were uninformative. We used maximum parsimony (MP) to infer the timing of each HGT event by incorporating data from synteny analyses and protein phylogenies. Briefly, we assumed a single acquisition of the HTG in the common ancestor if there was evidence of shared synteny among the taxa in which the HTG was found (**Table S3**). In the absence of synteny data, we examined the protein phylogenies to determine timing of HGT events (**Figure S1a**). We interpret monophyly of cecidomyiid protein sequences as a single acquisition event in a common ancestor under a MP model. Acquisition events that are only supported by protein phylogeny data are indicated on the species tree with dashed ticks (**Figure 1**).

Protein phylogeny construction.

Representative toxin sequences were queried against the NCBI refseq protein database on 11/20/2020 using BLASTP (Altschul et al. 1997) with a maximum of 500 top hits per query (see below for a list of query sequences used per toxin). Top hits were extracted for each sequence, and redundant sequences were removed with cd-hit (Li & Godzik 2006; Huang et al. 2010) with a 0.8 similarity cutoff, unless they were genes specifically identified in this manuscript.

Sequences were aligned with MAFFT v. 7.312 using the E-INS-I strategy and the BLOSUM62 amino acid scoring matrices (Katoh & Standley 2013). Sequences were trimmed to include only the conserved protein domains (i.e., domains in which <50% of the sequences had gaps). After trimming, sequences were re-aligned with the earlier MAFFT settings.
Gene topologies were inferred using maximum likelihood as implemented in W-IQ-TREE (http://iqtree.cibiv.univie.ac.at/) (Nguyen et al. 2015; Trifinopoulos et al. 2016) using the best-fit model as assessed by BIC in ModelFinder (Kalyaanamoorthy et al. 2017). The resultant consensus tree was constructed from 1000 ultrafast-bootstrapped trees (Hoang et al. 2018). Nodes with <50% bootstrap support were collapsed to polytomies using the di2multi function in ape v5.4 (Paradis & Schliep 2019).

Life history of represented sequences in each phylogeny were determined by searching the literature (see **Supplementary File 2**) while Phylum information was taken from NCBI Taxonomy (Sayers et al. 2019; Schoch et al. 2020). Life history and phylum information were used to annotate the phylogeny. Phylogenies were visualized and annotated using ggtree v. 2.5.0.991 (Yu et al. 2017; Yu 2020).

Specifics of each phylogeny are shown below:

1. **Aip56 queries:** Queries: *Drosophila bipectinata* (XP_017099943.1, AIP56 domain), *C. nasturtii* (XP_031636937.1, XP_031641113.1), *M. destructor* (this manuscript), *S. mosellana* (this manuscript), APSE1 (NP_050970.1), APSE4 (ACJ10093.1), APSE5 (ACJ10079.1). Total tips: 90. Model: LG+G4. LogL= -40095.6183, BIC=81322.3145.
2. **CdtB.** Queries: Candidatus *Hamiltonella defensa* (XP_016857353.1), *C. nasturtii* (XP_031641203), *D. ananassae* (XP_014760894.1), *D. biarmipes* (XP_016950904.1), *Myzus persicae* (XP_022165116.1), and *Scaptomyza flava* (QDF82162.1). Total tips: 76. Model: LG+F+I+G4. LogL = -27554.2607, BIC = 56112.4328.
3. **Lysozyme queries:** Queries: APSE2 (YP_002308525.1), *C. nasturtii* (XP_031638744.1), *M. destructor* (this manuscript). To make the tree clearer, we removed one large and highly divergent clade from the phylogeny which included *Hamiltonella defensa* and APSE-2 phage sequences. Total tips: 172. Model: WAG_I_G4. LogL=-43487.1457, BIC = 90213.0519.
4. **SltxB queries:** Queries: APSE1 (NP_050968.1), APSE5 (ACJ10077.1), *C.nasturtii* (XP_031619577.1, XP_031619578), *Burkholderia ambifaria* (WP_175804727.1), and *S. mosellana* (copies identified in this manuscript). Total tips: 28. Model: VT+G4, LogL=-3729.8063, BIC=7016.7700.
5. **RHS queries:** Queries: Bacteriophage APSE3 (CAB3775397.1), *Candidatus* Hamiltonella defensa (ATW32053.1), *S. mosellana* (copies identified in this manuscript), and *M. destructor* (Mdes011034). Total tips: 188. Model: WAG+F+I+G4. LogL = -312544.0798, BIC = 628028.7246.

Measuring phylogenetic signal.

For all species in a phylogeny, we assigned a "Lifestyle" trait that fell under *Arthropod*, *Plant*, *Nematode*, *Plant Root*, *Soil*, or *Other*, assignments that were meant to generally describe the lifestyle of the species. *Other* included mammalian pathogens, free-living oceanic bacteria, synthetic constructs, or other lifestyles that did not fall under the named categories. We searched the literature to determine these assignations, and citations for all species that are not *Other* are shown in **Supplementary File 2.**

We utilized Borges' δ value to evaluate phylogenetic signal of the species' lifestyle traits (Borges et al. 2019). The value of δ can be any positive real number. The higher the number, the higher the phylogenetic signal (Borges et al. 2019). This can be compared with the δ value of the same tree with randomized or shuffled traits to assess significance. To determine whether to "shuffle" traits (i.e. re-arrange the traits) or randomly assign traits, we piloted this analysis with shuffled and randomized trait sets, using lambda = 0.1, se=0.5, sim = 10,000, thin=10 and burn=100 in R (R Core Team 2017). We found that the shuffled trait set has a higher δ value, and as such is a more conservative method that we continued to implement.

We calculated the δ value using lambda = 0.1, se=0.5, sim = 10,000, thin=10 and burn=100 in R (R Core Team 2017). The originally calculated phylogenies were used, with one modification. We did not utilize the di2multi() function in ape (Paradis & Schliep 2019) which was implemented in our original phylogenies in order to be compatible with the δ calculation. To determine whether the realized δ value is statistically significant, we randomized the trait *n*=100 times along the phylogeny and calculated δ for each shuffling using the replicate() function in R (R Core Team 2017). The real value was compared to the randomized distribution of δ values. P-value was calculated as the number of simulations in which the shuffled δ is higher than the realized δ, a strategy utilized in several recent studies (Gruson et al.; Pinna et al. 2020; Ronget et al. 2020). δ values for the actual tree and trait assignments are reported in **Table S6.**

One major caveat is that not all sampled tips represent HGT events, and the overabundance of vertical transfer events may limit our ability to use this metric to test the hypothesis that similar lifestyles between organisms facilitates HGT. To improve the robustness of our conclusions, we removed vertically inherited tips from our phylogenies. To analyze our phylogenies *without* vertical descendance, we used the phylogenies calculated above. Then, we used the drop.tip() function in ape (interactive=TRUE) to manually remove one tip per sister taxa, *if* the two sister taxa were both from the same genus. One exception is in the case of the SltxB phylogeny, where we removed multiple cecidomyiid tips due to strong evidence of vertical descendance, as described in the manuscript. This process was repeated iteratively until the final trimmed tree had no sister taxa from the same genus. We show a representative example of the two trees, with vertical descendance and without, side-by-side in **Figure S3**.We then calculated the real and shuffled δ values as described above on the pruned tree (shown in **Table S6**).

We acknowledge there are still flaws in the above designs. First, our protocol may not eliminate all instances of vertical transfer, as vertical descendance will often link members of different genera. Furthermore, our protocol may also erroneously remove some instances of horizontal gene transfer, as we consider there may theoretically be transfer of genes *within* a genus. Additionally, the oversimplification of "Lifestyles" could lead to artificially high δ values.

Structural analysis.

To model and predict protein structure and function for representative proteins, we used the Phyre2 web portal (Kelley et al. 2015) using the "Normal" modeling mode (**Table S5**).

To determine at the residue level whether or not vital catalytic residues are preserved in disparate lineages, we used the MAFFT aligner as described above with representative sequences (Katoh & Standley 2013). The representative sequences shown in **Figure 3** are indicated in **Table S8.** Note that due to the low conservation between Aip56 insect and characterized sequences, we did not show this alignment. Alignments shown in **Figure 3** were visualized using BoxShade (https://embnet.vital-it.ch/software/BOX_form.html).

Supplementary Text

Domestication of various bacterial toxins following horizontal gene transfer from prokaryotes to eukaryotes

The presence of eukaryotic transcriptional motifs in putative HTGs may indicate domestication of a gene of prokaryotic origin in its novel eukaryotic context. Here, we analyze HGT sequences for motifs related to eukaryotic transcription, largely following methods described in Verster et al. (2019). Briefly, we analyzed the regions flanking our candidate HTGs for core promoter elements identified by transcription initiation factors TFIID and TFIIB (summarized in (Thomas & Chiang 2006)), alternative transcription initiation elements such as the GC box (Blake et al. 1990) or CAAT box (Graves et al. 1986; Raymondjean et al. 1988), and transcription termination elements such as polyadenylation signals, cleavage sites (CA), and upstream and downstream sequence elements (summarized in (Proudfoot 2011)). We also searched the sequences for the Shine-Dalgarno sequence, a motif essential for bacterial ribosome binding (Shine & Dalgarno 1974), which can indicate that our HTG may be a bacterial contaminant, as well as motifs for eukaryotic translational start sites, like Kozak sequences (Cavener 1987). This list is not exhaustive, nor will every element described above necessarily be found in all eukaryotic genes (Kutach & Kadonaga 2000).

We did not analyze candidate HTGs from *S. mosellana* since the genome was unannotated, making it difficult to accurately predict gene boundaries. Additionally, we did not analyze the horizontally transferred lysozyme copies, since phylogenetic analyses indicate these were transferred from a eukaryotic donor (see **Main Tex**t, **Figure 3** and **Figure S1**).

C. nasturtii

Legend:

- Predicted exons are highlighted in blue and predicted introns are highlighted in yellow. Exon/intron boundaries for *C. nasturtii* are taken from the GenBank assembly annotations.
- Coding sequences are indicated in bold text. 5' and 3' UTRs are therefore designated by unbolded text highlighted in blue.
- Poly(A) signals or cleavage sites are underlined. Upstream and downstream sequence elements are italicized.

- Intergenic regions (between *sltxB* copies) are designated by lowercase text.
- TATA box motifs are designated in orange. Initiator sequences are highlighted in white text when found within annotated mRNAs, or in blue text if found outside the gene boundaries. Kozak sequences are highlighted in green.

cdtB on NW_022197544.1:

TCGAAATTTATTTTTTGTTTTTGCTTTTCATTACATTTTCAGACATTGAATTGTGCGTG
TATCTTTTTGTGTGTTGGACAAAA**ATGAAGGGAGTTCTATTTTTTGGATTCATGTG**
TGCAACATTCATGGTAAATTTTAAATATAAAACTTAAATGTTCATTCCAGTTTGAAA
TTTCAATATTTGCTTTTTTTTTTTGTTTTTGTTCAG**AATTGTTATGGAAGGGTTTGTAG**
TGATATGGCATATGATATCCCGATGGCTACGTGGAATAGTCAAGGCGGAAGAT
GGGGTACGGTGAAAACATTGTTATCACATACGTATCCTGATGTCGAGGTACTTG
CATTGCAAGAATGCGGCAATCCGCCTATTGATCCAGCTATAGCACTTGTTGGAA
ATGGTAATATACCAACACAATGGTCTGATAATAGACCATATTTGACTGTGCAGT
GAGTAAAACTTCTGTTGCATTTTTATTTATAAATTATGTTGATAACTATAAGTGAATC
TAACTAATATAAAACAATTTCTATAG**ATATATATCCTATTTCTATAATAATGGTGA**
TATTAATCAAAGAGATAATGGAAGTGGCGCAAAGGAATATACTATCAATGTGA
GAGATAGAGGCCATGTTCAAAGAATATATTATTTATATCATTACGAATTACAAT
TGGCTGGAAATGTAAGAACAAATACTGCCATAATCACGAAAACAAGAGCGAAT
GAAGTGTTTGTTTTAATTGATCCAAATGAATCAAGACCAGTAATTGGCTTCCGA
ATTGGAAGTAATTATTATTTTAGTATGCATGCTGGGGCTTATCCGAAAAATCCT
TCTCCAGATACAGTCTCACAAATTGCACAATTTGTTAGCAACAACGCAATGCCT
ATGGAAGACGTTTCATTTGTAGTAATGGGCGATTTCAACACAGAGCCTAATTAT
TTTAACCCCACCGTAATACCTCGGGGTTTCCATTTTACAAAAGTTCTACCATCT
GAGAAAACACAAGGTCTTGGCAATACTGTAGTTAGGTTATACGATTATGCATTT
ATTGGAACAAGAAACACTTGTGAATTTCCAAATTTTATTGTAAACGCTGGGAAT
AAGTACGACAGTGATCATCGAGTTGTTGTATTTCGAAGACAATAGCAAG*TTTTAA*
GCATAGGACTGGACTAATAAACGTCAATAGGGGTTGTCATTC*TTGTTT*AATCAATTTA
CTTCAAATTTTCAATAAAGAAAAAAATGTTTCAAA

cdtB on NW_022203704.1:

TATAAATATGAACTTTATAATTTTGTGAAATCATTCATATTTTGAATGTCGAATTGTG
AATCGATTCTCAAATATGCAAGGGAAATTAAAACGATCAGTGTTATTATTGTCGAAG
GAATAATTCTATCAAGCATCAAATATTAGCATTGCGAGAGTTAGTATCAAGTTATTT
TGAAGAATTTTGGGAAAATTAACTACCACACAAAAACAGATGTTCGATGTTGAAGA
GTATAATAATAATTTAATCCTTTATAATTATTTCAAATATCCACAGAATAAAT**ATGA**
ATACTGAGGTCAGCGATTTTGTTTTCGTTACATGGAACACCGAAGGTTCCAATT
GGGAGAATGTCGCAAAGTTGATGCTCAGCAAAGAAGGTGTAGACCGAATCGAT
GTTATGGCTCTTCAAGAGTGTAGTGAACATCCCATCTCTGATGAGCATCCACAG
GCGGCAAAATCAGGTATTGGAACCATATCTCTACCTGCCAATGAAGTAGAAAAT
ATTCAAATAAATCCTACACCGAATCAACGAGACAATTCTTCAGGAATCACTACC
TACCTATGGCATTTCAACACAGAAGCAGAGTTTTTTTTGTATTATCGAAACAAC
AAGATATTTACTGAAATTGGTGCAGTTGGTGGATCAGGAAAAAAGAGCCTGAG

TAGTGCCTTTGTAACTAGAGTCGAGGCTACTAGAACTTTTTACATGGCTCCAGT
CGATAATAACGGTAGTCCAGTAAATAATAGTGACTACAATAAAAACCGTCCTGT
TATTGGGATTGAGGTGAATGGTGTCGTGTATTTCAACATTCATGGGTCGCATCA
GTATAATAATTCTGTAAATAATACCATTAGAATTATAAAAGAGTTTATGGCAAG
GAACCATCCTGCAATAAAATGGGTAATGATGGGTGACTTCAATAAAACGCCACT
AGAGATTAATACTGCGGGATTGTATTTGATGGAACCAAACACAGCCACTCGTG
CTAAGAGCGGTAAGATAATTGATTATGCCATATCTAATGATTCAACCGTTGGGA
ATATGCATATACATGTTTCACGTGACAATCATGGATCAGACCATTTTCCAGTGG
AATTTATCAATAAAAATATAAACCACAACAGTTTTGGATCATTTTAGCTTTCATT
AATGACGGCACTAAAGAAGCCATAAGAAAACCATAGA*TTTATT*GCCTAAAACGATA
AAAAATAAAATTCACATATAGTTGGCGCTGCATTCTATGCAAAC*TTGTTTT*CCGTTGC
A

aip56 on NW_022197544.1:

TTTTTTTAAATTTCGCTTGCTTTTGAGAAAGAAAAAATATTATAATATAGTGTTTGAT
TTATACGGAAACATTTAAAA**ATGTTAAAATGTATACTTTTGGCTCTTGCATTTAC**
GCATTTGTCTCAAACCAACTTTTTAGATAGCGGTAAATTGTTTATTTTTTTCCAAA
TTTTATTGAATTTTTATATAATTTGTTATTGTGTACGTACAACAG**GTCCATTGGACT**
GGACTTGCATGTCATGGCTTGCCGGTTCACGGAACAGGCCGTCAGTTCAAGAT
TCTACTTTTTATAACTTGGTGCACAATATTCCTTCTACAAGAAATGCAGATTTAC
ACTTCTCATTATCGCGACATCACATAATACCTTATAAAGTGTTATATAAATTTTT
TAATACCGTATTGGAGTTGGGTGAATCTAACTCAAGAATTCATTTTTTGCTTGG
GCAATTCTTGTCGAATCTACTTGTCGATTTAATGGTTCGCGCTCCCGGCGCTCA
AGATCAGACGGAAACCGAAGTTTTACAATCCATTCGCAATTTATTTCAAGGCTT
TGATGGTGCTGGATTTGGTGATACAGTGCAGGGAGAAATCCGAACATTTCGTT
TACCGAACTTTCAAGAAAACATAGTGGCATTCCGTCATCTACCGCTCACCGAGC
TAACAAATACACAACTCGTAAGGAATATTGCTCAGGTCAGAATTCAAATGACAC
TCCAATCTCTATTGACATGGATGCCATTTAATTATTTTGAGGGGCCCAGTGGAA
TGTATCGTACCGACGAACCCGGTCCAAATTTTGAAGATAATGCGTATGTAATAA
TCGGTGAAGAAAATTCGGTTCGACTACATGACGCATACGAAACAATGATTGCA
ATTGATGAAATACGGACGCAGCATCCGACAATTATTCCGTTGCAATCTATACAA
TTTGTATTTGACAATTTAAATGTAGTTCGAGGAAACGCACCTAACGGATATGGG
CAATTTAACCTGGGATATTGGGAAGAACTCGAACCGGTTGAAAAGAAAAAAGG
TAAATTGGAATGTCAATTCAAAATCCAAACAACAAGACCAACAACCACGCCACC
GCCCCATCACAATCACCGACCTCATGAGGAATTCCGTCGAAAAAACAAAGTTTC
TTCAGGAACGAACAGCTTAGACCCGCTTTGTTTAAATACAACGTTTGAGGCTAT
CAAACTTTATACTGATAATTTATATTTAGCCTCTAAACCTGGTAAAGTGCCCAA
TGTACATCAATACATATGTAGAGGAATATGGGAGAGCTATTTAGCAGCTTTTGG
TATATCGCAATGTCTTCCGTAATACTATAATTTTTTGAATCACCATATATGGCACGT
GAATTAAAACTTTTCTTCTTTCAAAATCCGTAAATTTTTTTTAAAAAATCCTTAGTTA
GTAAAAATATATGAAAAGAAATTGAATTGGACGTTCTCGCTTTTTGAACATATATAA
ACTGGTTGGTAGTTAGCTATGGAAATTAGTTATGGGTTGCAACGTATTACATTCGCA
ATTAGATTTTTAGAAAAATTTATTGAACAACGAATCTCTTAGAACTACAATCGGATT
CTTAAATCTCATGTTGTGCTTTTAAATTCAACGAAACTATCAACTATTGATCATTTTA
AACACAATCAAAATAGTTTTCTGCTCATTTTAATTAACATTTTATACTTAATCATCAT

GTTTTAGACATTTATTAAACACTAGGCAAACAAACTTAAGTTTTTAAGATGAATTCA
TTTTTTTTCTTTTCATGTGTTCACTCCAACGAGATACATTTTAAAATAAAGAATTTTA
TTGAGTTTGAATGCTATTTTCTGCTCA*TTTTTTATTGTTTTTGTTTTTTGT*ATCGCTTTTG
ATGGAAAAAAAATAGCATTGAAGAGGATTTAATTG

aip56 on NW_022200251.1:

TTCGAGCAAGAGTTATCAATATCAATACATTAGAA**ATGATTAGACTTACATTGCTT**
CTGATTTTCGTGGCTATTCTCTCTGCGGCTCGTGAAATCGAACAGTAAATTCTAG
TTTTTGTTTTCGCTTTGTTTTGATAATTGAATATTTTTACATTTTAACGATTACGACAT
ATTTGTAAGGCGTATATATATGACATAGTACATCGTATTTAGTATCGTATAAAACCA
TAAATTTCCTTCGCCGTGGATTCACTGCTTCATTGCCGTTTTGAACTAGTTTTGTTCA
CATCAAAATTTATCTGTAGAATCTAAATAATGGTATGTGTGAATGTGTTTCATAGAA
CAAATACTTCGTGACAACAGAAATAAGTTTATTCGTTGAAAGATAAAACTTCAGATT
TTCGTTGCGGAAGCGAATGAATAATTTTTAATTCAAATAAAAAATTGCTCGTGATTC
ACATATCTTATCGACTTTTCTTCAATTCAAAAGTTACGATGTTATTTCTATAAAGCCA
TGTTGATATATATTTTCTGTTTAGGTTAATCGTTATATAAATTCGATCCGAAGTCAAT
GTTTGATAAAGAGTCTTTTCAAACTATCTTTTTTTCCATAGGTAGTTTGAATATTTCG
ATATTTTCTTTTGACATTTTGGCTCTCCCCAATTTTGACACTGAAATAGTCTAAGGAA
CTTGATTTTTGTTTCAAAGTAAAAGTATGAAATTTATTCAATATATTAAATTTATTTG
TTTTCAGAAG**TATCGAAGATGAACCGTGTACGTATGACTTGGCGCGTCATCATA**
TCATTGCATACAGTAAGGTGAAAGAGTTTTTCGAAACAGCCGCTGTGAATGTGA
AAAATCGTGAATTACGACGGCGATTGGCAAAACTATTTGAAAAATTAGCGACAC
ATCCAAATGAACCAATGAACGAAGACGACCACACTAGTCTGTTCGAACGAAAT
GCATTCGAAAAGGAGTTATCAACACCGTTGCAGTTACTTGGAATTGGCAGTTAT
GATGCGAGAAGAGTGGCTATCTCGCTGATTCGATGGATACCATTTAACATTTTC
AAGGGTCCTGCAGCTAAAAATCGAGTTGATGATCCAAAAAATGGATTCGAAGA
AAATGCTGGCCGAATTGTAAATGCGGACACACGTGACAATTTAGAAAATTTGCA
CATTTTATATGATAACATGATAAATTATGTCAACACAGACAGTGTGGATAATTT
CAAGAATTCAATTAATTTAATGGAACATTTACTGATCGCCGTACCGAAGGGATA
CAGAGATTTTACAATGTCTGATTGGGAATTGGTTGGAATTCAAATTAAAGGACA
TACTAAACTGTGTAAATTTCGAATTAAAAGTATCGATGAAAGTTAACATTTATGT
TGTGGGTGCCATTTTGAAGTCATCATTTAATGTAATGCAAAAGTATTAGAGGCTTTTT
TTTGATTTTGAATGTTTTTAAAATTGATGATTTTTACAACTTCAAGTGCGTTGTATTC
ATAGAAGTACTCTGATTTTGCAATTGTAAAAATGAATTTTTGTGGAAATTTTCATGAT
ATTCTTGTGTTATATGGAACGAAAATAGTAGACATTTTTGAAAACTAGATATTTCCA
AG*TTCTTTT*ATTCAACAAAATCATAAAAATAAAAAAAAATATTGCACTTCAAAATCA
CGAA

sltxB on NW_022197768.1:

AATCACATCCAGCACTTTCCATCTGAGTATTTGAAAAGAGTTTAATAGACCATAAAT
ATGTACGTAAAAATGATTTTACTGATTACATTCAGCCCATATGTCACAGAGAGAACA
CAGATTTAATTGACTTGCATTATATAAGTAAAAGATTTATCAGCAAT**ATGTTTTGGC**
AACTTATACTGCTTGCTTTCTTTGCAAGTGTTCGGGCAACTGAAGAAGAGCAAG

CAGACGGTAAGATGTAGCTACAACTATTATAAATTAATTTATATTTGAAACAAGTTT
CTGTTTTTTTTAAG**GAAACGGAACACCATATGAAGACGTGTGCGTCGGACTGATT**
GAATCAATAAGGTTTCAATATGAGTTTATAGAAGGGACCGGCAGGAAAGATAT
TATATGGGTGAACAAGAAAACAATGAAACTCTTTTCATATGTGTCTATTATTTCTAT
TATTTGTTGATTATTTTGAATTATTTTTATTCGATTCTCCGCTAG**ATAAAATTGGATT**
CAGAAATGCAAGAGAACTATACTAATAAGCAAGATTTTGCGTACGCACTAAATG
AAGCATTCCTGAGAAAGGCCCGAGTACGTTTGCTGACCAAAACATGTCGTACT
GGTGATTTTGGATTTGGACAATTTGAAATATTGGACATAGGACAGCAGAAATAAA
AATATTTATTTAAATTTAAAAACAGCCGTTTCGCAATCGGTTTTTATGGTCATTTGTA
ACAATATTCCGTTAATTGTAATTCGATTTATA*TTGTTT*AGAGAAGAATTCTGAATGAA
TAAATGTATACACAATTAATAAtgccttaaaattaattcattggtgagtaaatatatactaacattatcactacagcataga
acataaagaaattgtgaccaaaatctcacatattttcgatgacacacaacctgaatttgcaatttgagtatcatgagtcaacaatgttcatttttcat
tttcaatttaaatgattgtgtttcaaactaaagaaaataagtccaagtttttttttaaattcatttgtgaattctttgaagcgacatttttttaaatatccaa
gcaagcgtttggtaccttttccaaaaaaaattttcgttaaaattcaatcaaaattaactcatttatggaaatatttatttcatacatttttcacacttaatta
ttacacatattgctgcctcaacatatcataagaaaagtccatgaaaattttctttttcgaaatttatttttatcaattctatagaacgtttccatttaaatt
agtttatggttgattagctatgacggcgttgcttatatatttgagtatttataagacttgacttattaagcactttcccacttaTACTTGAAG
TGCAG**ATGACATCACAACTAT**GTAAGTCAGATATAATTTCGTTGCTCGAAAATATA
TTGCTGGAAGTCCACCCTGTTTAAACGAATATTTTTTCAATTTTTTTTTTTTTTTTCAA
ATGTATACTAG**CAAATAAGATTGATAGCAATTTGGCGGATATGTTTTTGCTCGTG**
ATACTATCGATTTACCTTTCGATTTCATCAGCTTATGTCAACGTAAGCCATGCA
GTCTCTTTCCTTCATCCATATTTTGGAGGTAAGTTGCAACAATATAGCAATAAAG
TCGAAATTACTATTTGAATTCAATAAGCTTAGATCTCACGATTATACACGCTTATAA
AGATGTCATGTAATAATATCTCACATTTCTATTGTTTCTGCCTTAG**AATATGCTGGA**
AATGATTATAAATATGAGAATTTAATGTCGAGATCGATGCCAACGACTACACCA
CAGCCAAACATACCTGATTTTGATGAAGGAGATGATAGCTGTGTTGGAAAAATA
GAATCAATCCGATTTTTAACTGGACCGTTGCATGGACATATTCGTGATCACATA
ACTGTATTTAAATTGAATACCAAAAACTTTGAACATCATACGGATGTAAAAGAT
CTTCGAATTCCCCTATATTTGTCATACAAAGCACAGTCTATTATAAGATTAGAT
ATATCAAATTGTAAAAACAACGAAATTTCAGTTAAGGGATTTACTATAGGTTTC
ATAGGCAATCCACCAGCTATAACAAATTTTCTTTTATAACAAACACGCTTTATTT
GTTCAA*TTTAAATT*CCAAATAAACTTTACCTAAAATATTAAATCTACTTTTGGAAGCA
ACGACA*TTTTT*AATTCAAACCAAATAATACATTTTCATCAGAAGGACTCTGATAATTA
CTGACCAGAGAAAACAAGGTT

M. destructor

Legend:

- Predicted exons are highlighted in blue and predicted introns are highlighted in yellow. Exon/intron boundaries for *M. destructor* are taken from Ensembl.
- Coding sequences are indicated in bold text. 5’ and 3’ UTRs are therefore designated by unbolded text highlighted in blue. The sequences for unannotated HGT candidates are highlighted in orange, with putative in-frame start and stop codons indicated in bold orange text.
- Poly(A) signals or cleavage sites are underlined. Upstream and downstream sequence elements are italicized.

- Initiator sequences are highlighted in white text when found within annotated mRNAs, or in blue text if found outside the gene boundaries.

RHS on AEGA01002600.1:

TTGATTATATGTGTCAGCATTATTTTAACCACCACAAAGATATGCACTATTAATCGA
GAACAAAGAGATTTGGAAAAGCAGTATTATGATTTTGTTGTGCGGCGTCTAAACATT
TAAATATTAAATAAATAAATAAAATAAAAATTGGAAAAATTGGAAAGGAGAGTCTT
CAGAAAGTGAAGAAATTATTTCACATAGTATGTTGAATTGATAAAGTTGTTATTGTT
GTGCCTTAACTTTTTTTTTATTAATCCCTTTTTATCTTAAATTTCGGTAACAGGTACAA
GAATCCATTAGCAACCATTTTCA**ATGGCCGTAGACAATGGATTTTTTTCGAATGC**
CAGTAATTTTCCTGCTCGCCACACGCACCACAGAGAAAAATGGTTTAAAGCGTC
GACAAGAAAAATTCAGTTATGATAATCGCAATCGATTAATTAGCTACAATGCAT
CGGGCGATAGCCTTCCGGTGGATTCGTATGGAAATTTGATGACTTCACAAACAT
ATCGATATGATGCACTGAATAATATCATTTCTATAAGAACAACATTGTTTGATA
ATTCTGTGGATAATGTTACGTATCATTATCTCAATCCCGATGATCCAACACAGT
TGACGAAAGTAACACATACACACAAGAAATATCCCGAAACCATTATGTTGAGTT
ATGACACTGAAGGCAGGCTGACTTGCGATGAAGCAGGAAGATTACTTTCTTAT
GATGTTTTGGGACGATTGATCAGCGTCAATGGAAACAACGATCGTTCAAGAAT
GTACAGCTATGATGCACTCAATCGTTTAATCGCCAAAAAAACCAGTAAAAATAA
TGAGATCCAAGAATTGTATTATCGTGGTTCGGAATTGGTGAATGAAGTGATAAA
TTCACAGAAAAAAGAAAAACGTTTCATCAAAAATGGCCACGAATGTCTGGGCG
TAAGTGACATCAATGGTCTTACAATAACGGTTGGCGATAAAAACAACAGCCTTT
TGTTGTCCAAAAATGTGAATTTCGGCAGTGAAGATATTCAGTGCCATGTTTGGT
CACCCTATGGCAGCAGTACCTCGACTGATAATCGTCTTTTAGGTTTAAACGGTG
AACGTTTTGATTTAGCCAGTGGTACATACCATCTTGGAAATGGTTATCGTGCAT
ATAACCCAGTTTTGATGCGTTTTAACTGCCCGGACAATTTAAGTCCATTTAGTG
CCGGTGGAATAAATCCATATGCTTATTGTGCAGGAGATCCAGTAAATCATATTG
ATCCATCTGGCCATTTCAGTTGGATAGCTATGACTGGAATAACTCTGAATATAG
TTGGACTCGCCCTATCTGTTTTCACGGCTGGAGCGTCTATTGCGGCTGCAGGTA
GTGTGATGGCAGCTATAAGTTCGGCTTCGGCTTCCGGATTGATCATCGGTTCAT
TGAACGTAGCTTCAGATATAACGGGGATAGCAGGTGGATCAATTGCAATTTTTA
ATCCGGAAGCGTCGTCAGCATTGGCTTGGATGTCGTTGGCATTAGGATTTTATA
GCATGCGAAGACCCATCGGAGAATTCAAATGGCTTAAGACTGGTATCGACGAT
TCCATATTTGCAGAACATATTGTCGATGATTTACCCAAAAACAGATTCACTGCA
GTCACAAGTTGATCCACGTACTGGCCAATTTATGTTGAATTTTCCAGTTGCTGA
GTTAATTGGAAATAATCAACTTGGTCCTGTATTATCGCTGTCTTTGAAGTATTC
GCCGTTGAATGGAGAAAATGAAGGATTTGGAATCGGATTTTCGATTGGACTTA
CGCGATTCAACAGTCGAACCCATTCACTAAATCTTAGTAATGGCGAACAATATC
GTGTCGGCATAGGGGCATATTGAAACTGGCAAGTGGTACGAAACTGCTTCATG
CGTTCGAAATTCATTCGGAGAAATCATAGCAGAAACTGCACAAGATTGGTTGAC
AAATAGCTCAGACAGTGAACAGTTGTTTTCAATAAAATCAGAAATGATTGATGG
TGGATGGGAAGTAATCACTAAAAATATCAGCAATAAATCATTTTCATTAAAATC
AAATATAATTCATGATGGATGGGGAATGAAGAAAGAAATTAATTTCTCCAATGG
CGTGAAACGTCTTCAGAAAATCGATCCTATTAATTTGACTGAATCGGTTTATAA
TCGTGAAATGTCGGGATCCAATCCATTGATTTGCGCTGAAATGTTCATCGAAAA

```
GGATAAAACAAGCGAGTTTCCTGTCAGAATGATCATGAACGACACAATGGGTA
ACGAATACAATCATTGTACCTATAGCTGGAACGGACTGGGCCAATTGCTCGAA
GAACAAGATGAACTACAAGCTGTCACAAAAAGAACTTATGATCCATATGACCGA
GTATTGACACAAATATTGCCTGATGGAACTATTTTAAAACAAACTTATGCACCA
CATTCAACCGAAAATAAGATTGCTTCGATTAGTGTGACTGGTATGAATGGCGAA
GGAAATGTAAAGACTTGGTTGTTGGGTACACAAAAGTTTGATAGCTTAGGAAG
ATTGACGGAGTGTGCTAGTGGTGGCCGCACTACAGTGTATAGCTATTCAGATG
CTTCGGCTGTTCCATCATTAGTTACTTTGCCATCCGGTAAAACTGTGACATACA
CTTACATCCCAGAATTGGGCAATTCAATCAGAAGTATGAAAGCGGATGGTATCA
GTCAAGATTTTCGCTATGCCCCAAGGTCAGGAAAATTATTAATGGCGCGAGAA
GCCGAATCGAAAGTCGAAAAAAATTGGTCGTCAAATGGTCAGCTGAAATATGA
GGTATTCTCGCTTAATGAAGATGCTCGTCGAGCCGAATATAAATACACATTGAA
TGGAACACTAGTTACGTATACTGATGTAGCAGGCAAAAAAATGCAATACATTAT
GGATGAGCATGGTCGAACCATTCAAATAATCGATGATAGCTTAACGATTGATTT
GTCATATGATGCTTTGGGCCGTTTAACAGCACAGAAAATCCAGAGCTGTGGAT
CACCTACTTTTGTGAAAACTGCATTAAACTACGACATTTTCGGCCGGGAGATCG
AGCGCCACATTATTGATAATAAGGATATGACCTTGATTTTATCGCAGACGTGGC
AGAAAAATGGTTCAGATAAAGTGCTTCATCAAAAGCTTAAAAATTTTCGCTTTT
GTTATACAAATGGCCATAATGAAGATGAAGGCTACACTATATTTTGGAAAGAGG
GTAAAATAGAAAAACTATCGAAAACCGAAGATGATGTCACATTTGTCACCTCGT
TCATTATGTCACCATTGGGCCGGCACATGAAATTATCTTGGGACTGGAGTGGTC
AATATGCACGACTAATAAAAATCGAAGACGAATTCAAAATTTTATGCAAAATAA
ATTACAATACAAACGTACAAGTTGAAATCTGGCCGAATACATCAGATGCATATA
AATTGAATTTTGAACTCATCAATGATGGTCAATTGGATACAATAGCCCGTAAAG
TATCAGATTCGAACACTCTTAATTGGAATTTTATTTACGAGGATGTTGGTCACA
TGAGTAACCGATTGACATTGACTGGTGTTAATTATCCTACTCACATGCAAGATA
CCGTTGAATATAAAACACAAGATGGATTACCATTTCCAAACAAATCTGGTCGTC
ACTTAAAACTTCCTTGTGTTCAGACATACACACGCAACGTGGGATTTGGTCAAC
CGGAAACTATTTGTTTCTACGAATACACTCGAAATAATTTTCTGGGATATGATG
GTGATTTCGGTGACTGGTCTGCTGACAGCGATTACCTTTATACAACTCTAACAG
ATTACACATATGGATCAAAAGAGAAGTCAGTGTGTGGTGACATCTCCATTGTTA
CTGAGCGAACATACAACAACTATCATTTACTAATCGCTGAAGAGATCAACAGAC
AAAACCATATTCATCGTACAGAATACAGTTATTATGCCGTGAAAGATTGTTTCA
TTGATGGCCAACCAGCTCAATTCCAATTGCCCAAGGTGAAAAAAGAGATCATG
CGAGATCCAACGGGTAATGTTCGTACGTTAGTAATGCATAGTGAATTCGATGAA
AATGGCAATCCAACGAAAGAAGTACATCCAAATGAAACGGTTACGATAACAAC
TTGGTATAAAGCAGAAGGTGAAAATGGATGTCCAGCCGAACCAAATGGCTTTG
TTCGTTTCATGAAGGAACAACGAAATATTCCCCGTCATATTACTGCATTTGAAC
CGGAATATTCTACAAAGTACCGTTACACTAAGCTCAGTGACACCAGATACGTAG
TACAAGAAAGTAAAACATCTTTTTGTGATGATTTCACACTTAGTGAGCGGCGTT
GGGCATATGATGAGAATAAAAGTAATGAATTGGGCCGGATTGTGTCGATTAGT
GATACTGTGTACGATTTGTCTGATAAATCAAAGTGTTACACTTCCACACAACGT
TTTACGACAACAGTTATAGACAATCAAATGACACAGAAAATCATTTTCACTGGA
CATGATGATTTGCAACAAACATCAACGCGATGTCAATCTGTCTTTAATAGTCGA
TTGTTTAGTGAAATCTCTTCACTTGGACTAGAAACACAGTATAGTTATGATCAA
TTGGGAAGGTTGGTATACCGCAAGCTATGTCCAAATTCGGATTATGAAAACCTC
```

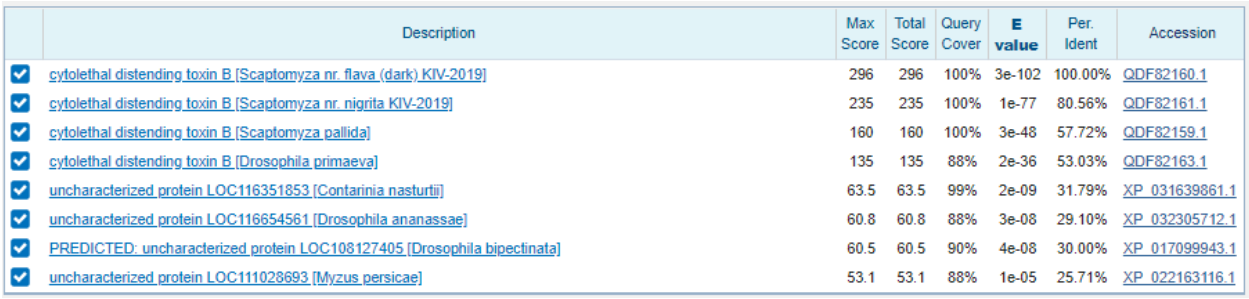
ACCACATGGGAATATGATATCAATGATGAGGGCTTACATTCAATCAAAACGGAT
GCTTCGGGAAATAAAGAGAAGACATGTTTTGATGGTACTGGACGCGCTATAAG
TTGGGAAAATTTTGAGTATTCTATCAAAAGAGTTGTTTTAAAATAGAATACAATAA
TTATTATTCGTATTATTACTTACAG**ATCTTTCTATGGAAAAAAAGGGCTTGAAGTA**
TATGGAGGAGCGTGGAAGCCACCAAAGTGGACAATTCCGCATAATAAATTAGA
TGGTTTGCAAGTAGGACAGGAAGCGAATGCTTTTCGTGTTGAGTCTAAATTGTT
TGCCGTTCGGAAAGATTCTATATTTAATGAAAATGGCACAAGAATTGCTGAGAA
AAGAAACTGGATAAAGAGCAGAAGTTCCAACACTTTTGAAAATGGAATTGATTC
ACAAAAACCAGCCTTCGCCCGGATGAATCGTATGGCCAATATGCGAAATAGAA
ACGTAACGAAACGTATTGAAAAATCACGACAACGAACACCACTCATATATAGCT
TTGGCAGGAATGACCCCATCTGGTATCTTCTTGACTAAATGAAATCCTGAGTTAT
ACACATATTTTGACTATTACATATGATCTTACTGTAACAACTAACATTAAAAATGAT
TTATAAATAAATAATCATAACAAATATGCTTGATCAATCATCAT

Aip56 on AEGA01003780.1:

ATTCAGTCTTTGACACGAACCGTATCATTACGTTTTTGGTGGATGTGGAAATGAGTC
TAAAAATTTTTGTTTTGTGCTTTTGTTGCATTACAATCGCCAATAGTGAACCATTCAG
GGCA**ATG**ATTGAATGTTTAAGACTTTGTACTCAACCCCCAATTGGACCGCCTGTGAT
CCATTATAGCAGTTTGTCGCTTCGCTATGGTGAAGATACGCGCATCCCGCCTCAAAT
CCATCATATTATACCTGCATCAAGACTTAGAAGTTTCTTCAACTTTGCTTTAGGTGAT
AGGAGAGCAACCGTAGCCCGGGATTTCATTGAATTTTTAAAGGAGTTAGCAGAATTT
GCAATGTATAGCTATTTGCCAAGTCATGTGCCAGCCGATTTATGCGATTTTATTAATG
ATATAGACAATCCAAATAGAAATGTTTTTCACAATTGGTCGGAATTGACGCCCGAAC
GACGAACGGAAATTTTAAATGCCCTAACTGGCTATGGACGCATTGTAGTTGATATGT
TCCAAATGATGCCATTCAACTTGGCCATTGGCCCAATGGAAAGATCTGACGACCCTG
GTGGAGCATTCGATCAACCCATGCAGTATATAGTTGAATCAGAACATTACAACGCGT
TGCAAAGCATAAATAGATATATGGGTGGTGCTAATAGCCATGCAGATCTTCATTATG
TTCAAGAAACAATAAGGTTATTAAGAAAAATGTTGGCGAATAATATAAGACGAGAA
CCGTACAGATATAATGAGGACGATTGGGACCGCAATAAATACACCAAAAAGTTTAG
TGTTAAAAAACAATCGAGACCTCGCCGCCAAACGGCTGAACTGTTCCCTGATATTGC
ATATTCACTGTTTGGTTTTTTGCCGGCGAGAGTGCGCAGGGAATCACAATCGTGTCA
GGACAGAACCCGTTATTCTGAAATAACTATCAAAAGCGACGAATCCTATTTTGAGAA
CGATCCTATGTGTCTAAATTCGTTTGCCAATTATAACCACGCAAGGAAAGTAGTTTT
ATTTATTATCAAC**TAA**TTTAGCAAAGTCTCTAATGTTGACATCGAATGAATTCCGGTT
TTTTTAAAAATACTACTACATATTACATAAATAATACTAACGACGAAATATTGTATG
ATCGGAATTTTTATAGCTTTTCTCATTAATTTGATCCATAGGTGTATAATCTGTACCC
TTTCTATTACAGCCAGATTTCGGCGGTGTGA*TTTGATTTT*AAATAAATAATTTTTTTTT
TTTTAAATAAACATATA*TTTTTTTTT*ATTGA

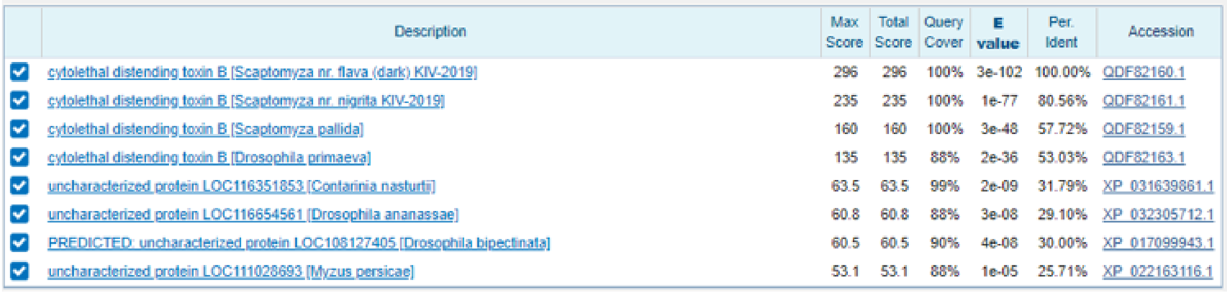
Table S1. CdtB sequence QDF82160 from *Scaptomyza* nr. *flava* was used as a BLASTP query to the NCBI GenBank: Eukarya (taxid: 2759) database on 9/28/2020. These results show that *C. nasturtii* was a hit (**Table S1a).** Another search with this newly discovered sequence, XP_031639861.1, as query shows that there are two possible *cdtB* copies in the *C. nasturtii* genome (**Table S1b**).

Table S1a.



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	cytolethal distending toxin B [Scaptomyza nr. flava (dark) KIV-2019]	296	296	100%	3e-102	100.00%	QDF82160.1
☑	cytolethal distending toxin B [Scaptomyza nr. nigrita KIV-2019]	235	235	100%	1e-77	80.56%	QDF82161.1
☑	cytolethal distending toxin B [Scaptomyza pallida]	160	160	100%	3e-48	57.72%	QDF82159.1
☑	cytolethal distending toxin B [Drosophila primaeva]	135	135	88%	2e-36	53.03%	QDF82163.1
☑	uncharacterized protein LOC116351853 [Contarinia nasturtii]	63.5	63.5	99%	2e-09	31.79%	XP_031639861.1
☑	uncharacterized protein LOC116654561 [Drosophila ananassae]	60.8	60.8	88%	3e-08	29.10%	XP_032305712.1
☑	PREDICTED: uncharacterized protein LOC108127405 [Drosophila bipectinata]	60.5	60.5	90%	4e-08	30.00%	XP_017099943.1
☑	uncharacterized protein LOC111028693 [Myzus persicae]	53.1	53.1	88%	1e-05	25.71%	XP_022163116.1

Table S1b.



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	cytolethal distending toxin B [Scaptomyza nr. flava (dark) KIV-2019]	296	296	100%	3e-102	100.00%	QDF82160.1
☑	cytolethal distending toxin B [Scaptomyza nr. nigrita KIV-2019]	235	235	100%	1e-77	80.56%	QDF82161.1
☑	cytolethal distending toxin B [Scaptomyza pallida]	160	160	100%	3e-48	57.72%	QDF82159.1
☑	cytolethal distending toxin B [Drosophila primaeva]	135	135	88%	2e-36	53.03%	QDF82163.1
☑	uncharacterized protein LOC116351853 [Contarinia nasturtii]	63.5	63.5	99%	2e-09	31.79%	XP_031639861.1
☑	uncharacterized protein LOC116654561 [Drosophila ananassae]	60.8	60.8	88%	3e-08	29.10%	XP_032305712.1
☑	PREDICTED: uncharacterized protein LOC108127405 [Drosophila bipectinata]	60.5	60.5	90%	4e-08	30.00%	XP_017099943.1
☑	uncharacterized protein LOC111028693 [Myzus persicae]	53.1	53.1	88%	1e-05	25.71%	XP_022163116.1

Table S2. Genomic and transcriptomic resources utilized in the text. **Table S2a** shows APSE genomes whose proteomes were used as queries against cecidomyiid genomes. **Table S2b** includes interrogated cecidomyiid genome and transcriptome information.

Table S2a. APSE genomic resources.

APSE Type	***Hamiltonella strain***	**Insect Host**	**Toxin**	**Assembly ID**	**Fragment ID**
APSE-1	NA	*Acyrthosiphon pisum*	Shiga-like	ENA: AF157835.1; GCF_000837745.1	
APSE-2	5AT; NY26; 82B; ZA17; WA4	*Acyrthosiphon pisum*	CdtB	GCA_000882435.1; ENA EU794049.1	EU794049
APSE-3	A1A; A2F; AS3; AS5; R7; H76	*Acyrthosiphon pisum; Aphis fabae*	YD-repeat	LR794150; GCA_902859955.1	EU794053.1; EU794057.1
APSE-4	5ATac	*Aphis craccivora*	Shiga-like		EU794051; EU794056.1
APSE-5	NA	*Uroleucon rudbeckiae*	Shiga-like		EU794050; EU794055.1
APSE-6	N4; H402	*Chaitophorus sp.; Aphis fabae*	CdtB		EU794054.1
APSE-7	NA		CdtB	ENA: LR794147; GCA_902859665	EU794052
APSE-8	3293	*Cinara watanabei*		ENA: LR794148	

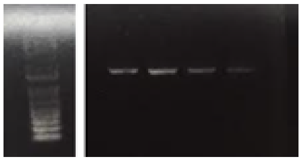
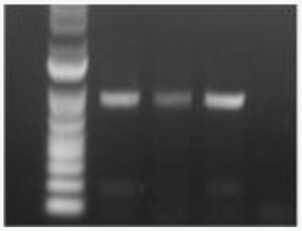
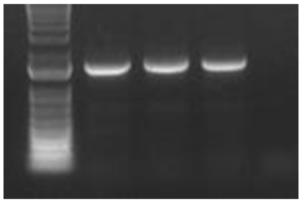
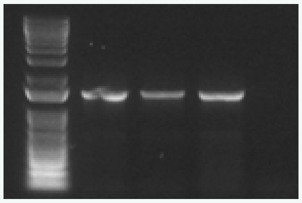
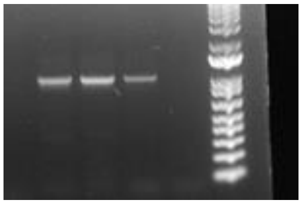
Table S2b. Cecidomyiidae genomic resources.

Species	**Taxonomy**	**Type**	**Notes**	**GenBank/SRA ID**
Catotricha subobsoleta	Cecidomyiidae: Lestremiinae: Catotrichini: Catotricha	Genome		GCA_011634745.1
Catotricha subobsoleta	Cecidomyiidae: Lestremiinae: Catotrichini: Catotricha	gDNA reads		SRX7007124
C. nasturtii	Cecidomyiidae: Cecidomyiinae: Cecidomyiini: Contarinia	Genome		GCA_009176525.2
C. nasturtii	Cecidomyiidae: Cecidomyiinae: Cecidomyiini: Contarinia	gDNA reads		SRX6846370
C. nasturtii	Cecidomyiidae: Cecidomyiinae: Cecidomyiini: Contarinia	transcriptome	L1	SAMN12767267
M. destructor	Cecidomyiidae: Cecidomyiinae: Oligotrophini: Mayetiolia	Genome		GCA_000149185.1
M. destructor	Cecidomyiidae: Cecidomyiinae: Oligotrophini: Mayetiolia	gDNA reads		SRX020192
M. destructor	Cecidomyiidae: Cecidomyiinae: Oligotrophini: Mayetiolia	Transcriptome	male L1	SRX545609
S. mosellana	Cecidomyiidae: Cecidomyiinae: Clinodiplosini: Sitodiplosis	Genome	Single pupa	GCA_009176505.1
S. mosellana	Cecidomyiidae: Cecidomyiinae: Clinodiplosini: Sitodiplosis	gDNA reads	Single pupa	SRX6820651
S. mosellana	Cecidomyiidae: Cecidomyiinae: Clinodiplosini: Sitodiplosis	Transcriptome	Non-diapause larvae	SRX312060

Table S3. Patterns of micro-synteny show that *sltxB* was likely transferred in Cecidomyiidae prior to the divergence of *C. nasturtii* and *S. mosellana*, but fragmented genomes make the timing of the other horizontally transferred genes difficult to determine. Text highlighted in yellow is the locus (or protein GenBank ID) of interest. "Position relative to GOI" indicates the number of genes upstream or downstream of the gene of interest (GOI). For example, -5 would indicate the gene is five genes upstream of the GOI. Cells highlighted in blue indicate genes are syntenic with the GOI in the considered species.

	Species	**ID type**	**-5**	**-4**	**-3**	**-2**	**-1**	**0 (GOI)**	**1**	**2**	**3**	**4**	**5**
aip56 (1)													
Query	***C. nasturtii***	**Query ID**	XP_031641173.1	XP_031616390.1	XP_031641173.1	XP_031640490.1	XP_031640489.1	XP_031641113.1	XP_031616417.1	XP_031640808.1	ref|XP_031619577.1	XP_031640287.1	XP_031640285.1
DB	***M. destructor***	**Scaffold**	GL501538.1	GL501545.1	GL501538.1	AEGA01028632.1	GL501538.1	[GL501532]	GL501450.1	GL501523.1	GL502152.1	GL502152.1	GL502152.1
DB	***S. mosellana***	**Scaffold**	VUAH01006196.1	VUAH01006196.1	VUAH01006196.1	VUAH01006196.1	VUAH01006196.1	NA	VUAH01000002.1	VUAH01006225.1	VUAH01006165.1	VUAH01006165.1	VUAH01006165.1
aip56 (2)													
Query	***C. nasturtii***	**Query**	XP_031636924.1	XP_031636923.1	XP_031636931.1	XP_031636947.1	XP_031636945.1	XP_031636937	XP_031636922.1	XP_031636912.1	XP_031636910.1	XP_031636972.1	XP_031636906.1
DB	***M. destructor***	**Scaffold**	NA	AEGA01015109.1	AEGA01031561.1	AEGA01015432	AEGA01025736.1	AEGA01034427.1	NA	AEGA01025736.1	AEGA01025736.1	AEGA01008845.1	AEGA01020694.1
DB	***S. mosellana***	**Scaffold**	NA	VUAH01000002.1	VUAH01006169.1	VUAH01006169.1	VUAH01006169.1	VUAH01001532.1	NA	VUAH01006169	VUAH01006169.1	VUAH01000002.1	VUAH01005947.1
sltxB													
Query	***C. nasturtii***	**Query**	XP_031620142.1	XP_031620140.1	XP_031619571.1	XP_031619573.1	XP_031619575.1	XP_031619578.1	XP_031619577.1	XP_031619799.1	XP_031620054.1	XP_031619946.1	XP_031619607.1
DB	***S. mosellana***	**Scaffold**	VUAH01006166.1	VUAH01006166.1	VUAH01006166.1	VUAH01006166.1	VUAH01006166.1	VUAH01006166.1	VUAH01000029.1	VUAH01006166.1	VUAH01006225.1	VUAH01006225.1	VUAH01006225.1
lysozyme													
Query	***C. nasturtii***	**Query ID**	XP_031638452.1	XP_031638329.1	XP_031638327.1	XP_031638801.1	XP_031638801.1	XP_031638744.1	XP_031638337.1	XP_031638715.1	XP_031638692.1	XP_031638673.1	XP_031638439.1
DB	***M. destructor***	**Scaffold**	GL501497.1	GL501436.1	AEGA01033347.1	AEGA01003791.1	GL501532.1	GL501497.1	GL502903.1	GL502903.1	GL502903.1	GL502903.1	GL502903.1
rhs													
Query	***M. destructive***	**Query**	Mdes011029	Mdes011030	Mdes011031	Mdes011032	Mdes011033-RA	Mdes011034	Mdes011036	Mdes-011038	Mdes011041	Mdes011042	Mdes011045
	S. mosellana	**Scaffold**	VUAH01000041.1	VUAH01000041.1	VUAH01000003.1	VUAH01000005.1	VUAH01000005.1	VUAH01006483.1	VUAH01006196.1	VUAH01000005.1	VUAH01000005.1	NA	NA
	C. nasturtii	**Scaffold**	NW_022203985.1	NW_022198578.1	NW_022198383.1	NW_022197885.1	NW_022198040.1	NA	NW_022197981.1	NW_022198574.1	NW_022198574.1	NW_022198340.1	NW_022198038.1

Table S4. Primers, reaction details, and gel images for amplification of horizontally transferred genes in the *C. nasturtii* nuclear genome.

Primers (5'->3')						
F	**R**	**Amplicon**	**Tm (˚C)**	**Extn. time (min:sec)**	**Size (bp)**	**Gel Image**
TCCATGATTGTCACGTGAAACA	TGCAAGGGAAATTAAAACGATCAGT	*cdtB* copy on short scaffold (LOC116351853)	52	1:10	1000	
GGCGATTTCAACACAGAGCC	CCCCGAAATGCCTCTACCAT	*cdtB* copy on long scaffold (LOC116352617) and nearest eukaryotic gene (LOC116352014)	52	1:10	908	
TCCGAAGACATGACAGTGCC	TTCAATCAGTCCGACGCACA	*aip56* copy on NW_022197544.1 (LOC116352557) and nearest eukaryotic gene (LOC116352216)	57	1:40	1604	
TGAATCCACGGCGAAGGAAA	ACCCACTAACGCAACCGAAT	*aip56* copy on NW_022200251.1 (LOC116349575) and nearest eukaryotic gene (LOC116349581)	57	1:40	1433	
CCTACGAAGGGCGCTAACTG	TTCAATCAGTCCGACGCACA	*sltxB* copy 1 on NW_022197768.1 (LOC116338454) and nearest eukaryotic gene (LOC116338452)	57	1:15	1054	

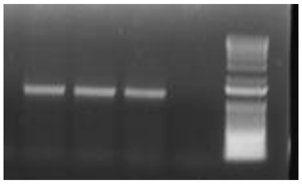
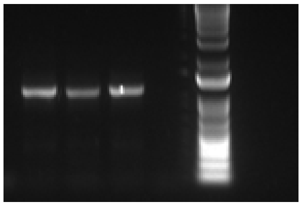
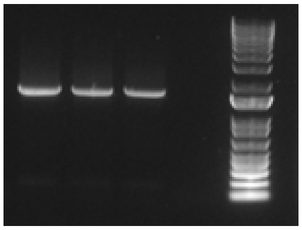
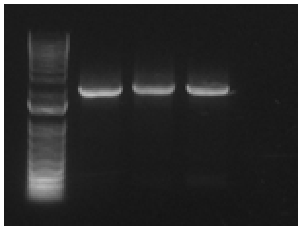
AGCATTCCTGAGAAAGGCCC	GGTGTAGTCGTTGGCATCGA	*sltxB* copy 2 on NW_022197768.1 (LOC116338453) and *sltxB* copy 1 (LOC116338454)	57	1:30	1244	
TCACCATTGTCCGTGCTCTC	CAAAGCGGGATCGTGCATTT	*lysozyme* copy 1 on NW_022201606.1 (LOC116350899) and nearest eukaryotic gene (LOC116350655)	57	1:15	1118	
GATGCAACGTTACCACAGCC	ACAACGTGCATTTCGGAAGC	*lysozyme* copy 2 (NW_022201606.1: 2605394-2605828) and lysozyme copy 3 on NW_022201606.1(LOC116350797)	57	1:30	1508	
CCAATGTCACTGCAATCGCC	ACATCCGAAGCCTCATCGTC	*lysozyme* copy 3 on NW_022201606.1 (LOC116350797) and nearest eukaryotic gene (LOC116350628)	57	2:15	2166	

Table S5. Phyre2 analyses of proteins encoded by horizontally transferred genes can provide clues as to the function of these proteins, even in novel eukaryotic contexts. **Table S5a-e** are Phyre2 analyses for, respectively: Aip56, CdtB, Lysozyme, RHS, and SltxB.

Table S5a. Phyre2 analyses of Aip56 sequences.

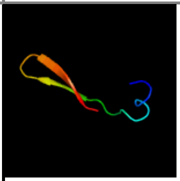
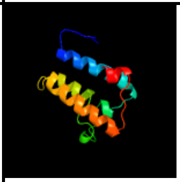
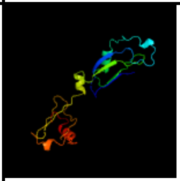
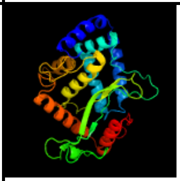
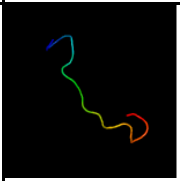
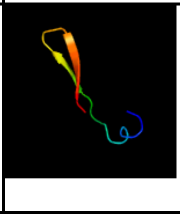
Query Sequence	**Template**	**3D Model**	**Confidence**	**% id/coverage**	**Template Information**
Hamiltonella defensa (WP_1000096555)	c6nobA		65.4	43/9	PDB header:hydrolase Chain: A: PDB Molecule:beta-fructofuranosidase; PDBTitle: structure of glycoside hydrolase family 32 from *Bifidobacterium adolescentis*
Photobacterium damselae (WP_012954632)	c4lgjA		100.0	40/50	PDB header:hydrolase Chain: A: PDB Molecule:uncharacterized protein; PDBTitle: crystal structure and mechanism of a type iii secretion protease
C. nasturtii (XP_031641113.1)	c2k19A		72.2	19/22	**PDB header**:antimicrobial protein Chain: A: PDB Molecule:putative piscicolin 126 immunity protein; PDBTitle: nmr solution structure of pisi
Bactrocera dorsalis (XP_014102626)	c2zx3B		96.7	21/31	PDB header:immune system, sugar binding protein Chain: B: PDB Molecule:csl3; PDBTitle: rhamnose-binding lectin csl3
Arsenophonus nasoniae (WP_051297127)	c4jgjA		100	32/46	PDB header:hydrolase Chain: A: PDB Molecule:uncharacterized protein; PDBTitle: crystal structure and mechanism of a type iii secretion protease
M. destructor (GL501532)	c5lnkW		47.3	28/8	PDB header:oxidoreductase Chain: W: PDB Molecule:mitochondrial complex i, sgdh subunit; PDBTitle: entire ovine respiratory complex i
Hamiltonella defensa (WP_10009655)	c6nobA		65.4	43/9	PDB header:hydrolase Chain: A: PDB Molecule:beta-fructofuranosidase; PDBTitle: structure of glycoside hydrolase family 32 from *Bifidobacterium adolescentis*

Table S5b. Phyre2 analyses of CdtB sequences.

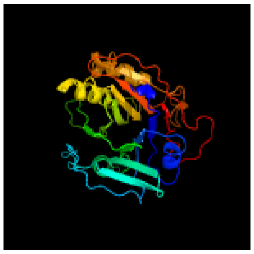
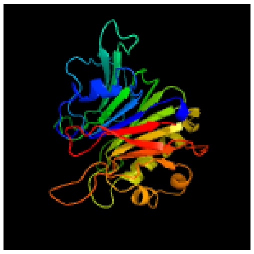
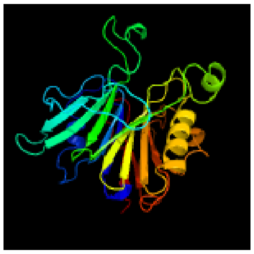
Query Sequence	Template	3D Model	Confidence	% id/coverage	Template Information
Bacteriophage APSE-2 (AGX01517.1)	d2f1na1		100.0	36/85	Fold:DNase I-like Superfamily:DNase I-like Family:DNase I-like
C. nasturtii (XP_031641203.1)	d2f1na1		100.0	29/91	Fold: DNase I-like Superfamily: DNase I-like Family: DNase I-like
C. nasturtii (XP_031639861.1)	c4k6lF_		100.0	28/95	PDB header: toxin Chain: F: PDB Molecule:cytolethal distending toxin subunit b homolog; PDBTitle: structure of typhoid toxin

Table S5c. Phyre2 analyses of lysozyme sequences.

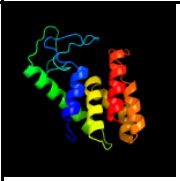
Query Sequence	**Template**	**3D Model**	**Confidence**	**% id / coverage**	**Template Information**
C. nasturtii XP_031638744	c6et6a		100	36 / 99	PDB header:hydrolase Chain: B: PDB Molecule:lysozyme; PDBTitle: muramidase domain of spmx from *Asticaccaulis excentricus*
M. destructor (GL501497 [230410-230853]	c6et6a		100	42 / 98	PDB header: antimicrobial protein Chain: A: PDB Molecule:lysozyme; PDBTitle: crystal structure of muramidase from *Acinetobacter baumanni*i ab 5075uw2 prophage
Coprinopsis cinerea okayama7 #130 (XP_001840847)	d1xtja		100	26 / 21	Fold:Lysozyme-like Superfamily:Lysozyme-like Family:Phage lysozyme

Table S5d. Phyre2 analyses of RHS sequences.

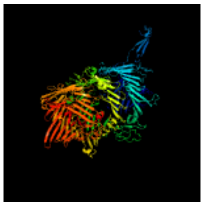
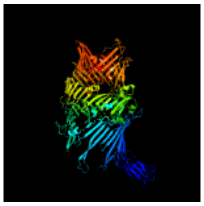
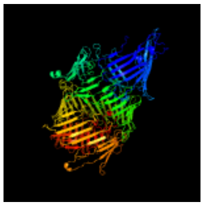
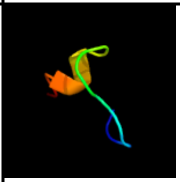
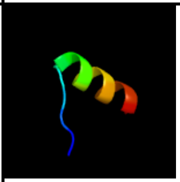
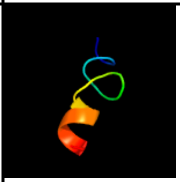
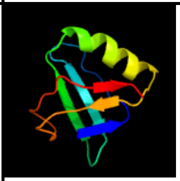
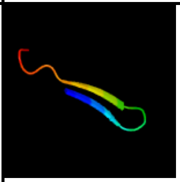
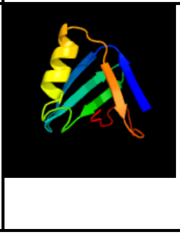
Query Sequence	**Template**	**3D Model**	**Confidence**	**% id/coverage**	**Template Information**
Xenorhabdus vietnamensis (WP_086110724.1)	c4o9xA_		100.0	17/80	PDB header:toxin Chain: A: PDB Molecule:tcdb2, tccc3; PDBTitle: crystal structure of tcdb2-tccc3
M. destructor (GL501425[246422,251816])	c4o9xA_		100.0	17/89	PDB header:toxin Chain: A: PDB Molecule:tcdb2, tccc3; PDBTitle: crystal structure of tcdb2-tccc3
S. mosellana (VUAH01006948[8,4950])	c4o9xA_		100.0	16/74	PDB header:toxin Chain: A: PDB Molecule:tcdb2, tccc3; PDBTitle: crystal structure of tcdb2-tccc3

Table S5e. Phyre2 analyses of SltxB sequences.

Query Sequence	**Template**	**3D Model**	**Confidence**	**% id / coverage**	**Template Information**
C. nasturtii (XP_03169577.1)	c5mgfC		42.8	69/14	PDB header:splicing Chain: C: PDB Molecule:snw domain-containing protein 1; PDBTitle: cryo-em structure of a human spliceosome activated for step 2 of2 splicing (c* complex)
C. nasturtii (XP_031619578.1)	c4n6cB		20.2	22/19	PDB header:structural genomics, unknown function Chain: B: PDB Molecule:uncharacterized protein; PDBTitle: crystal structure of the b1rzq2 protein from *Streptococcus pneumoniae*. northeast structural genomics consortium (nesg) target spr36
S. mosellana (VUAH01006166)	c3vgxD		28.3	67/25	PDB header: membrane protein Chain: D PDB Molecule: envelope glycoprotein gp160; PDBTitle: structure of gp41 t21/cp621-652
S. mosellana (VUAH1000060.1 [425037-425252])	d2bosa		92.4	31/97	Fold:OB-fold Superfamily:Bacterial enterotoxins Family:Bacterial AB5 toxins, B-subunits
S. mosellana (VUAH1000060.1 [423461-423673])	d2ogga2		42.1	38/21	Fold:Polo-box domain Superfamily:Polo-box domain Family:Polo-box duplicated region
S. mosellana (VUAH1000016.1 [901146-901313])	c3pl0B		27.0	23/73	PDB header:biosynthetic protein Chain: B: PDB Molecule:uncharacterized protein; PDBTitle: crystal structure of a bsma homolog (mpe_a2762) from *Methylobium petroleophilum* pm1 at 1.91 a resolution
S. mosellana (VUAH1000016.1 [899256-899522])	d1r4pb		37.7	30/78	Fold:OB-fold Superfamily:Bacterial enterotoxins Family:Bacterial AB5 toxins, B-subunits

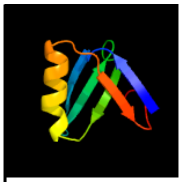
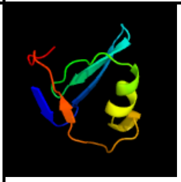
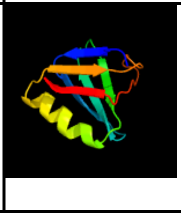
S. mosellana (VUAH1000016.1 [900260-900502])	d2bosa		78.5	31/81	Fold:OB-fold Superfamily:Bacterial enterotoxins Family:Bacterial AB5 toxins, B-subunits
S. mosellana (VUAH1000016.1 [896355-896597])	d2bosa		92.0	27/71	Fold:OB-fold Superfamily:Bacterial enterotoxins Family:Bacterial AB5 toxins, B-subunits
APSE-5 (ACJ10077.1)	d1c4ga		84.3	30/73	Fold:OB-fold Superfamily:Bacterial enterotoxins Family:Bacterial AB5 toxins, B-subunits
Bacterium associated with poplar plant (RYX79668.1)	d1c4ga		96.5	30/85	Fold:OB-fold Superfamily:Bacterial enterotoxins Family:Bacterial AB5 toxins, B-subunits

Table S6. δ values for gene phylogenies demonstrate that there is a relationship between ecological habitat and horizontal gene transfer. δ values for both complete trees and trees for which we removed vertical descendance ("HGT-Only") are shown. P-value is calculated as the number of simulations (*n*=100) in which the shuffled δ is higher than the realized δ.

	Real Phylogeny			**HGT-Only Phylogeny**		
	Tips	**δ**	**p-value**	**Tips**	**δ**	**p-value**
Aip56	90	9.55	<0.01*	63	16.8	0.01*
SltxB	28	6.86	<0.01*	11	35.7	0.07
CdtB	76	10.3	<0.01*	41	5.92	<0.01*
RHS	188	4.85	<0.01*	109	3.98	<0.01*
Lysozyme	172	13.2	<0.01*	155	13.4	<0.01*

Table S7. Accession numbers for marker genes included on Cecidomyiidae species tree.

Species	**Genome**	**Transcriptome**	**CO1 accession #**	**CAD accession #**	**EF1a accession #**	**28S accession #**
*Bibio marci**	NA	SRX314826	KT316846.1	KX453730.1	NA	KJ136761.1
Catotricha subobsoleta	GCA_011634745.1	NA	KT316873.1	KX453747.1	MG684878.1	KP288821.1
C. nasturtii	GCA_009176525.2	SRX6853821	EU812560.1	XM_031771980.1	XM_031764621.1	NA
M. destructor	GCA_000149185.1	SRX516926	EU375697.1	FJ040625.1	AF085227.1	FJ040514.1
S. mosellana	GCA_009176505.1	SRX252524	KC769209.1	MN191486.1	VUAH01006101.1	MN201531.1

*Outgroup.

Table S8. Accession IDs for protein sequences used in alignments and structure prediction.

Taxon	CdtB	Lysozyme	RHS	SltxB
Metazoa				
C. nasturtii	XP_031641203.1, XP_031639861.1	XP_031638744, NW2201606 [2605394,2605837], NW02201606 [2607199,2606738]	–	XP_031619577.1, XP_031619578.1
M. destructor		AEG01007297 [120-560], AEG01024319 [3788,4225], GL501497 [230410,230853]	GL501425[246422,251816]	
S. mosellana			VUAH01006948[8,4950]	UAH1000016.1 [900260,900502], [896355,896597], VUAH01000060.1 [425037,425252], [423461,423673], VUAH01000016.1 [899256,899522]
D. ananassae	XP_014760894.1			
D. biarmipes	XP_016950904.1			
Scaptomyza flava	QDH44045.1			
Myzus persicae	XP_022163116.1			
Fungi				
Fusarium decemcellulare		KAF5009760.1)		
Coprinopsis cinerea		XP_001840847.1		
Prokaryote				
Enterobacteriophage T4		ENLYS_BPT4		
Enterobacteriaphage H19				STXB_BPH19

Staphylococcus phage SA1		D2K0A1		
Bacteriophage APSE1				NP_050968.1
Bacteriophage APSE2	AGX01517.1	YP_002308525.1		
Bacteriophage APSE3			CAB3775397.1	
Bacteriophage APSE5		ACJ10082.1		ACJ10077.1
Escherichia coli	Q46669.1		AAB18570.1	EF|6751557.1
Poplar bacterium				RYX79668.1
Haemophilus ducreyi	AKO43951.1			
Vibrio cholerae			Q9KS45.1	
Salmonella enterica			ECF7068452.1	
Bacillus subtilis			WP_192858111.1	
Yersinia entomophaga			ABG33864.1	
Xenorhabdus vietnamensis			WP_086110724.1	
Outgroups				
Bovine DNAse I	P00639.3			

Fig S1. Complete Maximum Likelihood protein phylogenies for five genes transferred to the Cecidomyiidae lineage reveal possible donor provenance. Phylogenies show taxonomic and lifestyle information for each tip. HTG clades discussed in the manuscript are highlighted in blue. Branch bootstrap values are shown as darkness of nodes and are based on 1000 ultrafast bootstrap replicates. **Figs S1a-e** show, respectively: Aip56, CdtB, Lysozyme, RHS, and SltxB. For more information about the phylogenies, see **Supplementary Methods**.

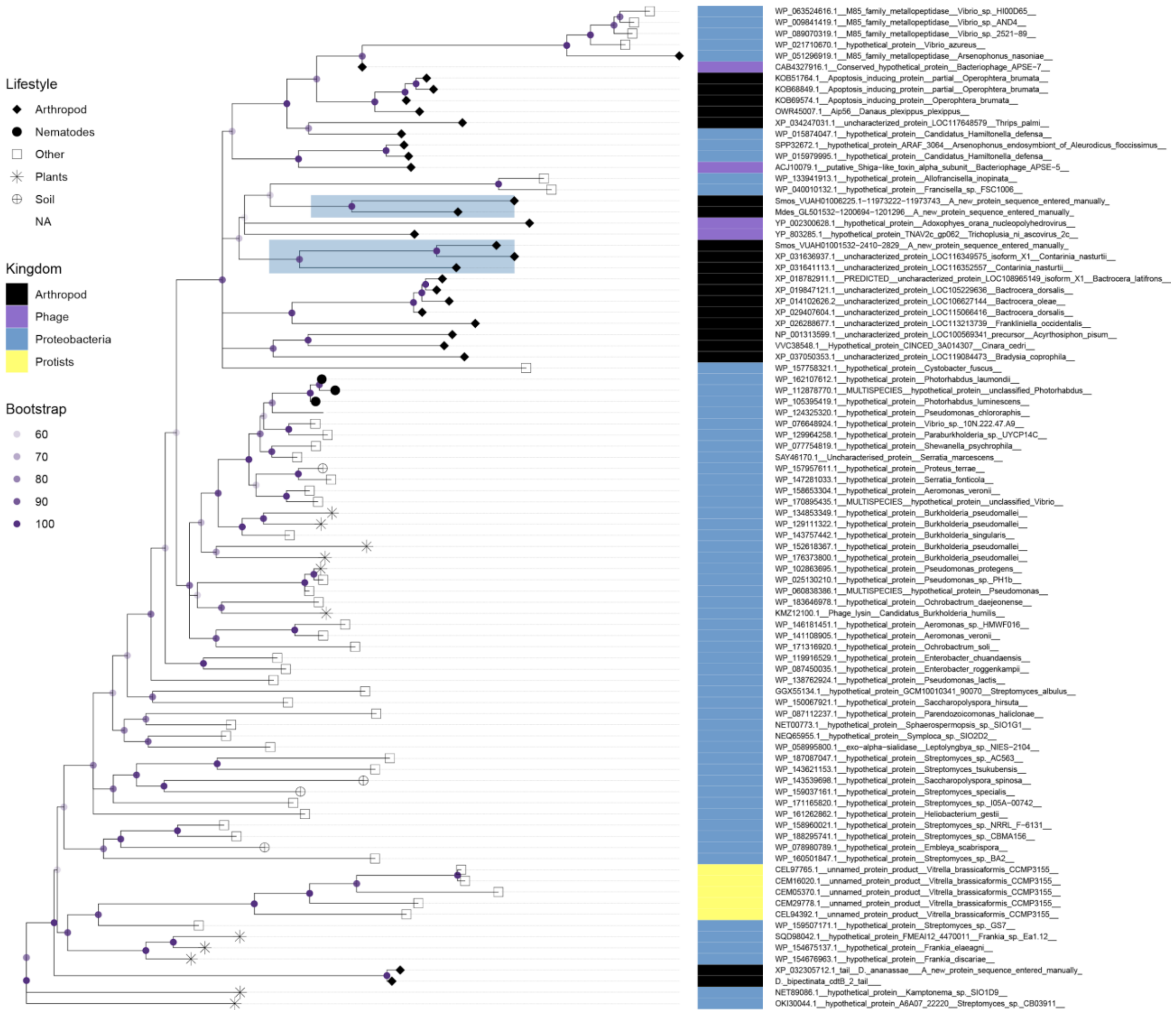

Lifestyle
◆ Arthropod
● Nematodes
□ Other
✳ Plants
⊕ Soil
NA
Kingdom
■ Arthropod
■ Phage
■ Proteobacteria
■ Protists
Bootstrap
● 60
● 70
● 80
● 90
● 100
WP_063524616.1__M85_family_metallopeptidase__Vibrio_sp._HI00D65__
WP_009841419.1__M85_family_metallopeptidase__Vibrio_sp._AND4__
WP_089070319.1__M85_family_metallopeptidase__Vibrio_sp._2521-89__
WP_021710670.1__hypothetical_protein__Vibrio_azureus__
WP_051296919.1__M85_family_metallopeptidase__Arsenophonus_nasoniae__
CAB4327916.1__Conserved_hypothetical_protein__Bacteriophage_APSE-7__
KOB51764.1__Apoptosis_inducing_protein__partial__Operophtera_brumata__
KOB68849.1__Apoptosis_inducing_protein__partial__Operophtera_brumata__
KOB69574.1__Apoptosis_inducing_protein__Operophtera_brumata__
OWR45007.1__Aip56__Danaus_plexippus_plexippus__
XP_034247031.1__uncharacterized_protein_LOC117648579__Thrips_palmi__
WP_015874047.1__hypothetical_protein__Candidatus_Hamiltonella_defensa__
SPP32672.1__hypothetical_protein_ARAF_3064__Arsenophonus_endosymbiont_of_Aleurodicus_floccissimus__
WP_015979995.1__hypothetical_protein__Candidatus_Hamiltonella_defensa__
ACJ10079.1__putative_Shiga-like_toxin_alpha_subunit__Bacteriophage_APSE-5__
WP_133941913.1__hypothetical_protein__Allofrancisella_inopinata__
WP_040010132.1__hypothetical_protein__Francisella_sp._FSC1006__
Smos_VUAH01006225.1-11973222-11973743__A_new_protein_sequence_entered_manually_
Mdes_GL501532-1200694-1201296__A_new_protein_sequence_entered_manually_
YP_002300628.1__hypothetical_protein__Adoxophyes_orana_nucleopolyhedrovirus__
YP_803285.1__hypothetical_protein_TNAV2c_gp062__Trichoplusia_ni_ascovirus_2c__
Smos_VUAH01001532-2410-2829__A_new_protein_sequence_entered_manually_
XP_031636937.1__uncharacterized_protein_LOC116349575_isoform_X1__Contarinia_nasturtii__
XP_031641113.1__uncharacterized_protein_LOC116352557__Contarinia_nasturtii__
XP_018782911.1__PREDICTED__uncharacterized_protein_LOC108965149_isoform_X1__Bactrocera_latifrons__
XP_019847121.1__uncharacterized_protein_LOC105229636__Bactrocera_dorsalis__
XP_014102626.2__uncharacterized_protein_LOC106627144__Bactrocera_oleae__
XP_029407604.1__uncharacterized_protein_LOC115066416__Bactrocera_dorsalis__
XP_026288677.1__uncharacterized_protein_LOC113213739__Frankliniella_occidentalis__
NP_001313599.1__uncharacterized_protein_LOC100569341_precursor__Acyrthosiphon_pisum__
VVC38548.1__Hypothetical_protein_CINCED_3A014307__Cinara_cedri__
XP_037050353.1__uncharacterized_protein_LOC119084473__Bradysia_coprophila__
WP_157758321.1__hypothetical_protein__Cystobacter_fuscus__
WP_162107612.1__hypothetical_protein__Photorhabdus_laumondii__
WP_112878770.1__MULTISPECIES__hypothetical_protein__unclassified_Photorhabdus__
WP_105395419.1__hypothetical_protein__Photorhabdus_luminescens__
WP_124325320.1__hypothetical_protein__Pseudomonas_chlororaphis__
WP_076648924.1__hypothetical_protein__Vibrio_sp._10N.222.47.A9__
WP_129964258.1__hypothetical_protein__Paraburkholderia_sp._UYCP14C__
WP_077754819.1__hypothetical_protein__Shewanella_psychrophila__
SAY46170.1__Uncharacterised_protein__Serratia_marcescens__
WP_157957611.1__hypothetical_protein__Proteus_terrae__
WP_147281033.1__hypothetical_protein__Serratia_fonticola__
WP_158653304.1__hypothetical_protein__Aeromonas_veronii__
WP_170895435.1__MULTISPECIES__hypothetical_protein__unclassified_Vibrio__
WP_134853349.1__hypothetical_protein__Burkholderia_pseudomallei__
WP_129111322.1__hypothetical_protein__Burkholderia_pseudomallei__
WP_143757442.1__hypothetical_protein__Burkholderia_singularis__
WP_152618367.1__hypothetical_protein__Burkholderia_pseudomallei__
WP_176373800.1__hypothetical_protein__Burkholderia_pseudomallei__
WP_102863695.1__hypothetical_protein__Pseudomonas_protegens__
WP_025130210.1__hypothetical_protein__Pseudomonas_sp._PH1b__
WP_060838386.1__MULTISPECIES__hypothetical_protein__Pseudomonas__
WP_183646978.1__hypothetical_protein__Ochrobactrum_daejeonense__
KMZ12100.1__Phage_lysin__Candidatus_Burkholderia_humilis__
WP_146181451.1__hypothetical_protein__Aeromonas_sp._HMWF016__
WP_141108905.1__hypothetical_protein__Aeromonas_veronii__
WP_171316920.1__hypothetical_protein__Ochrobactrum_soli__
WP_119916529.1__hypothetical_protein__Enterobacter_chuandaensis__
WP_087450035.1__hypothetical_protein__Enterobacter_roggenkampii__
WP_138762924.1__hypothetical_protein__Pseudomonas_lactis__
GGX55134.1__hypothetical_protein_GCM10010341_90070__Streptomyces_albulus__
WP_150067921.1__hypothetical_protein__Saccharopolyspora_hirsuta__
WP_087112237.1__hypothetical_protein__Parendozoicomonas_haliclonae__
NET00773.1__hypothetical_protein__Sphaerospermopsis_sp._SIO1G1__
NEQ65955.1__hypothetical_protein__Symploca_sp._SIO2D2__
WP_058995800.1__exo-alpha-sialidase__Leptolyngbya_sp._NIES-2104__
WP_187087047.1__hypothetical_protein__Streptomyces_sp._AC563__
WP_143621153.1__hypothetical_protein__Streptomyces_tsukubensis__
WP_143539698.1__hypothetical_protein__Saccharopolyspora_spinosa__
WP_159037161.1__hypothetical_protein__Streptomyces_specialis__
WP_171165820.1__hypothetical_protein__Streptomyces_sp._I05A-00742__
WP_161262862.1__hypothetical_protein__Heliobacterium_gestii__
WP_158960021.1__hypothetical_protein__Streptomyces_sp._NRRL_F-6131__
WP_188295741.1__hypothetical_protein__Streptomyces_sp._CBMA156__
WP_078980789.1__hypothetical_protein__Embleya_scabrispora__
WP_160501847.1__hypothetical_protein__Streptomyces_sp._BA2__
CEL97765.1__unnamed_protein_product__Vitrella_brassicaformis_CCMP3155__
CEM16020.1__unnamed_protein_product__Vitrella_brassicaformis_CCMP3155__
CEM05370.1__unnamed_protein_product__Vitrella_brassicaformis_CCMP3155__
CEM29778.1__unnamed_protein_product__Vitrella_brassicaformis_CCMP3155__
CEL94392.1__unnamed_protein_product__Vitrella_brassicaformis_CCMP3155__
WP_159507171.1__hypothetical_protein__Streptomyces_sp._GS7__
SQD98042.1__hypothetical_protein_FMEAI12_4470011__Frankia_sp._Ea1.12__
WP_154675137.1__hypothetical_protein__Frankia_elaeagni__
WP_154676963.1__hypothetical_protein__Frankia_discariae__
XP_032305712.1_tail__D._ananassae___A_new_protein_sequence_entered_manually_
D._bipectinata_cdtB_2_tail___
NET89086.1__hypothetical_protein__Kamptonema_sp._SIO1D9__
OKI30044.1__hypothetical_protein_A6A07_22220__Streptomyces_sp._CB03911__

Fig. S1a. Aip56 protein phylogeny.

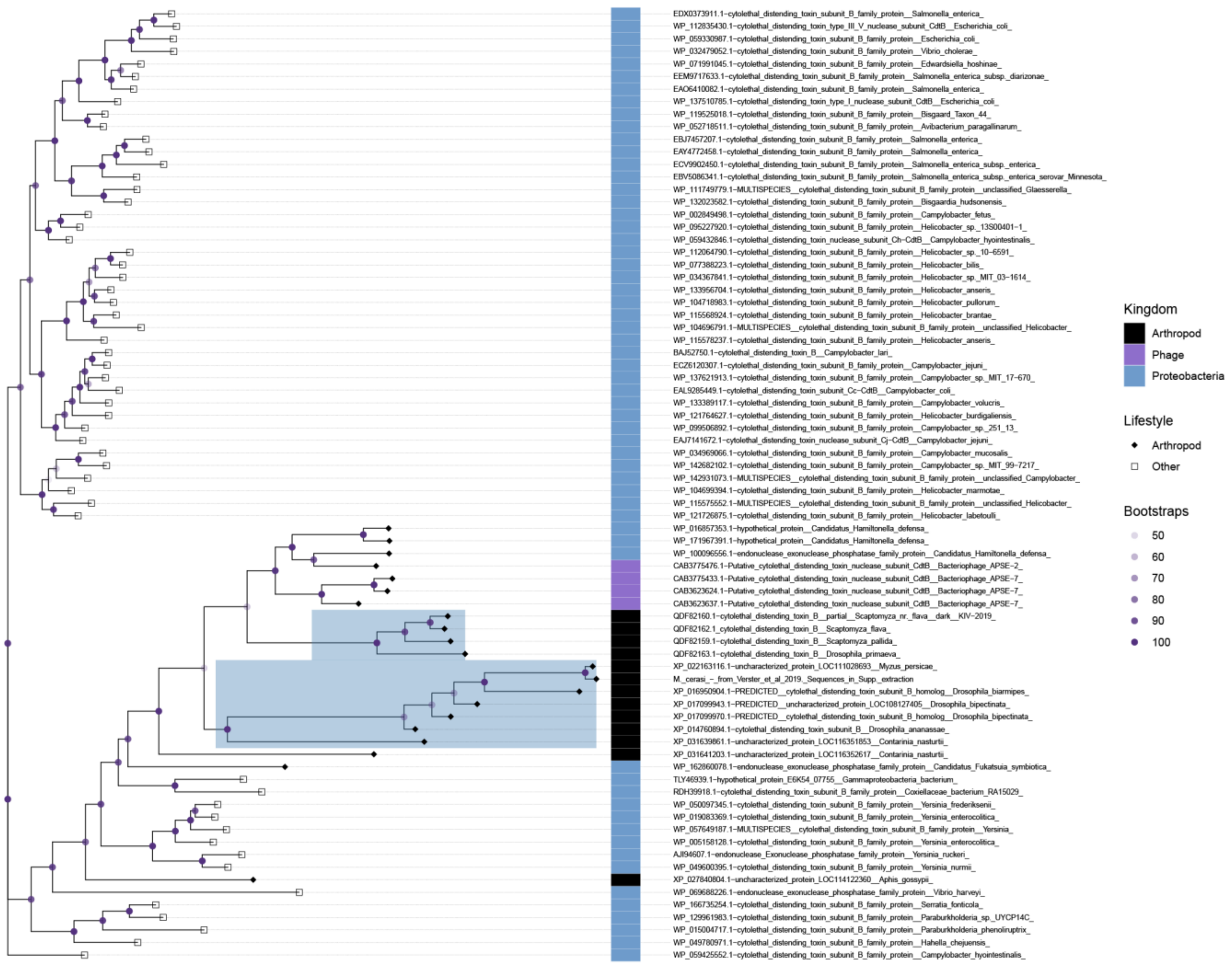

EDX0373911.1−cytolethal_distending_toxin_subunit_B_family_protein__Salmonella_enterica_
WP_112835430.1−cytolethal_distending_toxin_type_III_V_nuclease_subunit_CdtB__Escherichia_coli_
WP_059330987.1−cytolethal_distending_toxin_subunit_B_family_protein__Escherichia_coli_
WP_032479052.1−cytolethal_distending_toxin_subunit_B_family_protein__Vibrio_cholerae_
WP_071991045.1−cytolethal_distending_toxin_subunit_B_family_protein__Edwardsiella_hoshinae_
EEM9717633.1−cytolethal_distending_toxin_subunit_B_family_protein__Salmonella_enterica_subsp._diarizonae_
EAO6410082.1−cytolethal_distending_toxin_subunit_B_family_protein__Salmonella_enterica_
WP_137510785.1−cytolethal_distending_toxin_type_I_nuclease_subunit_CdtB__Escherichia_coli_
WP_119525018.1−cytolethal_distending_toxin_subunit_B_family_protein__Bisgaard_Taxon_44_
WP_052718511.1−cytolethal_distending_toxin_subunit_B_family_protein__Avibacterium_paragallinarum_
EBJ7457207.1−cytolethal_distending_toxin_subunit_B_family_protein__Salmonella_enterica_
EAY4772458.1−cytolethal_distending_toxin_subunit_B_family_protein__Salmonella_enterica_
ECV9902450.1−cytolethal_distending_toxin_subunit_B_family_protein__Salmonella_enterica_subsp._enterica_
EBV5086341.1−cytolethal_distending_toxin_subunit_B_family_protein__Salmonella_enterica_subsp._enterica_serovar_Minnesota_
WP_111749779.1−MULTISPECIES__cytolethal_distending_toxin_subunit_B_family_protein__unclassified_Glaesserella_
WP_132023582.1−cytolethal_distending_toxin_subunit_B_family_protein__Bisgaardia_hudsonensis_
WP_002849498.1−cytolethal_distending_toxin_subunit_B_family_protein__Campylobacter_fetus_
WP_095227920.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_sp._13S00401−1_
WP_059432846.1−cytolethal_distending_toxin_nuclease_subunit_Ch−CdtB__Campylobacter_hyointestinalis_
WP_112064790.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_sp._10−6591_
WP_077388223.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_bilis_
WP_034367841.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_sp._MIT_03−1614_
WP_133956704.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_anseris_
WP_104718983.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_pullorum_
WP_115568924.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_brantae_
WP_104696791.1−MULTISPECIES__cytolethal_distending_toxin_subunit_B_family_protein__unclassified_Helicobacter_
WP_115578237.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_anseris_
BAJ52750.1−cytolethal_distending_toxin_B__Campylobacter_lari_
ECZ6120307.1−cytolethal_distending_toxin_subunit_B_family_protein__Campylobacter_jejuni_
WP_137621913.1−cytolethal_distending_toxin_subunit_B_family_protein__Campylobacter_sp._MIT_17−670_
EAL9285449.1−cytolethal_distending_toxin_subunit_Cc−CdtB__Campylobacter_coli_
WP_133389117.1−cytolethal_distending_toxin_subunit_B_family_protein__Campylobacter_volucris_
WP_121764627.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_burdigaliensis_
WP_099506892.1−cytolethal_distending_toxin_subunit_B_family_protein__Campylobacter_sp._251_13_
EAJ7141672.1−cytolethal_distending_toxin_nuclease_subunit_Cj−CdtB__Campylobacter_jejuni_
WP_034969066.1−cytolethal_distending_toxin_subunit_B_family_protein__Campylobacter_mucosalis_
WP_142682102.1−cytolethal_distending_toxin_subunit_B_family_protein__Campylobacter_sp._MIT_99−7217_
WP_142931073.1−MULTISPECIES__cytolethal_distending_toxin_subunit_B_family_protein__unclassified_Campylobacter_
WP_104699394.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_marmotae_
WP_115575552.1−MULTISPECIES__cytolethal_distending_toxin_subunit_B_family_protein__unclassified_Helicobacter_
WP_121726875.1−cytolethal_distending_toxin_subunit_B_family_protein__Helicobacter_labetoulli_
WP_016857353.1−hypothetical_protein__Candidatus_Hamiltonella_defensa_
WP_171967391.1−hypothetical_protein__Candidatus_Hamiltonella_defensa_
WP_100096556.1−endonuclease_exonuclease_phosphatase_family_protein__Candidatus_Hamiltonella_defensa_
CAB3775476.1−Putative_cytolethal_distending_toxin_nuclease_subunit_CdtB__Bacteriophage_APSE−2_
CAB3775433.1−Putative_cytolethal_distending_toxin_nuclease_subunit_CdtB__Bacteriophage_APSE−7_
CAB3623624.1−Putative_cytolethal_distending_toxin_nuclease_subunit_CdtB__Bacteriophage_APSE−7_
CAB3623637.1−Putative_cytolethal_distending_toxin_nuclease_subunit_CdtB__Bacteriophage_APSE−7_
QDF82160.1−cytolethal_distending_toxin_B__partial__Scaptomyza_nr._flava__dark__KIV−2019_
QDF82162.1_cytolethal_distending_toxin_B__Scaptomyza_flava_
QDF82159.1−cytolethal_distending_toxin_B__Scaptomyza_pallida_
QDF82163.1−cytolethal_distending_toxin_B__Drosophila_primaeva_
XP_022163116.1−uncharacterized_protein_LOC111028693__Myzus_persicae_
M._cerasi_−_from_Verster_et_al_2019._Sequences_in_Supp._extraction
XP_016950904.1−PREDICTED__cytolethal_distending_toxin_subunit_B_homolog__Drosophila_biarmipes_
XP_017099943.1−PREDICTED__uncharacterized_protein_LOC108127405__Drosophila_bipectinata_
XP_017099970.1−PREDICTED__cytolethal_distending_toxin_subunit_B_homolog__Drosophila_bipectinata_
XP_014760894.1−cytolethal_distending_toxin_subunit_B__Drosophila_ananassae_
XP_031639861.1−uncharacterized_protein_LOC116351853__Contarinia_nasturtii_
XP_031641203.1−uncharacterized_protein_LOC116352617__Contarinia_nasturtii_
WP_162860078.1−endonuclease_exonuclease_phosphatase_family_protein__Candidatus_Fukatsuia_symbiotica_
TLY46939.1−hypothetical_protein_E6K54_07755__Gammaproteobacteria_bacterium_
RDH39918.1−cytolethal_distending_toxin_subunit_B_family_protein__Coxiellaceae_bacterium_RA15029_
WP_050097345.1−cytolethal_distending_toxin_subunit_B_family_protein__Yersinia_frederiksenii_
WP_019083369.1−cytolethal_distending_toxin_subunit_B_family_protein__Yersinia_enterocolitica_
WP_057649187.1−MULTISPECIES__cytolethal_distending_toxin_subunit_B_family_protein__Yersinia_
WP_005158128.1−cytolethal_distending_toxin_subunit_B_family_protein__Yersinia_enterocolitica_
AJI94607.1−endonuclease_Exonuclease_phosphatase_family_protein__Yersinia_ruckeri_
WP_049600395.1−cytolethal_distending_toxin_subunit_B_family_protein__Yersinia_nurmii_
XP_027840804.1−uncharacterized_protein_LOC114122360__Aphis_gossypii_
WP_069688226.1−endonuclease_exonuclease_phosphatase_family_protein__Vibrio_harveyi_
WP_166735254.1−cytolethal_distending_toxin_subunit_B_family_protein__Serratia_fonticola_
WP_129961983.1−cytolethal_distending_toxin_subunit_B_family_protein__Paraburkholderia_sp._UYCP14C_
WP_015004717.1−cytolethal_distending_toxin_subunit_B_family_protein__Paraburkholderia_phenoliruptrix_
WP_049780971.1−cytolethal_distending_toxin_subunit_B_family_protein__Hahella_chejuensis_
WP_059425552.1−cytolethal_distending_toxin_subunit_B_family_protein__Campylobacter_hyointestinalis_
Kingdom
Arthropod
Phage
Proteobacteria
Lifestyle
◆ Arthropod
□ Other
Bootstraps
50
60
70
80
90
100

Fig S1b. CdtB protein phylogeny.

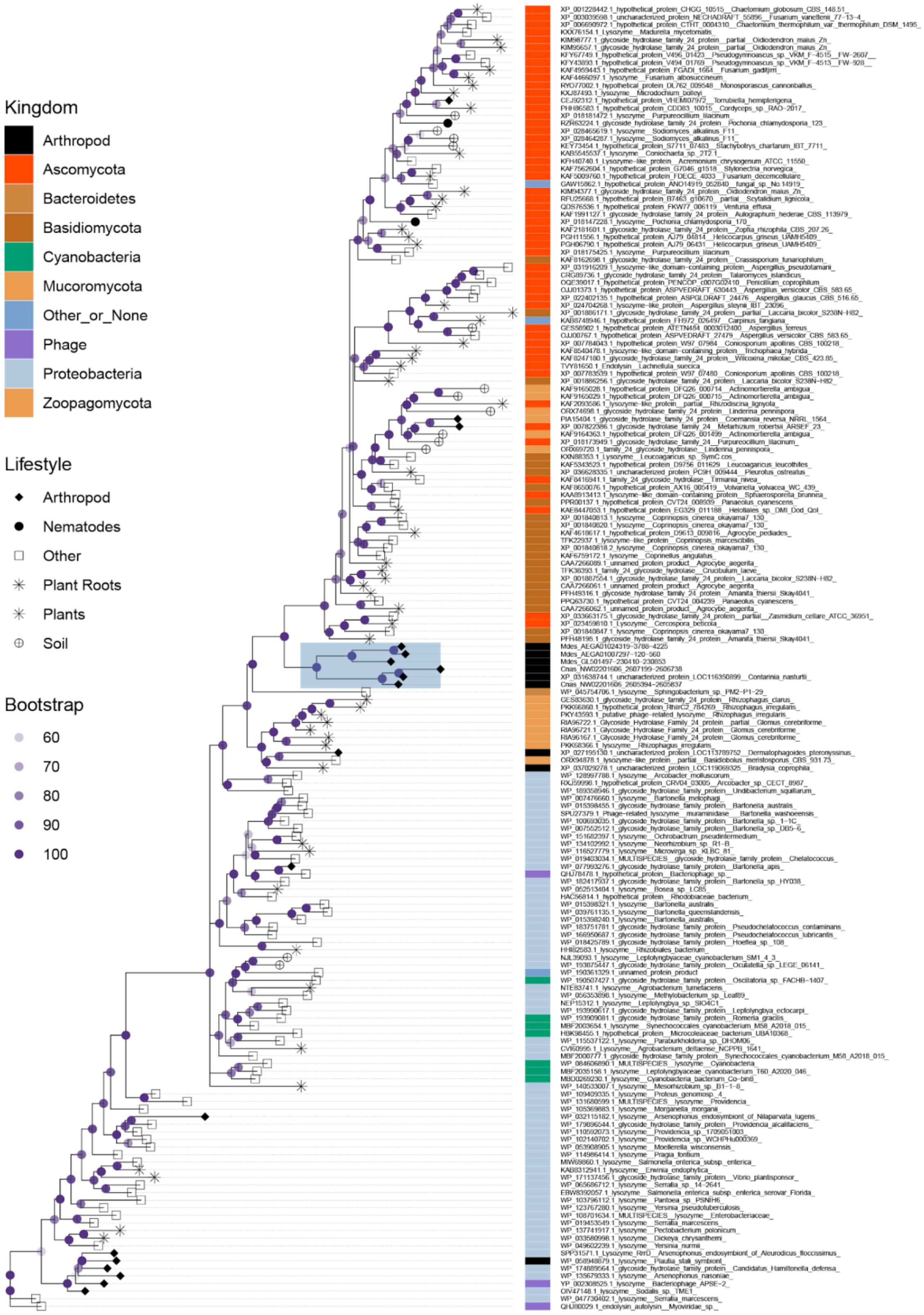

Kingdom
Arthropod
Ascomycota
Bacteroidetes
Basidiomycota
Cyanobacteria
Mucoromycota
Other_or_None
Phage
Proteobacteria
Zoopagomycota
Lifestyle
Arthropod
Nematodes
Other
Plant Roots
Plants
Soil
Bootstrap
60
70
80
90
100
XP_001228442.1_hypothetical_protein_CHGG_10515__Chaetomium_globosum_CBS_148.51_
XP_003039598.1_uncharacterized_protein_NECHADRAFT_55896__Fusarium_vanettenii_77-13-4_
XP_006690972.1_hypothetical_protein_CTHT_0004310__Chaetomium_thermophilum_var._thermophilum_DSM_1495_
KXX76154.1_Lysozyme__Madurella_mycetomatis_
KIM98777.1_glycoside_hydrolase_family_24_protein__partial__Oidiodendron_maius_Zn_
KIM95657.1_glycoside_hydrolase_family_24_protein__partial__Oidiodendron_maius_Zn_
KFY67749.1_hypothetical_protein_V496_01423__Pseudogymnoascus_sp._VKM_F-4515__FW-2607__
KFY43893.1_hypothetical_protein_V494_01769__Pseudogymnoascus_sp._VKM_F-4513__FW-928__
KAF4959443.1_hypothetical_protein_FGADI_1664__Fusarium_gaditjirri_
KAF4466097.1_lysozyme__Fusarium_albosuccineum_
RYO77002.1_hypothetical_protein_DL762_009548__Monosporascus_cannonballus_
KXJ87493.1_lysozyme__Microdochium_bolleyi_
CEJ92312.1_hypothetical_protein_VHEMI07972__Torrubiella_hemipterigena_
PHH86583.1_hypothetical_protein_CDD83_10015__Cordyceps_sp._RAO-2017_
XP_018181472.1_lysozyme__Purpureocillium_lilacinum_
RZR63224.1_glycoside_hydrolase_family_24_protein__Pochonia_chlamydosporia_123_
XP_028465619.1_lysozyme__Sodiomyces_alkalinus_F11_
XP_028464287.1_lysozyme__Sodiomyces_alkalinus_F11_
KEY73454.1_hypothetical_protein_S7711_07483__Stachybotrys_chartarum_IBT_7711_
KAB5545537.1_lysozyme__Coniochaeta_sp._2T2.1_
KFH40740.1_Lysozyme-like_protein__Acremonium_chrysogenum_ATCC_11550_
KAF7562604.1_hypothetical_protein_G7046_g1518__Stylonectria_norvegica_
KAF5009760.1_hypothetical_protein_FDECE_4033__Fusarium_decemcellulare_
GAW15862.1_hypothetical_protein_ANO14919_052840__fungal_sp._No.14919_
KIM94377.1_glycoside_hydrolase_family_24_protein__Oidiodendron_maius_Zn_
RFU25668.1_hypothetical_protein_B7463_g10670__partial__Scytalidium_lignicola_
QDS76536.1_hypothetical_protein_FKW77_006119__Venturia_effusa_
KAF1991127.1_glycoside_hydrolase_family_24_protein__Aulographum_hederae_CBS_113979_
XP_018147228.1_lysozyme__Pochonia_chlamydosporia_170_
KAF2181601.1_glycoside_hydrolase_family_24_protein__Zopfia_rhizophila_CBS_207.26_
PGH11556.1_hypothetical_protein_AJ79_04814__Helicocarpus_griseus_UAMH5409_
PGH06790.1_hypothetical_protein_AJ79_06431__Helicocarpus_griseus_UAMH5409_
XP_018175425.1_lysozyme__Purpureocillium_lilacinum_
KAF8162698.1_glycoside_hydrolase_family_24_protein__Crassisporium_funariophilum_
XP_031916209.1_lysozyme-like_domain-containing_protein__Aspergillus_pseudotamarii_
CRG89736.1_glycoside_hydrolase_family_24_protein__Talaromyces_islandicus_
OQE39017.1_hypothetical_protein_PENCOP_c007G02410__Penicillium_coprophilum_
OJJ01373.1_hypothetical_protein_ASPVEDRAFT_630443__Aspergillus_versicolor_CBS_583.65_
XP_022402135.1_hypothetical_protein_ASPGLDRAFT_24476__Aspergillus_glaucus_CBS_516.65_
XP_024704268.1_lysozyme-like_protein__Aspergillus_steynii_IBT_23096_
XP_001886171.1_glycoside_hydrolase_family_24_protein__partial__Laccaria_bicolor_S238N-H82_
KAB8748946.1_hypothetical_protein_FH972_026497__Carpinus_fangiana_
GES58902.1_hypothetical_protein_ATETN484_0003012400__Aspergillus_terreus_
OJJ00767.1_hypothetical_protein_ASPVEDRAFT_27479__Aspergillus_versicolor_CBS_583.65_
XP_007784043.1_hypothetical_protein_W97_07984__Coniosporium_apollinis_CBS_100218_
KAF8540478.1_lysozyme-like_domain-containing_protein__Trichophaea_hybrida_
KAF8247180.1_glycoside_hydrolase_family_24_protein__Wilcoxina_mikolae_CBS_423.85_
TVY81650.1_Endolysin__Lachnellula_suecica_
XP_007783539.1_hypothetical_protein_W97_07480__Coniosporium_apollinis_CBS_100218_
XP_001886256.1_glycoside_hydrolase_family_24_protein__Laccaria_bicolor_S238N-H82_
KAF9165028.1_hypothetical_protein_DFQ26_000714__Actinomortierella_ambigua_
KAF9165029.1_hypothetical_protein_DFQ26_000715__Actinomortierella_ambigua_
KAF2093586.1_lysozyme-like_protein__partial__Rhizodiscina_lignyota_
ORX74698.1_glycoside_hydrolase_family_24_protein__Linderina_pennispora_
PIA15404.1_glycoside_hydrolase_family_24_protein__Coemansia_reversa_NRRL_1564_
XP_007822386.1_glycoside_hydrolase_family_24__Metarhizium_robertsii_ARSEF_23_
KAF9164363.1_hypothetical_protein_DFQ26_001499__Actinomortierella_ambigua_
XP_018173949.1_glycoside_hydrolase_family_24__Purpureocillium_lilacinum_
ORX69720.1_family_24_glycoside_hydrolase__Linderina_pennispora_
KXN88353.1_Lysozyme__Leucoagaricus_sp._SymC.cos_
KAF5343523.1_hypothetical_protein_D9756_011629__Leucoagaricus_leucothites_
XP_036628335.1_uncharacterized_protein_PC9H_009444__Pleurotus_ostreatus_
KAF8416941.1_family_24_glycoside_hydrolase__Tirmania_nivea_
KAF8650076.1_hypothetical_protein_AX16_005419__Volvariella_volvacea_WC_439_
KAA8913413.1_lysozyme-like_domain-containing_protein__Sphaerosporella_brunnea_
PPR00137.1_hypothetical_protein_CVT24_008939__Panaeolus_cyanescens_
KAE8447053.1_hypothetical_protein_EG329_011188__Helotiales_sp._DMI_Dod_QoI_
XP_001840813.1_lysozyme__Coprinopsis_cinerea_okayama7_130_
XP_001840820.1_lysozyme__Coprinopsis_cinerea_okayama7_130_
KAF4618617.1_hypothetical_protein_D9613_009816__Agrocybe_pediades_
TFK22937.1_lysozyme-like_protein__Coprinopsis_marcescibilis_
XP_001840818.2_lysozyme__Coprinopsis_cinerea_okayama7_130_
KAF6759172.1_lysozyme__Coprinellus_angulatus_
CAA7266089.1_unnamed_protein_product__Agrocybe_aegerita_
TFK38393.1_family_24_glycoside_hydrolase__Crucibulum_laeve_
XP_001887554.1_glycoside_hydrolase_family_24_protein__Laccaria_bicolor_S238N-H82_
CAA7266061.1_unnamed_protein_product__Agrocybe_aegerita_
PFH49316.1_glycoside_hydrolase_family_24_protein__Amanita_thiersii_Skay4041_
PPQ63730.1_hypothetical_protein_CVT24_004239__Panaeolus_cyanescens_
CAA7266062.1_unnamed_protein_product__Agrocybe_aegerita_
XP_033663175.1_glycoside_hydrolase_family_24_protein__partial__Zasmidium_cellare_ATCC_36951_
XP_023459810.1_Lysozyme__Cercospora_beticola_
XP_001840847.1_lysozyme__Coprinopsis_cinerea_okayama7_130_
PFH48195.1_glycoside_hydrolase_family_24_protein__Amanita_thiersii_Skay4041_
Mdes_AEGA01024319-3788-4225
Mdes_AEGA01007297-120-560
Mdes_GL501497-230410-230853
Cnas_NW02201606_2607199-2606738
XP_031638744.1_uncharacterized_protein_LOC116350899__Contarinia_nasturtii_
Cnas_NW02201606_2605394-2605837
WP_045754706.1_lysozyme__Sphingobacterium_sp._PM2-P1-29_
GES83630.1_glycoside_hydrolase_family_24_protein__Rhizophagus_clarus_
PKK66860.1_hypothetical_protein_RhiirC2_784269__Rhizophagus_irregularis_
PKY43593.1_putative_phage-related_lysozyme__Rhizophagus_irregularis_
RIA96722.1_Glycoside_Hydrolase_Family_24_protein__partial__Glomus_cerebriforme_
RIA96721.1_Glycoside_Hydrolase_Family_24_protein__Glomus_cerebriforme_
RIA96167.1_Glycoside_Hydrolase_Family_24_protein__Glomus_cerebriforme_
PKK68366.1_lysozyme__Rhizophagus_irregularis_
XP_027195130.1_uncharacterized_protein_LOC113789752__Dermatophagoides_pteronyssinus_
ORX94878.1_lysozyme-like_protein__partial__Basidiobolus_meristosporus_CBS_931.73_
XP_037029278.1_uncharacterized_protein_LOC119069325__Bradysia_coprophila_
WP_128997788.1_lysozyme__Arcobacter_molluscorum_
RXJ59998.1_hypothetical_protein_CRV04_03005__Arcobacter_sp._CECT_8987_
WP_189358946.1_glycoside_hydrolase_family_protein__Undibacterium_squillarum_
WP_007476660.1_lysozyme__Bartonella_melophagi_
WP_015398455.1_glycoside_hydrolase_family_protein__Bartonella_australis_
SPU27379.1_Phage-related_lysozyme__muraminidase___Bartonella_washoeensis_
WP_100693035.1_glycoside_hydrolase_family_protein__Bartonella_sp._1-1C_
WP_007552512.1_glycoside_hydrolase_family_protein__Bartonella_sp._DB5-6_
WP_151682397.1_lysozyme__Ochrobactrum_pseudintermedium_
WP_134102992.1_lysozyme__Neorhizobium_sp._R1-B_
WP_116527779.1_lysozyme__Microvirga_sp._KLBC_81_
WP_019403034.1_MULTISPECIES__glycoside_hydrolase_family_protein__Chelatococcus_
WP_077993276.1_glycoside_hydrolase_family_protein__Bartonella_apis_
QHJ78478.1_hypothetical_protein__Bacteriophage_sp._
WP_182417937.1_glycoside_hydrolase_family_protein__Bartonella_sp._HY038_
WP_052513404.1_lysozyme__Bosea_sp._LC85_
HAC56814.1_hypothetical_protein__Rhodobiaceae_bacterium_
WP_015398321.1_lysozyme__Bartonella_australis_
WP_039761135.1_lysozyme__Bartonella_queenslandensis_
WP_015398240.1_lysozyme__Bartonella_australis_
WP_183751781.1_glycoside_hydrolase_family_protein__Pseudochelatococcus_contaminans_
WP_166950687.1_glycoside_hydrolase_family_protein__Pseudochelatococcus_lubricantis_
WP_018425789.1_glycoside_hydrolase_family_protein__Hoeflea_sp._108_
HHI82583.1_lysozyme__Rhizobiales_bacterium_
NJL39093.1_lysozyme__Leptolyngbyaceae_cyanobacterium_SM1_4_3_
WP_193875447.1_glycoside_hydrolase_family_protein__Oculatella_sp._LEGE_06141_
WP_190361329.1_unnamed_protein_product
WP_190507427.1_glycoside_hydrolase_family_protein__Oscillatoria_sp._FACHB-1407_
NTE83741.1_lysozyme__Agrobacterium_tumefaciens_
WP_056353898.1_lysozyme__Methylobacterium_sp._Leaf89_
NEP15312.1_lysozyme__Leptolyngbya_sp._SIO4C1_
WP_193990617.1_glycoside_hydrolase_family_protein__Leptolyngbya_ectocarpi_
WP_193909081.1_glycoside_hydrolase_family_protein__Romeria_gracilis_
MBF2003654.1_lysozyme__Synechococcales_cyanobacterium_M58_A2018_015_
HBK98455.1_hypothetical_protein__Microcoleaceae_bacterium_UBA10368_
WP_115537122.1_lysozyme__Paraburkholderia_sp._DHOM06_
CVI60995.1_Lysozyme__Agrobacterium_deltaense_NCPPB_1641_
MBF2000777.1_glycoside_hydrolase_family_protein__Synechococcales_cyanobacterium_M58_A2018_015_
WP_084606890.1_MULTISPECIES__lysozyme__Cyanobacteria_
MBF2035158.1_lysozyme__Leptolyngbyaceae_cyanobacterium_T60_A2020_046_
MBD0269230.1_lysozyme__Cyanobacteria_bacterium_Co-bin8_
WP_140533007.1_lysozyme__Mesorhizobium_sp._B1-1-8_
WP_109409335.1_lysozyme__Proteus_genomosp._4_
WP_131680599.1_MULTISPECIES__lysozyme__Providencia_
WP_105369883.1_lysozyme__Morganella_morganii_
WP_032115182.1_lysozyme__Arsenophonus_endosymbiont_of_Nilaparvata_lugens_
WP_179896544.1_glycoside_hydrolase_family_protein__Providencia_alcalifaciens_
WP_110592073.1_lysozyme__Providencia_sp._1709051003_
WP_102140702.1_lysozyme__Providencia_sp._WCHPHu000369_
WP_053908905.1_lysozyme__Moellerella_wisconsensis_
WP_114986414.1_lysozyme__Pragia_fontium_
MIW69860.1_lysozyme__Salmonella_enterica_subsp._enterica_
KAB8312941.1_lysozyme__Erwinia_endophytica_
WP_171137456.1_glycoside_hydrolase_family_protein__Vibrio_plantisponsor_
WP_065686712.1_lysozyme__Serratia_sp._14-2641_
EBW8392057.1_lysozyme__Salmonella_enterica_subsp._enterica_serovar_Florida_
WP_103796112.1_lysozyme__Pantoea_sp._PSNIH6_
WP_123767280.1_lysozyme__Yersinia_pseudotuberculosis_
WP_108701634.1_MULTISPECIES__lysozyme__Enterobacteriaceae_
WP_019453549.1_lysozyme__Serratia_marcescens_
WP_137741917.1_lysozyme__Pectobacterium_polonicum_
WP_033580998.1_lysozyme__Dickeya_chrysanthemi_
WP_049602239.1_lysozyme__Yersinia_nurmii_
SPP31571.1_Lysozyme_RrrD__Arsenophonus_endosymbiont_of_Aleurodicus_floccissimus_
WP_058948879.1_lysozyme__Plautia_stali_symbiont_
WP_174889564.1_glycoside_hydrolase_family_protein__Candidatus_Hamiltonella_defensa_
WP_135679333.1_lysozyme__Arsenophonus_nasoniae_
YP_002308525.1_lysozyme__Bacteriophage_APSE-2_
OIV47148.1_lysozyme__Sodalis_sp._TME1_
WP_047730402.1_lysozyme__Serratia_marcescens_
QHJ80029.1_endolysin_autolysin__Myoviridae_sp._

Fig S1c. Lysozyme protein phylogeny.

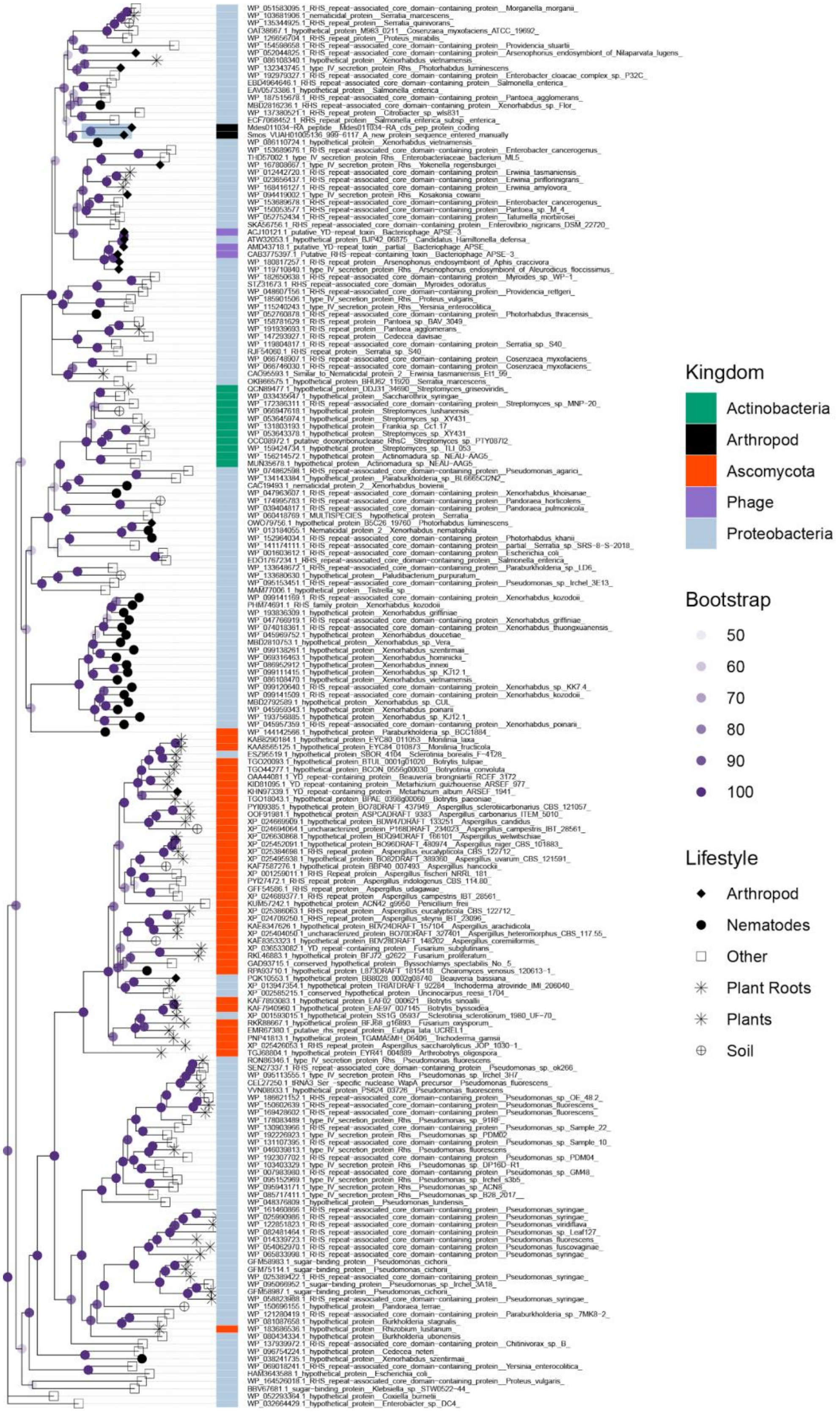

WP_051583095.1_RHS_repeat-associated_core_domain-containing_protein__Morganella_morganii_
WP_103681906.1_nematicidal_protein__Serratia_marcescens_
WP_135344925.1_RHS_repeat_protein__Serratia_quinivorans_
OAT38667.1_hypothetical_protein_M983_0211__Cosenzaea_myxofaciens_ATCC_19692_
WP_126656704.1_RHS_repeat_protein__Proteus_mirabilis_
WP_154598658.1_RHS_repeat-associated_core_domain-containing_protein__Providencia_stuartii_
WP_052044825.1_RHS_repeat-associated_core_domain-containing_protein__Arsenophonus_endosymbiont_of_Nilaparvata_lugens_
WP_086108340.1_hypothetical_protein__Xenorhabdus_vietnamensis_
WP_132343745.1_type_IV_secretion_protein_Rhs__Photorhabdus_luminescens_
WP_192979327.1_RHS_repeat-associated_core_domain-containing_protein__Enterobacter_cloacae_complex_sp._P32C_
EBD4964646.1_RHS_repeat-associated_core_domain-containing_protein__Salmonella_enterica_
EAV0573386.1_hypothetical_protein__Salmonella_enterica_
WP_187515678.1_RHS_repeat-associated_core_domain-containing_protein__Pantoea_agglomerans_
MBD2816236.1_RHS_repeat-associated_core_domain-containing_protein__Xenorhabdus_sp._Flor_
WP_137380521.1_RHS_repeat_protein__Citrobacter_sp._wls831_
ECF7068452.1_RHS_repeat_protein__Salmonella_enterica_subsp._enterica_
Mdes011034-RA_peptide__Mdes011034-RA_cds_pep_protein_coding
Smos_VUAH01005136_999-6117_A_new_protein_sequence_entered_manually
WP_086110724.1_hypothetical_protein__Xenorhabdus_vietnamensis_
WP_153689676.1_RHS_repeat-associated_core_domain-containing_protein__Enterobacter_cancerogenus_
THD57002.1_type_IV_secretion_protein_Rhs__Enterobacteriaceae_bacterium_ML5_
WP_167808667.1_type_IV_secretion_protein_Rhs__Yokenella_regensburgei_
WP_012442720.1_RHS_repeat-associated_core_domain-containing_protein__Erwinia_tasmaniensis_
WP_023656437.1_RHS_repeat-associated_core_domain-containing_protein__Erwinia_piriflorinigrans_
WP_168416127.1_RHS_repeat-associated_core_domain-containing_protein__Erwinia_amylovora_
WP_094419002.1_type_IV_secretion_protein_Rhs__Kosakonia_cowanii_
WP_153689678.1_RHS_repeat-associated_core_domain-containing_protein__Enterobacter_cancerogenus_
WP_150053577.1_RHS_repeat-associated_core_domain-containing_protein__Pantoea_sp._M_4_
WP_052752434.1_RHS_repeat-associated_core_domain-containing_protein__Tatumella_morbirosei_
SKA56756.1_RHS_repeat-associated_core_domain-containing_protein__Enterovibrio_nigricans_DSM_22720_
ACJ10121.1_putative_YD-repeat_toxin__Bacteriophage_APSE-3_
ATW32053.1_hypothetical_protein_BJP42_06875__Candidatus_Hamiltonella_defensa_
AMD43718.1_putative_YD-repeat_toxin__partial__Bacteriophage_APSE_
CAB3775397.1_Putative_RHS-repeat-containing_toxin__Bacteriophage_APSE-3_
WP_180817257.1_RHS_repeat_protein__Arsenophonus_endosymbiont_of_Aphis_craccivora_
WP_119710840.1_type_IV_secretion_protein_Rhs__Arsenophonus_endosymbiont_of_Aleurodicus_floccissimus_
WP_182650638.1_RHS_repeat-associated_core_domain-containing_protein__Myroides_sp._WP-1_
STZ31673.1_RHS_repeat-associated_core_domain__Myroides_odoratus_
WP_048607156.1_RHS_repeat-associated_core_domain-containing_protein__Providencia_rettgeri_
WP_185901506.1_type_IV_secretion_protein_Rhs__Proteus_vulgaris_
WP_115240243.1_type_IV_secretion_protein_Rhs__Yersinia_enterocolitica_
WP_052760878.1_RHS_repeat-associated_core_domain-containing_protein__Photorhabdus_thracensis_
WP_158781629.1_RHS_repeat_protein__Pantoea_sp._BAV_3049_
WP_191939693.1_RHS_repeat_protein__Pantoea_agglomerans_
WP_147293927.1_RHS_repeat_protein__Cedecea_davisae_
WP_119804817.1_RHS_repeat-associated_core_domain-containing_protein__Serratia_sp._S40_
RJF54060.1_RHS_repeat_protein__Serratia_sp._S40_
WP_066748907.1_RHS_repeat-associated_core_domain-containing_protein__Cosenzaea_myxofaciens_
WP_066746030.1_RHS_repeat-associated_core_domain-containing_protein__Cosenzaea_myxofaciens_
CAO95593.1_Similar_to_Nematicidal_protein_2__Erwinia_tasmaniensis_Et1_99_
OKB66575.1_hypothetical_protein_BHU62_11920__Serratia_marcescens_
QCN89477.1_hypothetical_protein_DDJ31_34690__Streptomyces_griseoviridis_
WP_033435647.1_hypothetical_protein__Saccharothrix_syringae_
WP_172386311.1_RHS_repeat-associated_core_domain-containing_protein__Streptomyces_sp._MNP-20_
WP_066947618.1_hypothetical_protein__Streptomyces_lushanensis_
WP_053645974.1_hypothetical_protein__Streptomyces_sp._XY431_
WP_131803193.1_hypothetical_protein__Frankia_sp._Cc1.17_
WP_053643378.1_hypothetical_protein__Streptomyces_sp._XY431_
OCC08972.1_putative_deoxyribonuclease_RhsC__Streptomyces_sp._PTY087I2_
WP_159424734.1_hypothetical_protein__Streptomyces_sp._TLI_053_
WP_156214572.1_hypothetical_protein__Actinomadura_sp._NEAU-AAG5_
MUN35678.1_hypothetical_protein__Actinomadura_sp._NEAU-AAG5_
WP_074862598.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_agarici_
WP_134143384.1_hypothetical_protein__Paraburkholderia_sp._BL6665CI2N2_
CAC19493.1_nematicidal_protein_2__Xenorhabdus_bovienii_
WP_047963607.1_RHS_repeat-associated_core_domain-containing_protein__Xenorhabdus_khoisanae_
WP_174995783.1_RHS_repeat-associated_core_domain-containing_protein__Pandoraea_horticolens_
WP_039404817.1_RHS_repeat-associated_core_domain-containing_protein__Pandoraea_pulmonicola_
WP_060418769.1_MULTISPECIES__hypothetical_protein__Serratia_
OWO79756.1_hypothetical_protein_B5C26_19760__Photorhabdus_luminescens_
WP_013184055.1_Nematicidal_protein_2__Xenorhabdus_nematophila_
WP_152964034.1_RHS_repeat-associated_core_domain-containing_protein__Photorhabdus_khanii_
WP_141174111.1_RHS_repeat-associated_core_domain-containing_protein__partial__Serratia_sp._SRS-8-S-2018_
WP_001603612.1_RHS_repeat-associated_core_domain-containing_protein__Escherichia_coli_
EDO1767234.1_RHS_repeat-associated_core_domain-containing_protein__Salmonella_enterica_
WP_133648672.1_RHS_repeat-associated_core_domain-containing_protein__Paraburkholderia_sp._LD6_
WP_133680630.1_hypothetical_protein__Paludibacterium_purpuratum_
WP_095153451.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_sp._Irchel_3E13_
MAM77006.1_hypothetical_protein__Tistrella_sp._
WP_099141169.1_RHS_repeat-associated_core_domain-containing_protein__Xenorhabdus_kozodoii_
PHM74691.1_RHS_family_protein__Xenorhabdus_kozodoii_
WP_193836309.1_hypothetical_protein__Xenorhabdus_griffiniae_
WP_047766919.1_RHS_repeat-associated_core_domain-containing_protein__Xenorhabdus_griffiniae_
WP_074018361.1_RHS_repeat-associated_core_domain-containing_protein__Xenorhabdus_thuongxuanensis_
WP_045969752.1_hypothetical_protein__Xenorhabdus_doucetiae_
MBD2810753.1_hypothetical_protein__Xenorhabdus_sp._Vera_
WP_099138261.1_hypothetical_protein__Xenorhabdus_szentirmaii_
WP_069316463.1_hypothetical_protein__Xenorhabdus_hominickii_
WP_086952912.1_hypothetical_protein__Xenorhabdus_innexi_
WP_099111415.1_hypothetical_protein__Xenorhabdus_sp._KJ12.1_
WP_086108470.1_hypothetical_protein__Xenorhabdus_vietnamensis_
WP_099120640.1_RHS_repeat-associated_core_domain-containing_protein__Xenorhabdus_sp._KK7.4_
WP_099141509.1_RHS_repeat-associated_core_domain-containing_protein__Xenorhabdus_kozodoii_
MBD2792589.1_hypothetical_protein__Xenorhabdus_sp._CUL_
WP_045959343.1_hypothetical_protein__Xenorhabdus_poinarii_
WP_193756885.1_hypothetical_protein__Xenorhabdus_sp._KJ12.1_
WP_045957359.1_RHS_repeat-associated_core_domain-containing_protein__Xenorhabdus_poinarii_
WP_144142566.1_hypothetical_protein__Paraburkholderia_sp._BCC1884_
KAB8290184.1_hypothetical_protein_EYC80_011053__Monilinia_laxa_
KAA8565125.1_hypothetical_protein_EYC84_010873__Monilinia_fructicola_
ESZ95519.1_hypothetical_protein_SBOR_4104__Sclerotinia_borealis_F-4128_
TGO20093.1_hypothetical_protein_BTUL_0001g01020__Botrytis_tulipae_
TGO44277.1_hypothetical_protein_BCON_0556g00030__Botryotinia_convoluta_
OAA44081.1_YD_repeat-containing_protein__Beauveria_brongniartii_RCEF_3172_
KID81095.1_YD_repeat-containing_protein__Metarhizium_guizhouense_ARSEF_977_
KHN97339.1_YD_repeat-containing_protein__Metarhizium_album_ARSEF_1941_
TGO18043.1_hypothetical_protein_BPAE_0398g00060__Botrytis_paeoniae_
PYI09385.1_hypothetical_protein_BO78DRAFT_437949__Aspergillus_sclerotiicarbonarius_CBS_121057_
OOF91981.1_hypothetical_protein_ASPCADRAFT_9383__Aspergillus_carbonarius_ITEM_5010_
XP_024669909.1_hypothetical_protein_BDW47DRAFT_133251__Aspergillus_candidus_
XP_024694064.1_uncharacterized_protein_P168DRAFT_234023__Aspergillus_campestris_IBT_28561_
XP_026630868.1_hypothetical_protein_BDQ94DRAFT_166101__Aspergillus_welwitschiae_
XP_025452091.1_hypothetical_protein_BO96DRAFT_480974__Aspergillus_niger_CBS_101883_
XP_025384698.1_RHS_repeat_protein__Aspergillus_eucalypticola_CBS_122712_
XP_025495938.1_hypothetical_protein_BO82DRAFT_389360__Aspergillus_uvarum_CBS_121591_
KAF7587276.1_hypothetical_protein_BBP40_007493__Aspergillus_hancockii_
XP_001259011.1_RHS_Repeat_protein__Aspergillus_fischeri_NRRL_181_
PYI27472.1_RHS_repeat_protein__Aspergillus_indologenus_CBS_114.80_
GFF54586.1_RHS_repeat_protein__Aspergillus_udagawae_
XP_024689377.1_RHS_repeat_protein__Aspergillus_campestris_IBT_28561_
KUM57242.1_hypothetical_protein_ACN42_g9950__Penicillium_freii_
XP_025386063.1_RHS_repeat_protein__Aspergillus_eucalypticola_CBS_122712_
XP_024709250.1_RHS_repeat_protein__Aspergillus_steynii_IBT_23096_
KAE8347626.1_hypothetical_protein_BDV24DRAFT_157104__Aspergillus_arachidicola_
XP_025404050.1_uncharacterized_protein_BO70DRAFT_327401__Aspergillus_heteromorphus_CBS_117.55_
KAE8353323.1_hypothetical_protein_BDV28DRAFT_148202__Aspergillus_coremiiformis_
XP_036533082.1_YD_repeat-containing_protein__Fusarium_subglutinans_
RKL46883.1_hypothetical_protein_BFJ72_g2622__Fusarium_proliferatum_
GAD93715.1_conserved_hypothetical_protein__Byssochlamys_spectabilis_No._5_
RPA93710.1_hypothetical_protein_L873DRAFT_1815418__Choiromyces_venosus_120613-1_
PQK10553.1_hypothetical_protein_BB8028_0002g08740__Beauveria_bassiana_
XP_013947354.1_hypothetical_protein_TRIATDRAFT_92284__Trichoderma_atroviride_IMI_206040_
XP_002585215.1_conserved_hypothetical_protein__Uncinocarpus_reesii_1704_
KAF7893083.1_hypothetical_protein_EAF02_000621__Botrytis_sinoallii_
KAF7940960.1_hypothetical_protein_EAE97_007145__Botrytis_byssoidea_
XP_001593015.1_hypothetical_protein_SS1G_05937__Sclerotinia_sclerotiorum_1980_UF-70_
RKK88667.1_hypothetical_protein_BFJ68_g16893__Fusarium_oxysporum_
EMR67380.1_putative_rhs_repeat_protein__Eutypa_lata_UCREL1_
PNP41813.1_hypothetical_protein_TGAMA5MH_06406__Trichoderma_gamsii_
XP_025426053.1_RHS_repeat_protein__Aspergillus_saccharolyticus_JOP_1030-1_
TGJ68804.1_hypothetical_protein_EYR41_004889__Arthrobotrys_oligospora_
RON86346.1_type_IV_secretion_protein_Rhs__Pseudomonas_fluorescens_
SEN27337.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_sp._ok266_
WP_095113555.1_type_IV_secretion_protein_Rhs__Pseudomonas_sp._Irchel_3H7_
CEL27250.1_tRNA3_Ser_-specific_nuclease_WapA_precursor__Pseudomonas_fluorescens_
VVN08933.1_hypothetical_protein_PS624_03726__Pseudomonas_fluorescens_
WP_186621152.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_sp._OE_48.2_
WP_150602639.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_fluorescens_
WP_169428602.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_fluorescens_
WP_178083489.1_type_IV_secretion_protein_Rhs__Pseudomonas_sp._91RF_
WP_130903966.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_sp._Sample_22_
WP_192226923.1_type_IV_secretion_protein_Rhs__Pseudomonas_sp._PDM02_
WP_131107395.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_sp._Sample_10_
WP_046039813.1_type_IV_secretion_protein_Rhs__Pseudomonas_fluorescens_
WP_192307702.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_sp._PDM04_
WP_103403329.1_type_IV_secretion_protein_Rhs__Pseudomonas_sp._DP16D-R1_
WP_007983980.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_sp._GM48_
WP_095152969.1_type_IV_secretion_protein_Rhs__Pseudomonas_sp._Irchel_s3b5_
WP_095943171.1_type_IV_secretion_protein_Rhs__Pseudomonas_sp._ACN8_
WP_085717411.1_type_IV_secretion_protein_Rhs__Pseudomonas_sp._B28_2017__
WP_048376809.1_hypothetical_protein__Pseudomonas_lundensis_
WP_161460866.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_syringae_
WP_025990986.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_syringae_
WP_122851823.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_viridiflava_
WP_082481464.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_sp._Leaf127_
WP_014339723.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_fluorescens_
WP_054062970.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_fuscovaginae_
WP_065833998.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_syringae_
GFM58983.1_sugar-binding_protein__Pseudomonas_cichorii_
GFM75114.1_sugar-binding_protein__Pseudomonas_cichorii_
WP_025389422.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_syringae_
WP_095066952.1_sugar-binding_protein__Pseudomonas_sp._Irchel_3A18_
GFM58987.1_sugar-binding_protein__Pseudomonas_cichorii_
WP_058823988.1_RHS_repeat-associated_core_domain-containing_protein__Pseudomonas_syringae_
WP_150696155.1_hypothetical_protein__Pandoraea_terrae_
WP_121280419.1_RHS_repeat-associated_core_domain-containing_protein__Paraburkholderia_sp._7MK8-2_
WP_081087658.1_hypothetical_protein__Burkholderia_stagnalis_
WP_183686536.1_hypothetical_protein__Rhizobium_lusitanum_
WP_080434334.1_hypothetical_protein__Burkholderia_ubonensis_
WP_137939972.1_RHS_repeat-associated_core_domain-containing_protein__Chitinivorax_sp._B_
WP_096754224.1_hypothetical_protein__Cedecea_neteri_
WP_038241735.1_hypothetical_protein__Xenorhabdus_szentirmaii_
WP_069018241.1_RHS_repeat-associated_core_domain-containing_protein__Yersinia_enterocolitica_
HAM3643588.1_hypothetical_protein__Escherichia_coli_
WP_164526018.1_RHS_repeat-associated_core_domain-containing_protein__Proteus_vulgaris_
BBV67681.1_sugar-binding_protein__Klebsiella_sp._STW0522-44_
WP_052293364.1_hypothetical_protein__Coxiella_burnetii_
WP_032664429.1_hypothetical_protein__Enterobacter_sp._DC4_
Kingdom
Actinobacteria
Arthropod
Ascomycota
Phage
Proteobacteria
Bootstrap
50
60
70
80
90
100
Lifestyle
Arthropod
Nematodes
Other
Plant Roots
Plants
Soil

Fig S1d. RHS protein phylogeny.

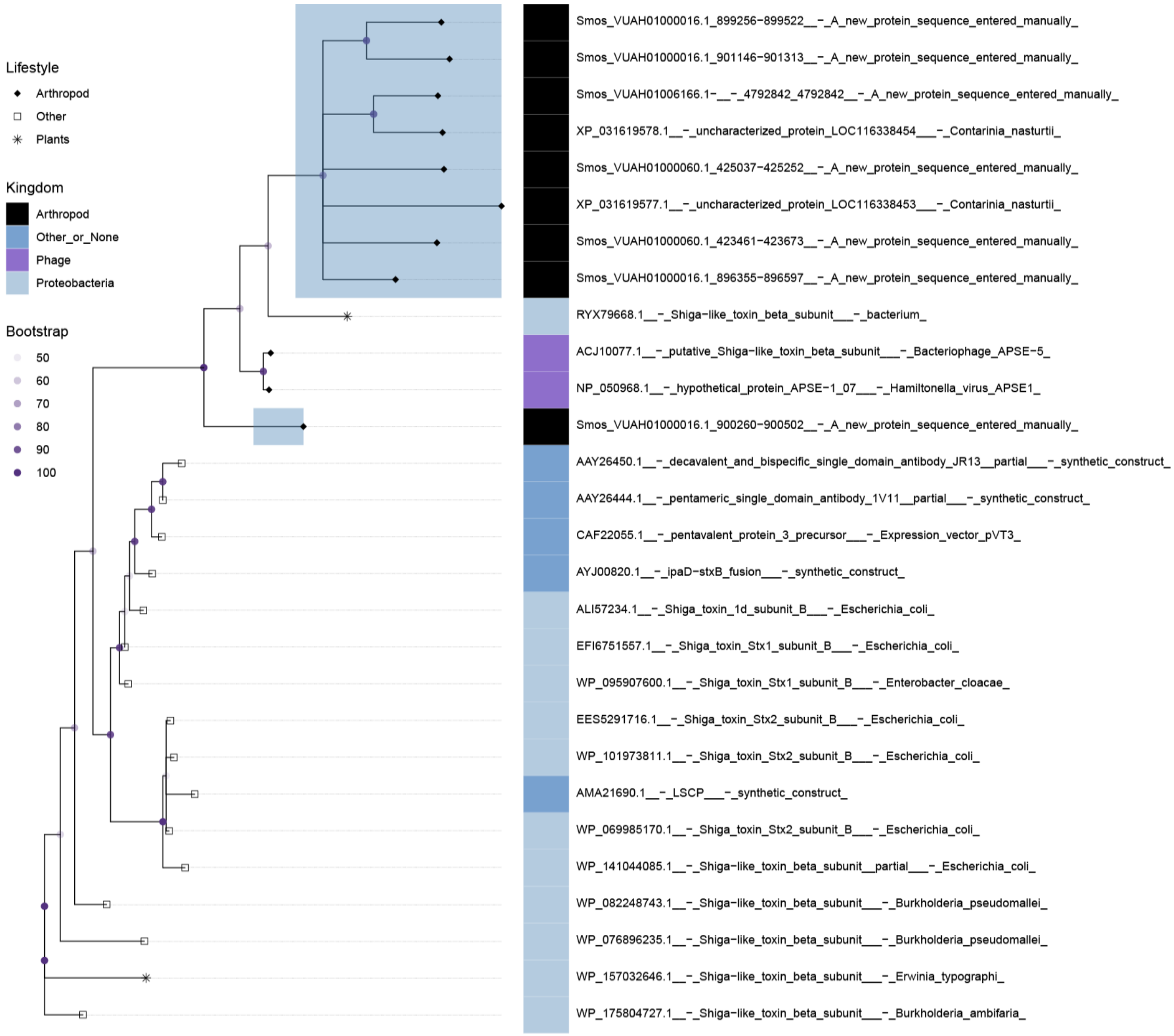
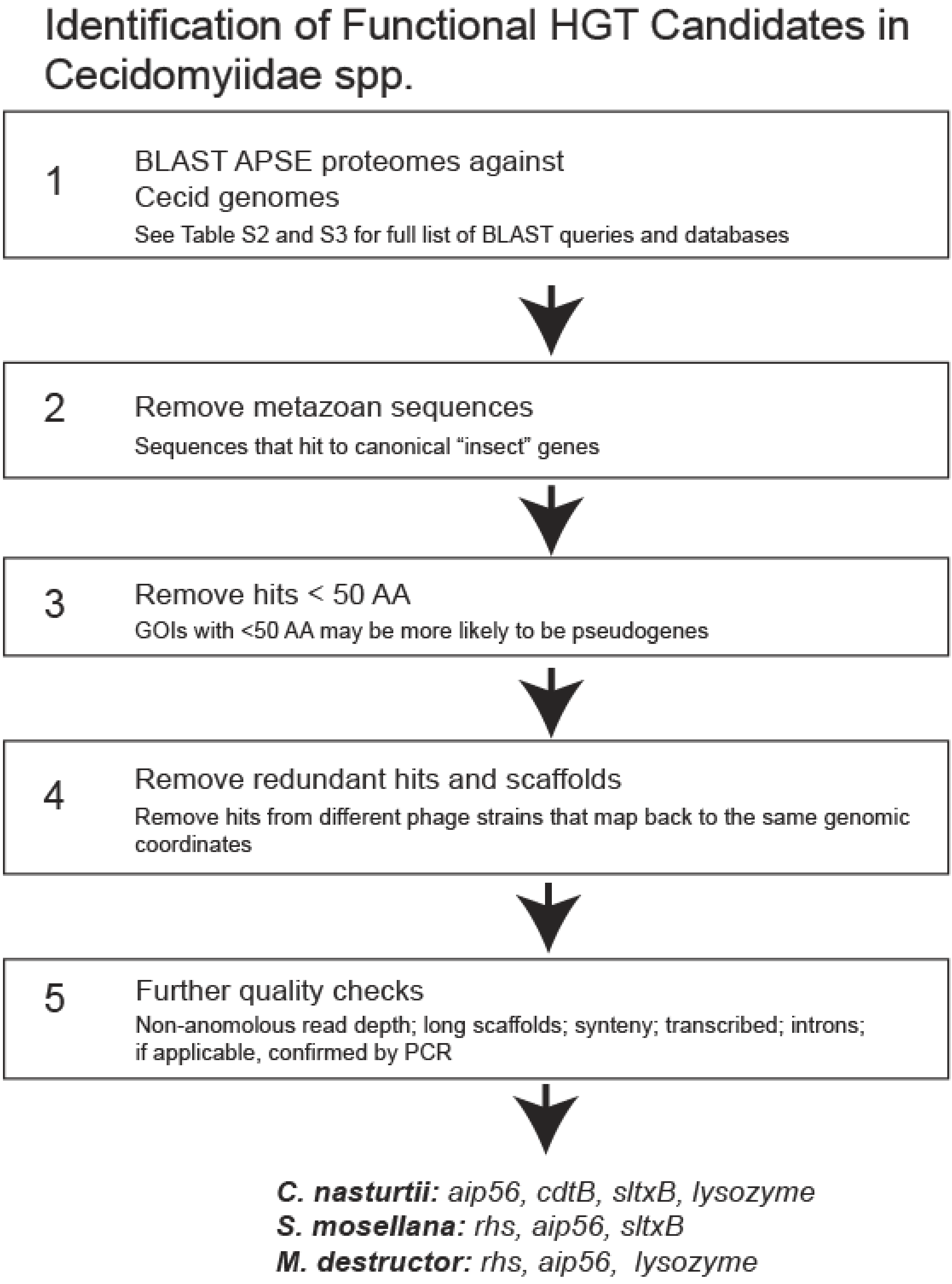

Lifestyle
◆ Arthropod
□ Other
✳ Plants
Kingdom
■ Arthropod
■ Other_or_None
■ Phage
■ Proteobacteria
Bootstrap
● 50
● 60
● 70
● 80
● 90
● 100
Smos_VUAH01000016.1_899256−899522__−_A_new_protein_sequence_entered_manually_
Smos_VUAH01000016.1_901146−901313__−_A_new_protein_sequence_entered_manually_
Smos_VUAH01006166.1−__−_4792842_4792842__−_A_new_protein_sequence_entered_manually_
XP_031619578.1__−_uncharacterized_protein_LOC116338454___−_Contarinia_nasturtii_
Smos_VUAH01000060.1_425037−425252__−_A_new_protein_sequence_entered_manually_
XP_031619577.1__−_uncharacterized_protein_LOC116338453___−_Contarinia_nasturtii_
Smos_VUAH01000060.1_423461−423673__−_A_new_protein_sequence_entered_manually_
Smos_VUAH01000016.1_896355−896597__−_A_new_protein_sequence_entered_manually_
RYX79668.1__−_Shiga−like_toxin_beta_subunit___−_bacterium_
ACJ10077.1__−_putative_Shiga−like_toxin_beta_subunit___−_Bacteriophage_APSE−5_
NP_050968.1__−_hypothetical_protein_APSE−1_07___−_Hamiltonella_virus_APSE1_
Smos_VUAH01000016.1_900260−900502__−_A_new_protein_sequence_entered_manually_
AAY26450.1__−_decavalent_and_bispecific_single_domain_antibody_JR13__partial___−_synthetic_construct_
AAY26444.1__−_pentameric_single_domain_antibody_1V11__partial___−_synthetic_construct_
CAF22055.1__−_pentavalent_protein_3_precursor___−_Expression_vector_pVT3_
AYJ00820.1__−_ipaD−stxB_fusion___−_synthetic_construct_
ALI57234.1__−_Shiga_toxin_1d_subunit_B___−_Escherichia_coli_
EFI6751557.1__−_Shiga_toxin_Stx1_subunit_B___−_Escherichia_coli_
WP_095907600.1__−_Shiga_toxin_Stx1_subunit_B___−_Enterobacter_cloacae_
EES5291716.1__−_Shiga_toxin_Stx2_subunit_B___−_Escherichia_coli_
WP_101973811.1__−_Shiga_toxin_Stx2_subunit_B___−_Escherichia_coli_
AMA21690.1__−_LSCP___−_synthetic_construct_
WP_069985170.1__−_Shiga_toxin_Stx2_subunit_B___−_Escherichia_coli_
WP_141044085.1__−_Shiga−like_toxin_beta_subunit__partial___−_Escherichia_coli_
WP_082248743.1__−_Shiga−like_toxin_beta_subunit___−_Burkholderia_pseudomallei_
WP_076896235.1__−_Shiga−like_toxin_beta_subunit___−_Burkholderia_pseudomallei_
WP_157032646.1__−_Shiga−like_toxin_beta_subunit___−_Erwinia_typographi_
WP_175804727.1__−_Shiga−like_toxin_beta_subunit___−_Burkholderia_ambifaria_

Fig S1e. SltxB protein phylogeny.

Fig S2. Workflow to identify functional horizontal gene transfer candidates in Cecidomyiidae species identifies at least five transfers of toxin-encoding genes.



Identification of Functional HGT Candidates in Cecidomyiidae spp.

1 BLAST APSE proteomes against Cecid genomes

See Table S2 and S3 for full list of BLAST queries and databases

2 Remove metazoan sequences

Sequences that hit to canonical "insect" genes

3 Remove hits < 50 AA

GOIs with <50 AA may be more likely to be pseudogenes

4 Remove redundant hits and scaffolds

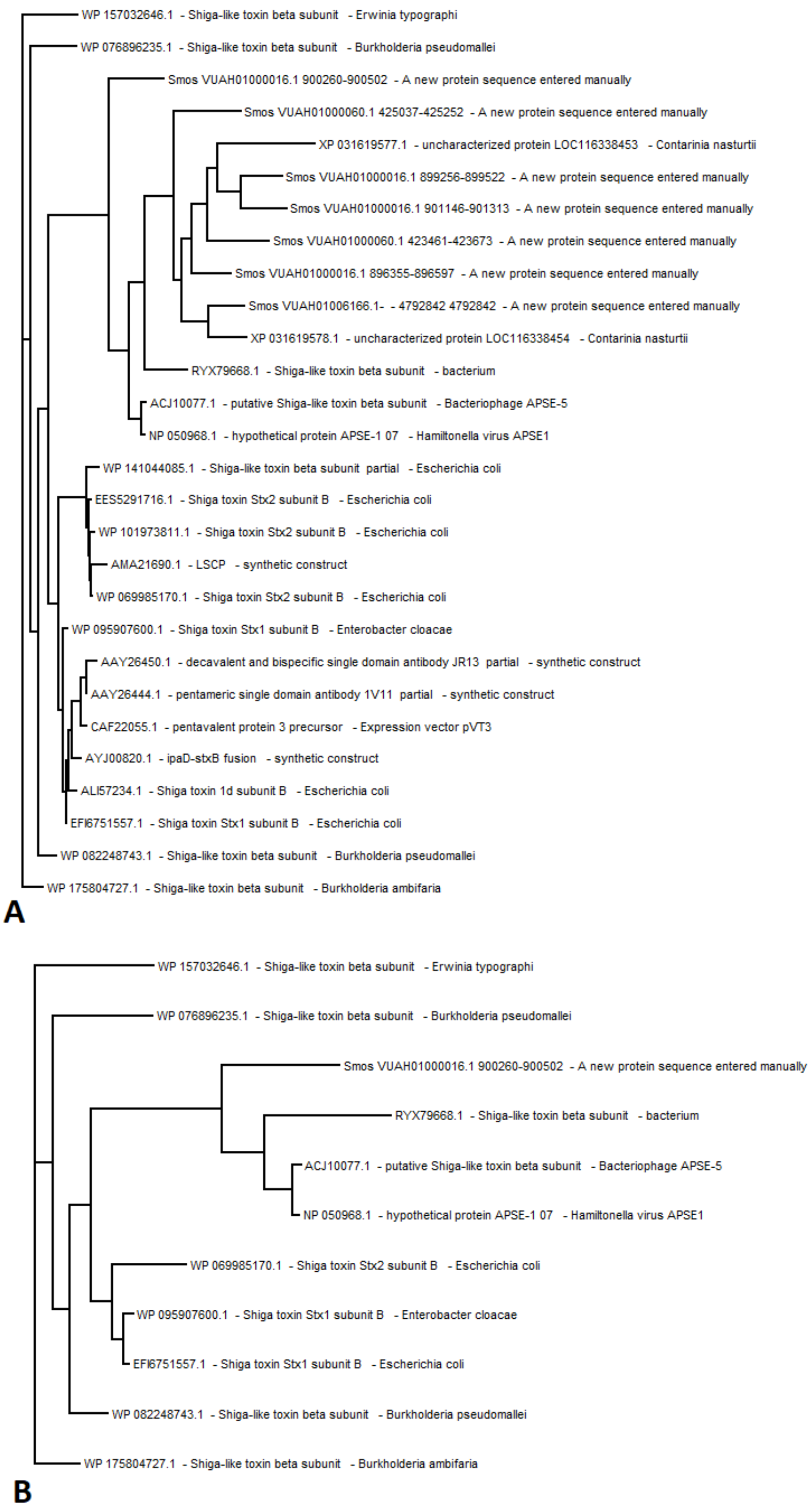
Remove hits from different phage strains that map back to the same genomic coordinates

5 Further quality checks

Non-anomolous read depth; long scaffolds; synteny; transcribed; introns; if applicable, confirmed by PCR

C. nasturtii: *aip56, cdtB, sltxB, lysozyme*
S. mosellana: *rhs, aip56, sltxB*
M. destructor: *rhs, aip56, lysozyme*

Fig S3. Two representative trees that were used in the comparison of phylogenetic signal. The tree labeled 'A' is the un-pruned tree, while the tree shown in 'B' has been trimmed to only include tips that were likely passed down via horizontal transfer.


WP 157032646.1 - Shiga-like toxin beta subunit - Erwinia typographi
WP 076896235.1 - Shiga-like toxin beta subunit - Burkholderia pseudomallei
Smos VUAH01000016.1 900260-900502 - A new protein sequence entered manually
Smos VUAH01000060.1 425037-425252 - A new protein sequence entered manually
XP 031619577.1 - uncharacterized protein LOC116338453 - Contarinia nasturtii
Smos VUAH01000016.1 899256-899522 - A new protein sequence entered manually
Smos VUAH01000016.1 901146-901313 - A new protein sequence entered manually
Smos VUAH01000060.1 423461-423673 - A new protein sequence entered manually
Smos VUAH01000016.1 896355-896597 - A new protein sequence entered manually
Smos VUAH01006166.1- - 4792842 4792842 - A new protein sequence entered manually
XP 031619578.1 - uncharacterized protein LOC116338454 - Contarinia nasturtii
RYX79668.1 - Shiga-like toxin beta subunit - bacterium
ACJ10077.1 - putative Shiga-like toxin beta subunit - Bacteriophage APSE-5
NP 050968.1 - hypothetical protein APSE-1 07 - Hamiltonella virus APSE1
WP 141044085.1 - Shiga-like toxin beta subunit partial - Escherichia coli
EES5291716.1 - Shiga toxin Stx2 subunit B - Escherichia coli
WP 101973811.1 - Shiga toxin Stx2 subunit B - Escherichia coli
AMA21690.1 - LSCP - synthetic construct
WP 069985170.1 - Shiga toxin Stx2 subunit B - Escherichia coli
WP 095907600.1 - Shiga toxin Stx1 subunit B - Enterobacter cloacae
AAY26450.1 - decavalent and bispecific single domain antibody JR13 partial - synthetic construct
AAY26444.1 - pentameric single domain antibody 1V11 partial - synthetic construct
CAF22055.1 - pentavalent protein 3 precursor - Expression vector pVT3
AYJ00820.1 - ipaD-stxB fusion - synthetic construct
ALI57234.1 - Shiga toxin 1d subunit B - Escherichia coli
EFI6751557.1 - Shiga toxin Stx1 subunit B - Escherichia coli
WP 082248743.1 - Shiga-like toxin beta subunit - Burkholderia pseudomallei
WP 175804727.1 - Shiga-like toxin beta subunit - Burkholderia ambifaria
A
WP 157032646.1 - Shiga-like toxin beta subunit - Erwinia typographi
WP 076896235.1 - Shiga-like toxin beta subunit - Burkholderia pseudomallei
Smos VUAH01000016.1 900260-900502 - A new protein sequence entered manually
RYX79668.1 - Shiga-like toxin beta subunit - bacterium
ACJ10077.1 - putative Shiga-like toxin beta subunit - Bacteriophage APSE-5
NP 050968.1 - hypothetical protein APSE-1 07 - Hamiltonella virus APSE1
WP 069985170.1 - Shiga toxin Stx2 subunit B - Escherichia coli
WP 095907600.1 - Shiga toxin Stx1 subunit B - Enterobacter cloacae
EFI6751557.1 - Shiga toxin Stx1 subunit B - Escherichia coli
WP 082248743.1 - Shiga-like toxin beta subunit - Burkholderia pseudomallei
WP 175804727.1 - Shiga-like toxin beta subunit - Burkholderia ambifaria
B

Supplementary File 1 Legend. These analyses support the finding that the horizontally transferred genes identified in this study are not due to contamination. 'Species' column shows the species in which the HGT occurred. 'APSE' and 'Protein ID' columns show the APSE strain and GenBank IDs of query sequences identified in the cecidomyiid genomes. For the 'Protein Name' column, we report a summary of the BLASTP results if they appear to correspond to one or more characterized proteins. In some cases, the identity and function of the protein are ambiguous and are labelled as 'Hypothetical Protein.' In the 'E-value' column, we report the lowest E-value in the case of multiple APSE protein queries. In some cases, a single TBLASTN query resulted in hits to multiple genomic 'ranges' on the same scaffold. If the subject sequences shared high AA identity **(>90%)** throughout multiple ranges, we considered these evidence of duplications of the GOI, and the E-values for each individual 'range' was reported in separate rows. 'Scaffold' and 'Scaffold Size' coordinates reflect GenBank accessions and associated lengths unless otherwise noted. 'GOI Coordinates' column reflects the TBLASTN reported ranges, unless the GOI has been annotated, in which case the annotation ID is shown. In the 'Other Eukaryotic Genes' column, we report if we found evidence of *bona fide* eukaryotic genes (Yes/No). Where the genome has been annotated, we report the Annotation IDs of the nearest eukaryotic gene proximal to the GOI. If the genome has not been annotated, we ran Augustus annotation on each scaffold under consideration using the 'fly' setting as implemented in Geneious (Stanke et al. 2004). In the 'Intron' and 'Exon Coordinates' column, we indicate the number of introns predicted by either annotations specific to the species or Augustus annotations. In some cases, Augustus did not predict any genes in the region of interest, in which case we reported 'NGP' for 'No Gene Predicted.' Note that Augustus relies on training on the appropriate gene sets (Stanke et al. 2004), and it may fail in cases of HGT due to the inherent differences of genes with lateral provenance. Where the GOI does not have an associated annotation ID, we report the Augustus-predicted exon coordinates. For the 'BWA' columns, please see **Supplementary Methods** - *BWA analysis* section. For 'Transcription' columns, please see **Supplementary Methods** - *Transcription analysis*. In the case of *C. nasturtii*, we indicate if we were able to successfully PCR the GOI (PCR primers and conditions are documented in **Table S4**).

Supplementary File 2 Legend. Taxonomic and lifestyle information for species in protein phylogenies shown in **Fig S1**. Included are citations if species are found on Arthropods, Plants or Soil.

Supplementary Bibliography

1. Altschul SF et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.
2. Andrews S 2010. FastQC: a quality control tool for high throughput sequence data.
3. Blake MC, Jambou RC, Swick AG, Kahn JW, Azizkhan JC. 1990. Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. Mol. Cell. Biol. 10:6632–6641.
4. Borges R, Machado JP, Gomes C, Rocha AP, Antunes A. 2019. Measuring phylogenetic signal between categorical traits and phylogenies. Bioinformatics. 35:1862–1869.
5. Cavener DR. 1987. Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. Nucleic Acids Res. 15:1353–1361.
6. Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 5:e11147.
7. Dorchin N, Harris KM, Stireman JO 3rd. 2019. Phylogeny of the gall midges (Diptera, Cecidomyiidae, Cecidomyiinae): Systematics, evolution of feeding modes and diversification rates. Mol. Phylogenet. Evol. 140:106602.
8. Graves BJ, Johnson PF, McKnight SL. 1986. Homologous recognition of a promoter domain common to the MSV LTR and the HSV tk gene. Cell. 44:565–576.
9. Gruson H et al. Hummingbird iridescence: an unsuspected structural diversity influences colouration at multiple scales.
10. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518–522.
11. Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 26:680–682.
12. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods. 14:587–589.
13. Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20:1160–1166.
14. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
15. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 10:845–858.
16. Koutsovoulos G et al. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini.* Proc. Natl. Acad. Sci. U. S. A. 113:5053–5058.
17. Kutach AK, Kadonaga JT. 2000. 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.
18. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25:1754–1760.
19. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22:1658–1659.
20. Lyons E et al. 2008. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. Plant Physiol. 148:1772–1781.

21. Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: 2010 Gateway Computing Environments Workshop (GCE).Vol. 1 IEEE p. 2.
22. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.
23. Nikoh N et al. 2010. Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLoS Genet. 6:e1000827.
24. Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.
25. Pinna C et al. 2020. Convergence in light transmission properties of transparent wing areas in clearwing mimetic butterflies. bioRxiv.
26. Proudfoot NJ. 2011. Ending the message: poly(A) signals then and now. Genes Dev. 25:1770–1782.
27. Raymondjean M, Cereghini S, Yaniv M. 1988. Several distinct 'CCAAT' box binding proteins coexist in eukaryotic cells. Proc. Natl. Acad. Sci. U. S. A. 85:757–761.
28. R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: URL https://www. R-project. org/.
29. Ronget V, Lemaître J-F, Tidière M, Gaillard J-M. 2020. Assessing the diversity of the form of age-specific changes in adult mortality from captive mammalian populations. Diversity. 12:354.
30. Sayers EW et al. 2019. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 47:D23–D28.
31. Schoch CL et al. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database . 2020.
32. Shine J, Dalgarno L. 1974. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. U. S. A. 71:1342–1346.
33. Sikora T, Jaschhof M, Mantič M, Kaspřák D, Ševčík J. 2019. Considerable congruence, enlightening conflict: molecular analysis largely supports morphology-based hypotheses on Cecidomyiidae (Diptera) phylogeny. Zool. J. Linn. Soc. 185:98–110.
34. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.
35. Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32:W309–12.
36. Thomas MC, Chiang C-M. 2006. The general transcription machinery and general cofactors. Crit. Rev. Biochem. Mol. Biol. 41:105–178.
37. Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Research. 44:W232–W235.
38. Verster KI et al. 2019. Horizontal transfer of bacterial cytolethal distending toxin B genes to insects. Mol. Biol. Evol. 36:2105–2110.
39. Yu G. 2020. Using ggtree to visualize data on tree-like structures. Curr. Protoc. Bioinformatics. 69:e96.
40. Yu G, Smith DK, Zhu H, Guan Y, Lam TT. 2017. Ggtree: An R package for visualization

and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8:28–36.