# The Toxicogenome of Hyalella azteca

Poynton, Helen C.; Colbourne, John K.; Hasenbein, Simone; Benoit, Joshua B.; Sepulveda, Maria S.; Poelchau, Monica F.; Hughes, Daniel S.T.; Murali, Shwetha C.; Chen, Shuai; Glastad, Karl M.; Goodisman, Michael A.D.; Werren, John H.; Vineis, Joseph H.; Bowen, Jennifer L.; Friedrich, Markus; Jones, Jeffery; Robertson, Hugh M.; Feyereisen, René; Mechler-Hickson, Alexandra; Mathers, Nicholas

*DOI:*
10.1021/acs.est.8b00837

# Supporting Information

## The Toxicogenome of *Hyalella azteca*: a model for sediment ecotoxicology and evolutionary toxicology

Helen C. Poynton, Simone Hasenbein, Joshua B. Benoit, Maria S. Sepulveda, Monica F. Poelchau, Daniel S.T. Hughes, Shwetha C. Murali, Shuai Chen, Karl M. Glastad, Michael A. D. Goodisman, John H. Werren, Joseph H. Vineis, Jennifer L. Bowen, Markus Friedrich, Jeffery Jones, Hugh M. Robertson, René Feyereisen, Alexandra Mechler-Hickson, Nicholas Mathers, Carol Eunmi Lee, John K. Colbourne, Adam Biales, J. Spencer Johnston, Gary A. Wellborn, Andrew J. Rosendale, Andrew G. Cridge, Monica C. Munoz-Torres, Peter A. Bain, Austin R. Manny, Kaley M. Major, Faith N. Lambert, Chris D. Vulpe, Padrig Tuck, Bonnie J. Blalock,Yu-Yu Lin, Mark E. Smith, Hugo Ochoa-Acuña, Mei-Ju May Chen, Christopher P. Childers, Jiaxin Qu, Shannon Dugan, Sandra L. Lee, Hsu Chao, Huyen Dinh, Yi Han, HarshaVardhan Doddapaneni, Kim C. Worley, Donna M. Muzny, Richard A. Gibbs, Stephen Richards

**Summary:**

The supporting information includes 142 pages of additional methods (S1, including Table S1), detailed annotation reports (S2-S4, including 17 tables and 22 figures), gene expression data tables and figures (Tables S5.1-S5.4 and Figure S5), supplemental sequence files (S6), and detailed author contributions (S7).

## Table of Contents

## S1. Supplemental Methods

### Calculation of genome size

The genome size of male and female *H. azteca* were determined by flow cytometry. One adult *H. azteca* and 1/5 of the head of a *D.virilis* standard strain female (1C = 328 Mbp) were placed in a 2 ml Kontes Dounce tissue homogenizer with 1ml of Galbraith buffer. Nuclei were released from the sample and standard by grinding with 15 strokes of the “B” or loose pestle at a rate of 3 strokes every two seconds. The ground mixture was filtered through a 20um nylon filter, and stained with 25 μl/mg propidium iodide. The stained samples were held in the cold and dark for at least 20 minutes, then the amount of red propidium iodide fluorescence of the 2C nuclei from the sample and standard was scored with a Partec Cytoflow cytometer equipped with a Cobalt samba laser emitting 100mw of light at 532 nm. The 1C amount of DNA per sample was determined as the average channel number of the sample 2C nuclei divided by the average channel number of the 2C nuclei of the standard times 328 mbp. A minimum of 1000 2C nuclei were scored for the sample and the standard; the CV of the standard and sample 2C peak was 2.0 or less for all samples.

### Library preparation for genome sequencing

Genomic DNA (gDNA) was pooled from several individuals to construct all of the libraries. To prepare the 180bp and 500bp libraries, we used a gel-cut paired end library protocol. Briefly, 1 µg of the DNA was sheared using a Covaris S-2 system (Covaris, Inc. Woburn, MA) using the 180-bp or 500-bp program. Sheared DNA fragments were purified with Agencourt AMPure XP beads, end-repaired, dA-tailed, and ligated to Illumina universal adapters. After adapter ligation, DNA fragments were further size selected by agarose gel and PCR amplified for 6 to 8 cycles using Illumina P1 and Index primer pair and Phusion® High-Fidelity PCR Master Mix (New England Biolabs). The final library was purified using Agencourt AMPure XP beads and quality assessed by Agilent Bioanalyzer 2100 (DNA 7500 kit) determining library quantity and fragment size distribution before sequencing.

The long mate pair libraries with 3kb or 8kb insert sizes were constructed according to the manufacturer’s protocol (Mate Pair Library v2 Sample Preparation Guide art # 15001464 Rev. A PILOT RELEASE). Briefly, 5 µg (for 2 and 3-kb gap size library) or 10 µg (8-10 kb gap size library) of gDNA was sheared to desired size fragments by Hydroshear (Digilab, Marlborough, MA), then end repaired and biotinylated. Fragment sizes between 3-3.7 kb (3kb) or 8-10 kb

(8kb) were purified from 1% low melting agarose gel and then circularized by blunt-end ligation. These size selected circular DNA fragments were then sheared to 400-bp (Covaris S-2), purified using Dynabeads M-280 Streptavidin Magnetic Beads, end-repaired, dA-tailed, and ligated to Illumina PE sequencing adapters. DNA fragments with adapter molecules on both ends were amplified for 12 to 15 cycles with Illumina P1 and Index primers. Amplified DNA fragments were purified with Agencourt AMPure XP beads. Quantification and size distribution of the final library was determined before sequencing as described above.

**Automated Gene Annotation Using a Maker 2.0 Pipeline Tuned for Arthropods**

The HAZT_1.0 genome assembly was subjected to automatic gene annotation using a Maker 2.0 annotation pipeline tuned specifically for arthropods. The pipeline is designed to be systematic providing a single consistent procedure for the species in the pilot study, scalable to handle 100's of genome assemblies, evidence guided using both protein and RNAseq evidence to guide gen models, and targeted to utilize extant information on arthropod gene sets. The core of the pipeline was a Maker 2[1] instance, modified slightly to enable efficient running on our computational resources. The genome assembly was first subjected to de-novo repeat prediction and CEGMA analysis to generate gene models for initial training of the ab-initio gene predictors. Three rounds of training of the Augustus[2] and SNAP[3] gene predictors within Maker were used to bootstrap to a high quality training set. Input protein data included 1 million peptides from a non-redundant reduction (90% identity) of Uniprot Ecdysozoa (1.25 million peptides) supplemented with proteomes from eighteen additional species (*Strigamia maritime, Tetranychus urticae, Caenorhabditis elegans, Loa loa, Trichoplax adhaerens, Amphimedon queenslandica, Strongylocentrotus purpuratus, Nematostella vectensis, Branchiostoma floridae, Ciona intestinalis, Ciona savignyi, Homo sapiens, Mus musculus, Capitella teleta, Helobdella robusta, Crassostrea gigas, Lottia gigantean, Schistosoma mansoni*) leading to a final nr peptide evidence set of 1.03 million peptides. We used CEGMA models for QC purposes: of 1,977 CEGMA single copy ortholog gene models, 1,749 were found in the assembly and 1,679 in the final predicted gene set – a reasonable result given the small contig sizes of the assembly (HAZT_1.0). Finally, the pipeline uses a nine-way homology prediction with human, *Drosophila* and *C. elegans*, and InterPro Scan5 to allocate gene names. The automated gene sets are available from the BCM-HGSC website (https://www.hgsc.bcm.edu/arthropods/hyalella-azteca-genome-project), from the Ag Data Commons[4] as well as the National Agricultural Library (https://i5k.nal.usda.gov/Hyalella_azteca) where a web-browser of the genome, annotations, and supporting annotation data is accessible.

### Identification of bacterial contamination in genome assembly

Putative bacterial contamination in the assembled *H. azteca* genome HAZT_1.0 was identified using two complementary approaches. In the first approach, contamination was computationally identified using two different python-based computational pipelines. The *H. azteca* assembly was first analyzed using the homology-based pipeline described in Wheeler *et al.*[5] and more details are provided in **S2**. This pipeline was developed to identify lateral gene transfer (LGT) events in an assembled genome, but it is applicable here because it identifies scaffolds (based on e-value) that appear to be of bacterial, not amphipod, origin. A total 140 scaffolds were marked as likely bacterial. In the second approach, contamination in the HAZT_1.0 contigs was identified via NCBI contamination screening with reduced stringency on the BLAST parameters (Terence Murphy, pers. comm.). Information from both approaches was combined, affecting 257 scaffolds, and gene models that overlapped contaminated regions by >1 bp were removed from the OGSv1.0.

### RNA sequencing and transcriptome libraries

Two sets of transcriptomic data were generated to assist in gene prediction, including the "mixed juveniles" library and "multi-aged, mixed" library. For the mixed juveniles library, conditions were created to reproduce those used in traditional toxicity testing.[6] Ten animals aged 7-8 days were cultured in individual beakers under standard conditions with 5 ml of sand substrate and 175 ml of overlying control water (15 mg/L Cl and 0.02 mg/L Br); animals were fed a diet of 0.5 mg diatoms (*Thalassiosira weissflogii*; Reed Mariculture, Campbell, CA, USA) and 0.25 mg Tetramin fishfood (Tetra, Blacksburg, VA, USA) daily, with water renewals at least three times per week. After ten days, animals were harvested from the beakers with a pipet and placed immediately in RNAlater (Ambion, ThermoFisher Scientific,Waltham, MA, USA). Twelve biological replicates were generated.

In addition, two separate collections of animals were also made from laboratory cultures and pooled. The first pool included three adult males and four adult females; all females contained embryos, and one male and one female was an actively mating pair. The second pool contained nine juveniles aged < 24-h to 3-days old. All animals were depurated in standard culture media for 4-h prior to RNA isolation. Following isolation, RNA from these two collections was combined in equal amounts to create a pool of RNA representing "multi-aged, mixed".

For RNA extraction, RNAlater was removed from each *H. azteca* sample and the sample subsequently rinsed with TRI Reagent (Molecular Research Center, Cincinnati, OH, USA) to remove residual RNAlater. Following washing, 0.5 ml of fresh TRI Reagent was added to each tube and the sample was homogenized using a TissueLyser II bead mill (QIAGEN, Valencia, CA, USA). RNA was extracted according to the manufacture-supplied protocol and DNAse I treated using QIAGEN RNAeasy on-column digestion. RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA) at the University of Massachusetts Boston Center for Personalized Cancer Therapy (CPCT) Genomics Core.

400 ng of RNA was used for library preparation using an Illumina TruSeq Stranded mRNA kit (Illumina, San Diego, CA, USA) following the manufacturer's guidelines and using 12 different Illumina adapters. Library quantity and fragment pool length was accessed using an Agilent 2100 Bioanalzyer. The final thirteen libraries were sequenced using a paired-end (PE75) Rapid Run protocol on an Illumina HiSeq2500 instrument at the University of Massachusetts Boston CPCT Genomics Core producing approximately 35 million reads per sample. Sequence files were parsed using bcl2fastq (v2.17, Illumina). Raw reads were assessed for quality using FastQC software (v0.10.1, Babraham Bioinformatics, Babraham Institute, Babraham, Cambridge, United Kingdom). Low quality reads with Phred quality scores below 20 (fewer than 1%) and Illumina adapters were removed from the dataset using Trim Galore (v0.37, stringency 3, error rate 0, paired) (Babraham Bioinformatics, Babraham Institute, Babraham, Cambridge, United Kingdom) at the Massachusetts Green High Performance Computing Center (Holyoke, MA, USA). RNAseq reads were aligned to the *H. azteca* genome scaffolds (HAZT_1.0) using tophat 2.0.14 with bowtie 2-2.1.0 and samtools 1.2. Overall mapping rate was 70.2%. Resulting bam files were transferred to NAL and added to the Apollo genome browser.

**Cadmium and PCB126 exposures and RNA isolation**

To identify genes that respond to cadmium and organic pollution exposure, 48-h exposures were performed to cadmium chloride ($CdCl_2$) (99.0%, Fisher Scientific, Pittsburgh, PA) and PCB126 (99.4%, Accustandard, New Haven, CT), a co-planer PCB with dioxin-like toxicity and a component of Aroclor 1254. Concentrations for these exposures were selected by considering the environmental relevance and toxicity of the chemicals. The concentration chosen for the PCB126 exposure, 7.0 $\mu$g/L, is equivalent to the invertebrate final acute value determined for Aroclor 1254[7] The concentration chosen for $CdCl_2$ exposure, 5.5 $\mu$g/L, is equal to the $LC_{10}$ determined experimentally in our laboratory. Methanol (final concentration equal to

0.06 %) was used as a solvent to solubilize PCB126 and was chosen for its low toxicity to *H. azteca*.[8] Methanol was also added at 0.06% to the control and $CdCl_2$ exposures. Four exposures were prepared in reformulated moderately hard reconstituted water (RMHRW) containing: methanol only (control), 7.0 $\mu$g/L PCB126 in methanol, 5.5 $\mu$g/L $CdCl_2$ with methanol, or a mixture of 7.0 $\mu$g/L PCB126 in methanol and 5.5 $\mu$g/L $CdCl_2$.

Thirty 10-d old *H. azteca* were added to 150 ml of each treatment in glass exposure vessels, sealed to decrease the volatilization of methanol, and placed in an diurnal environmental chamber for 48-h at 23$^{o}$C. Temperature, dissolved oxygen, pH and conductivity were monitored after 24-h and at the conclusion of the experiment. After 48-h, organisms were removed from the exposure vessels, washed in clean RMHRW and immediately placed in TriReagent (Molecular Research Center, Cincinnati, OH) for RNA Isolation following the methods described in main text. For each treatment vessel, RNA was isolated as a pool from 10 individuals.

**Microarray analysis and identification of differentially expressed transcripts**

RNA from three biological replicates per treatment was reverse transcribed and labeled using Nimblegen two-color DNA labeling kit (Roche Nimblegen, Madison, WI) following the protocol of Lopez & Colbourne.[9] Labeled cDNA was hybridized to a 133k *H. azteca* NimbleGen 12-plex oligonucleotide microarray (GPL17458; described in Weston et al.[10]) in a Nimblegen hybridization system following the manufacture's protocols. Following hybridization, microarrays were washed with the Nimblegen wash buffer kit (Roche Nimblegen, Madison, WI) according to the manufacturer's recommendations. Each treatment included three biological replicate samples, each of which was hybridized to two arrays as dye-swapped technical replicates. Therefore, six competitive hybridizations were performed per sample, following the microarray loop design[11]: $Cd_1$ vs. $solvent_1$, $Cd_2$ vs. $mixture_1$, $Cd_3$ vs. $PCB126_1$, $PCB126_2$ vs. $mixture_2$, $PCB126_3$ vs. $solvent_2$, and $mixture_3$ vs. $solvent_3$ (where the subscript indicates the biological replicate). Differentially expressed genes were identified through the following steps. Raw data from all six comparisons was analyzed and normalized using the statistical software packages R[12] and Bioconductor.[13] Differential expression was obtained from the median $\log_2$-fluorescence intensity of probes by LIMMA (Linear Models for Microarray Data) functions[14] and t-statistics from the empirical Bayes method[15] after quantile normalization of probes across chips, samples and replicates. To determine the significance of differential expression, *t*-values were calculated for each gene and adjusted for multiple testing[16] using the

Bioconductor LIMMA package in R. Differentially expressed contigs with a false discovery rate (FDR) of 1% were identified and included in the present analysis.

**S1. References:**

(1) Cantarel, B. L.; Korf, I.; Robb, S. M.; Parra, G.; Ross, E.; Moore, B.; Holt, C.; Alvarado, A. S.; Yandell, M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* **2008,** *18*, (1), 188-196.
(2) Stanke, M.; Diekhans, M.; Baertsch, R.; Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **2008,** *24*, (5), 637-644.
(3) Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **2004,** *5*, (1), 59.
(4) Hughes, D. S. T.; Richards, S.; Poynton, H. Hyalella azteca Genome Annotations v0.5.3. Ag Data Commons, 2018, http://dx.doi.org/10.15482/USDA.ADC/1415993.
(5) Wheeler, D.; Redding, A. J.; Werren, J. H. Characterization of an ancient lepidopteran lateral gene transfer. *PloS one* **2013,** *8*, (3), e59262.
(6) U.S. EPA. *Methods for measuring the toxicity and bioaccumulation of sediment-associated contaminants with freshwater invertebrates*; US Environmental Protection Agency: 2000.
(7) Fuchsman, P. C.; Barber, T. R.; Lawton, J. C.; Leigh, K. B. An evaluation of cause-effect relationships between polychlorinated biphenyl concentrations and sediment toxicity to benthic invertebrates. *Environ Toxicol Chem* **2006,** *25*, (10), 2601-2612.
(8) Bowman, M. C.; Oiler, W. L.; Cairns, T.; Gosnell, A. B.; Oliver, K. H. Stressed bioassay systems for rapid screening of pesticide residues. Part I: Evaluation of bioassay systems. *Archives of Environmental Contamination and Toxicology* **1981,** *10*, (1), 9-24.
(9) Lopez, J.; Colbourne, J. Dual-labeled expression microarray protocol for high-throughput genomic investigations. *CGB Technical Report* **2011,** *201*, (1), 2.
(10) Weston, D. P.; Poynton, H. C.; Wellborn, G. A.; Lydy, M. J.; Blalock, B. J.; Sepulveda, M. S.; Colbourne, J. K. Multiple origins of pyrethroid insecticide resistance across the species complex of a nontarget aquatic crustacean, Hyalella azteca. *Proc Natl Acad Sci U S A* **2013,** *110*, (41), 16532-7.
(11) Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* **2002,** *32*, 490-495.
(12) Ihaka, R.; Gentleman, R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **1996,** *5*, (3), 299-314.
(13) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **2004,** *5*, (10), R80.
(14) Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **2004,** *3*, (1), 1-25.
(15) Kendziorski, C.; Newton, M.; Lan, H.; Gould, M. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **2003,** *22*, (24), 3899-3914.
(16) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **1995**, 289-300.

## Table S1: Sequencing, assembly, annotation statistics and accession numbers

| | | |
|---|---|---|
| **Bio Projects** | i5K Pilot NCBI Bio-project | PRJNA163973<br>http://www.ncbi.nlm.nih.gov/bioproject/163973 |
| | *Hyalella Azteca* NCBI Bio-project | PRJNA243935<br>https://www.ncbi.nlm.nih.gov/bioproject/243935 |
| | NCBI Bio-sample | SAMN02978968<br>https://www.ncbi.nlm.nih.gov/biosample/SAMN02978968 |
| **Genome Sequence** | 180bp (132 bp actual) insert DNA | 2 Illumina HiSeq 2000 run: 236M read pairs, 47.7 Gbp and 120.4M read pairs, 24.3 Gbp |
| | 500bp (532 bp actual) insert DNA | 1 Illumina HiSeq 2000 run: 55.2M read pairs, 11.2 Gbp |
| | 3kb (3,139 bp actual) insert DNA | 2 Illumina HiSeq 2000 runs: 297.9M read pairs, 60.2 Gbp |
| | 8kb (7,518 bp actual) insert DNA | 1 Illumina HiSeq 2000 run: 249.6M read pairs, 50.4 Gbp |
| | 180bp insert NCBI SRA Accession | SRX685181 and SRX685176<br>https://www.ncbi.nlm.nih.gov/sra/SRX685181 &<br>https://www.ncbi.nlm.nih.gov/sra/SRX685176 |
| | 500bp insert NCBI SRA Accession | SRX685180<br>https://www.ncbi.nlm.nih.gov/sra/SRX685180 |
| | 3kb insert NCBI SRA Accession | SRX685179<br>https://www.ncbi.nlm.nih.gov/sra/SRX685179 |
| | 8kb insert NCBI SRA Accession | SRX685177<br>https://www.ncbi.nlm.nih.gov/sra/SRX685177 |
| **Genome Assembly Hazt_1.0 Allpaths/Atlas** | Number of contigs | 216,093 |
| | Contig N50 | 5,445 bp |
| | Number of scaffolds | 10,380 |
| | Scaffold N50 | 987,977 bp |
| | Size of final assembly | 1,178,848,281 bp |
| | Size of final assembly - without gaps | 596,612,604 bp |
| | NCBI Genome Assembly Accession | GCA_000764305.1<br>https://www.ncbi.nlm.nih.gov/assembly/GCA_000764305.1 |
| **Genome Assembly Hazt_2.0 Redundans Improvement** | Number of contigs | 23,280 |
| | Contig N50 | 115,114 bp |
| | Number of scaffolds | 18,000 |
| | Scaffold N50 | 215,427 bp |
| | Size of final assembly | 550,885,727 bp |
| | Size of final assembly - without gaps | 548,265,702 bp |
| | NCBI Genome Assembly Accession | GCA_000764305.2<br>https://www.ncbi.nlm.nih.gov/assembly/GCA_000764305.2 |
| **Automated Genome Annotation (Hazt_0.5.3)** | Genes (Hazt_0.5.3) | 12,906 |
| | Average Transcript length | 1,077.2 bp |
| | Average CDS length | 1,065.2 bp (355.1 aa) |
| | Exons per gene | 4.70 |
| | Genome Annotation Link | National Agricultural Library<br>https://i5k.nal.usda.gov/Hyalella_azteca |
| **Automated Genome Annotation (Hazt_2.0 Redundans Improvement)** | Genes (Hazt_2.0) | 19,936 |
| | Average Transcript length | 2,538 bp |
| | Average CDS length | 1,704 bp (568 aa) |
| | Exons per gene | 7.20 |
| | Genome Annotation Link | https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Hyalella_azteca/100/ |

# S2. Recovery of bacterial draft genomes and lateral gene transfer analysis from *Hyalella azteca* assembled scaffolds

John H. Werren

Biology Department, University of Rochester, Rochester, NY 14627 United States

Correspondence to: jack.werren@rochester.edu

Joseph H. Vineis and Jennifer L. Bowen
Department of Marine and Environmental Sciences, Marine Science Center, Northeastern University, Nahant, MA 01908 United States

Correspondence to: je.bowen@northeastern.edu

## Introduction

Genome sequencing of eukaryotic organisms often generates a significant portion of prokaryotic sequence information either due to contamination or biologically relevant associated microorganism.[1-3] Microbial sources of contamination originate from culture media and DNA sequencing reagents including those involved in extraction and sequencing library preparation. Biological sources of non-target genomes are from naturally occurring intracellular and extracellular inhabitants of the host or those that live in or as food.

There are significant challenges to identifying potential prokaryotic organisms that contaminate a eukaryotic genome assembly. Annotation tools can be useful in identifying foreign DNA that may be from alien sources including metazoan and prokaryotic lineages. However, these genes have the potential to exist in the eukaryote as a result of lateral gene transfer, so a close examination is necessary to tease out the true origin of the gene. For example, the tardigrade *Hypsibius dujardini* genome was originally published with 17.5% of all genes resulting from horizontal gene transfer (HGT).[4] Many of these genes were subsequently discovered to be bacterial contaminants and complete microbial genomes were assembled from the data.[2] We screened the *Hyalella azteca* assemblies for bacterial contamination and candidate lateral gene transfers using two different approaches, and detected two bacteria for which draft genomes were generated. One is in the *Flavobacteriaceae*, and the other in Comamonadaceae, with affinities to the genus *Ideonella*. Given the gene repertoires of these bacteria, they could be relevant to ecotoxicology of *H. azteca*. However, we are not able, at this stage, to differentiate whether these genomes are symbiotic, parasitic, or serve as a significant food source for *H. azteca.*

**Methods**

*DNA based pipeline of Wheeler:* A modification of the Wheeler et al.[5] DNA based homology pipeline was used to screen for bacterial contaminations and candidate LGTs. This original pipeline was developed by Dr. Dave Wheeler and John (Jack) Werren and has been subsequently modified by contributions from Sarah Kingan and Zhichao Yan. Versions of the pipeline have been used to detect bacterial to animal lateral gene transfers and bacterial contamination in a number of genome projects, including *Cimex lectularius*,[3] Hessian fly,[6] invasive ant *Cardiocondyla obscurior*,[7] sponge *Amphimedon queenslandica*,[8] Mediterranean fruitfly,[9] and *Trichogramma* wasps.[10]

The pipeline generates two outputs. One divides the target genome into 1kb fragments and screens each for bitscore similarity to an "animal" database (composed of gene models from vertebrates), and a bacterial database (composed of genome sequences from ~1000 microbial species with representations across different bacterial taxonomic groups). These bitscores are then compared to identify genome sections with strong bacterial bitscores but weak vertebrate scores. The comparison helps to separate highly conserved genes from those more likely to be bacterial in origin. All information for the regions of prokaryotic similarity are output into one text file, which can be exported into excel for viewing. The second output file is used to identify bacterial "contamination" within the genome assembly. The output provides information on the number of bacterial matches per scaffold, size of the bacterial matches, percent bacterial coverage, scaffold size, and the genus of the best bacterial hit. Based on the scaffold size and percent coverage, the scaffold is tentatively assigned as bacterial based on the proportion of the scaffold giving bacterial matches. Average read depth per scaffold and across bacterial-eukaryote junctions is used as an additional criterion. Follow-up manual curation identifies the extent and gene content of candidate LGTs and bacterial scaffolds. Employing this output, we have been able to identify complete and nearly complete bacterial scaffolds, as well as smaller contaminating scaffolds. These need to be removed from assemblies as they can create errors in gene annotation. But they also provide information of associated microbes in sequenced organisms.

*Tetranucleotide based method:* We next used a tetranucleotide frequency variability and short read mapping coverage among organisms to detect contaminating scaffolds in the metagenomic assemblies.[11] Illumina-utils (https://github.com/merenlab/illumina-utils) was used to filter low quality sequences from all illumina fastq files using "iu-filter-quality-minoche". Bowtie2[12] was then used to map the quality filtered reads to all scaffolds in the *H. azteca* assembly and samtools[13] was used to retain all reads that mapped to the assembly and to convert sam to bam format. We analyzed each scaffold greater than 2kb in the Hatz assembly to identify scaffolds containing prokaryotic characteristics. We used "anvi-gen-contigs-db" in Anvi'o[14] to call all genes in the assembly with prodigal,[15] calculate GC content, and tabulate tetranucleotide frequency of each scaffold. This information was stored in a contigs database. We detected prokaryotic single copy genes[16, 17] and ribosomal rRNA genes using "Anvi-run-hmms", which is an hmm model based search strategy.[18] Taxonomy of each scaffold was estimated using Centrifuge[19] and it was added to the contigs database. We used "anvi-profile" to add depth of coverage and single nucleotide variant (SNV) positions of each scaffold in the bam mapping file to the contigs database.

*Genome reconstruction*: Differences in the coverage and tetranucleotide frequency of each scaffold were used to calculate the Euclidean distances among all scaffolds and "anvi-interactive" visualized these distances as a tree according to "ward" clustering. Because tetranucleotide frequency composition is conserved within prokaryotic organisms[20] and the coverage depth of each bacterial scaffold should be consistently different from the *H. azteca* genome, prokaryotic derived scaffolds consistently cluster together in the tree. We explored each region of the tree that demonstrated an obvious prokaryotic signature and single copy gene hmm hits (**Figure S2.1**). We manually selected scaffolds from this region of the tree that showed a clear difference in coverage, summed to reasonable microbial genome size (3-5Mbp), and contained high completion and low redundancy of single copy genes.

## Results

We first used a modification of the DNA based pipeline of Wheeler et al.[5] to screen for both contaminating bacterial scaffolds and candidate lateral gene transfers. The pipeline has been successful in detecting bacterial associates in genome assemblies from other i5K projects genomes such as bedbug *Cimex lectularius*[3] and *Trichogramma* wasps.[10] The pipeline and follow-up analysis detected 140 likely bacterial scaffolds in Assembly HAZT_1.0. Among these were large scaffolds from the Citophaga-Flavobacterium-Bacteriodes (CFB) phylogroup, and a bacterium in the Comamonadaceae.

Tetranucleotide frequency variability and short read coverage was then used to screen the Redundans HAZT_2.0 assembly for candidate bacterial contaminating. From that assembly we recovered two draft genomes from the *H. azteca*, and there are possibly other bacterial contaminants in the assembly, but there is not enough support to report these metagenome assembled genomes (MAGs) with any confidence (Figure S2.1). The likely bacterial scaffolds have been removed from the assembly, and the draft genomes will be submitted as *H. azteca* associated metagenomes (NCBI BioProject: will be added prior to publication).

Annotation of the bacteria was performed on the two draft genomes. A summary of two high-quality bacterial MAGs is contained in Table S2.1 and Figure S2.2. MAG1 is divergent bacterium in the *Flavobacteriaceae*. The most similar 16S rRNA sequences currently available in the NCBI are to members of the genus Chishuiella, Weeksella, Algoriella and Flavobacterium, and only show ~92% identify. Therefore, based on the 16S rRNA sequences it likely represents a divergent bacterial type. Among the 2050 genes identified for MAG1, we found genes for nitrate and nitrite ammonification, ammonia assimilation, sulfur metabolism, and phosphorous metabolism.

16S rRNA sequence and gene annotation of MAG2 indicate that this bacterium is a gram-negative, motile heterotroph in the family Comamonadaceae, with close affinities (98% 16S rRNA identity) to members of the genus *Ideonella*. Like MAG1, MAG2, contains a significant number of genes for sulfur, nitrogen, and phosphorous metabolism. Some *Ideonella* are able to degrade polyethylene terephthalate (PET) plastics,[21] in particular *Idonella sakaiensis,* an isolate from the wax moth *Galleria mellonella*.[22] *I. sakaiensis* contains two genes critical for the metabolism of PET, a poly(ethylene terephthalate) hydrolase (PETase; A0A0K8P6T7) and mono(2-hydroxyethyl) terephthalate hydrolase (MHET hydrolase; A0A0K8P8E7). To determine if MAG2 contained PET metabolizing enzymes, using tblastn we blasted these *I. sakaiensis* genes against the MAG2 assembled genome. There were no significant homologies to the PETase, but we identified a protein with 38% similarity to the MHET hydrolase suggesting that this pathway may be partially conserved in MAG2.

There were no exact hits to any 16S rRNA sequences derived from either MAG to the NCBI or the SILVA119 database indicating that both are novel organisms. We found little nucleotide variability in the population of both MAGs indicating a nearly clonal population (Figure S2.2A). High regions of variability were, as expected, associated with ribosomal RNA genes (Figure S2.2B)

An analysis of broad functional categories indicates that both genomes contain multiple genes related to toxin and antibiotic resistance but MAG2 contains a significantly larger number (Figure S2.2C). Several of the categories explored, are only found in MAG2, but we cannot rule out the possibility that these genes exist in MAG1. MAG2 contains twice the number of nucleotides and genes (Table S2.1). The low coverage and similarity in coverage of these genomes to each other and the *H. azteca* genome presents challenges to accurately recruiting all scaffolds (Figure S2.1). However, the high completion estimates and 0% contamination (Table S2.1) provides evidence that these differences are real.

Several lateral gene transfer candidates were also found using the Wheeler et al.[5] pipeline (Table S2.2), which fell on five separate scaffolds of Hazt_1.0.  A LGT candidate was found on a sixth scaffold, but with lower confidence.  The most common prokaryotic match was to the genus *Rickettsia*, but evidence of LGT from the well characterized arthropod endosymbiont *Wolbachia* was also found.

| MAG | Length | | N50 | %GC | % comp | | genes | Order | Family | Genus |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.9 | 5 | 1 | 38 | 86 | 0 | 2050 | Flavobacteriales | Flavobacteriacea | . |
| 2 | 4.3 | 8 | 2.2 | 62 | 95 | 0 | 4518 | Burkholderiales | Comamonadaceae | Ideonella |

**Table S2.1. MAG ID, genome sizes, number and GC content of the scaffolds, N50 statistics, percent completion and redundancy and consensus taxonomy**. Completion and redundancy percentages are based on single copy genes from Campbell et al.[17] Consensus taxonomy was assigned based on comparison of 16S rRNA gene fragments from both genomes against SILVA119 and NCBI 16S refseq databases.

| LGT name | Scaffold: Position | e-value | Prokaryotic bit score | “Animal” bit score | Prokaryote Match | Genus | Species/strain | mRNA expression level |
|---|---|---|---|---|---|---|---|---|
| LGT_445 | 445: 8284-8783<br>445: 10333-10783<br>445: 15011-15322 | 99 | 367 | 0 | gi\|15603881\|ref\|NC_000963.1\| | *Rickettsia* | *prowazekii* str. Madrid E | none |
| LGT_557 | 557: 498654-498965<br>557: 499003-499176<br>557: 509076-509421 | 24<br>14<br>28 | 118<br>86<br>134 | 0<br>0<br>0 | gi\|157826385\|ref\|NC_009883.1\| | *Rickettsia* | *bellii* OSU 85-389 | low |
| LGT_633 | 663: 701912-702001 | 19 | 104 | 0 | gi\|157964072\|ref\|NC_009900.1\| | *Rickettsia* | *massiliae* MTU5 | none |
| LGT_884* | 884: 34748-34824<br>884: 61770-61846 | 12<br>11 | 80.6<br>77 | 0<br>0 | gi\|157826385\|ref\|NC_009883.1\| | *Rickettsia* | *bellii* OSU 85-389 | none<br>none |
| LGT_10_3.83 | 10: 3835537-3835667 | 27 | 129 | 0 | gi\|73667559\|ref\|NC_007355.1\| | *Methanosarcina* | *barkeri* str. Fusaro | none |
| LGT_332 | 332: 1152004-1152106 | 20 | 105 | 0 | gi\|190570478\|ref\|NC_010981.1\| | *Wolbachia* | endosymbiont of *Culex quinquefasciatus* Pel | none |

**Table S2.2: Candidate Bacteria-*Hyalella* Lateral Gene Transfers**. Candidates are shown along with the scaffold name and position, likely bacterial source, and matches strength based on bitscore relative to bit score to a reference animal set (see methods). These candidates will require further examination to determine validity. *Confidence in LGT_884 is weaker compared with the other candidate LGTs.

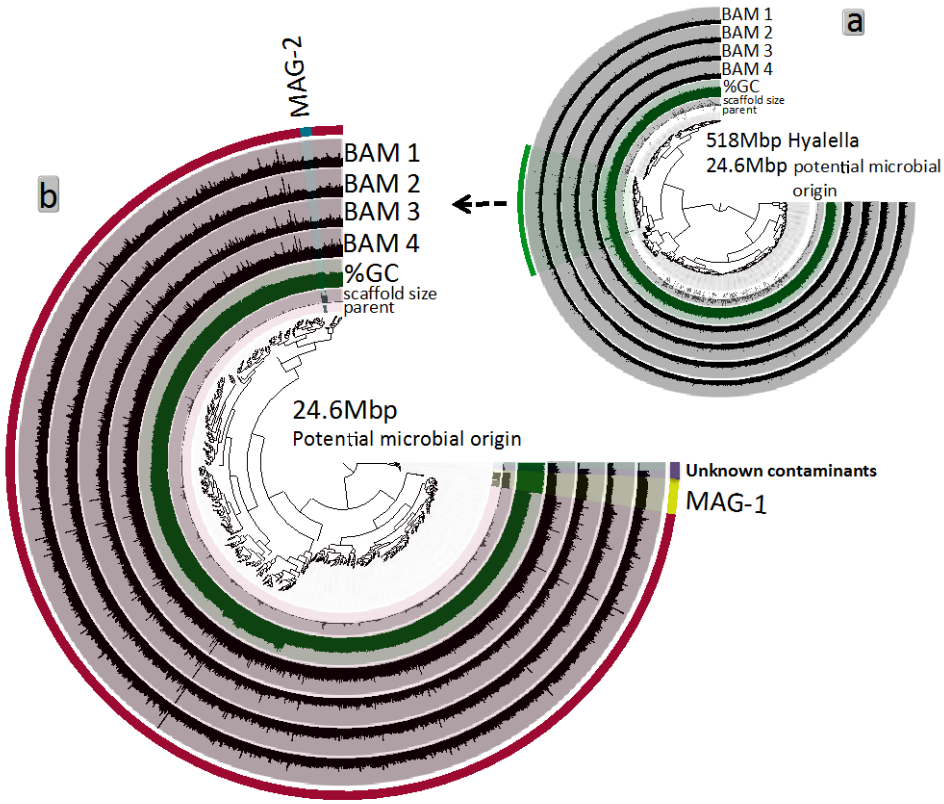



**Figure S2.1. Identification of microbial draft genomes**. (a) Visualization of all scaffolds in the *H. azteca* assembly organized by coverage and tetranucleotide sequence similarity. The scaffolds highlighted in green represent a portion of the tree that may be microbial based on the presence of single copy microbial genes and coverage profiles that are different from the majority of *H.* azteca scaffolds. (b) Selection of two metagenomic assembled draft microbial genomes (MAG)s based on their distinct coverage profiles and completion scores. The scaffolds included in the MAG are highlighted in yellow (MAG1) and blue (MAG2) in the outer ring of the display. All other contigs are highlighted in maroon. There are also two groups of scaffolds that are of unknown origin (purple) and have distinct coverage profiles.

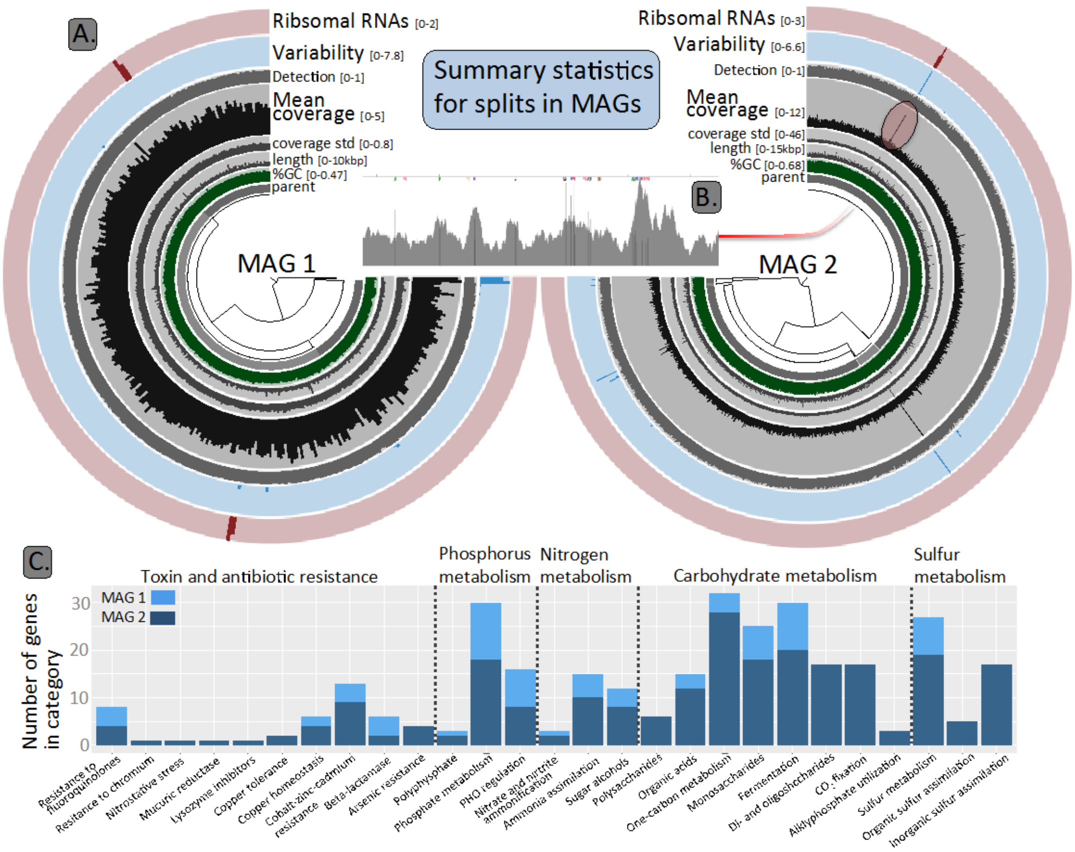



**Figure S2.2. Summary statistics and abundance of functional categories for each MAG**. (A.) Each ring around the tree displays a separate characteristic of the MAG at 5kbp intervals. Ribosomal RNAs: the count of ribosomal RNA genes; Variability: the number of single nucleotide variants per kbp; Detection: the percentage of the split with a coverage > 0; Mean coverage: the mean coverage of the split; coverage std: the standard deviation in mean coverage; length: the length of the split; %GC: the percent GC content of the split; parent: the id of the parent. (B.) Visualization of the single highlighted (pink circle) split in MAG2 at single nucleotide resolution. Positions of variability are shown as fine black lines. The coverage range of the plot is 0 – 30. (C.) The count of genes within broad functional categories detected in MAG1 and MAG2 based on annotation in the RAST database.

## S2. References


(1) Merchant, S.; Wood, D. E.; Salzberg, S. L. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* **2014,** *2*, e675.

(2) Koutsovoulos, G.; Kumar, S.; Laetsch, D. R.; Stevens, L.; Daub, J.; Conlon, C.; Maroon, H.; Thomas, F.; Aboobaker, A. A.; Blaxter, M. No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. *Proceedings of the National Academy of Sciences* **2016,** *113*, (18), 5053-5058.

(3) Benoit, J. B.; Adelman, Z. N.; Reinhardt, K.; Dolan, A.; Poelchau, M.; Jennings, E. C.; Szuter, E. M.; Hagan, R. W.; Gujar, H.; Shukla, J. N. Unique features of a global human ectoparasite identified through sequencing of the bed bug genome. *Nature Communications* **2016,** *7*, 10165.

(4) Boothby, T. C.; Tenlen, J. R.; Smith, F. W.; Wang, J. R.; Patanella, K. A.; Nishimura, E. O.; Tintori, S. C.; Li, Q.; Jones, C. D.; Yandell, M. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences* **2015,** *112*, (52), 15976-15981.

(5) Wheeler, D.; Redding, A. J.; Werren, J. H. Characterization of an ancient lepidopteran lateral gene transfer. *PloS One* **2013,** *8*, (3), e59262.

(6) Zhao, C.; Escalante, L. N.; Chen, H.; Benatti, T. R.; Qu, J.; Chellapilla, S.; Waterhouse, R. M.; Wheeler, D.; Andersson, M. N.; Bao, R. A massive expansion of effector genes underlies gall-formation in the wheat pest Mayetiola destructor. *Current Biology* **2015,** *25*, (5), 613-620.

(7) Klein, A.; Schrader, L.; Gil, R.; Manzano-Marín, A.; Flórez, L.; Wheeler, D.; Werren, J. H.; Latorre, A.; Heinze, J.; Kaltenpoth, M. A novel intracellular mutualistic bacterium in the invasive ant Cardiocondyla obscurior. *The ISME Journal* **2016,** *10*, (2), 376.

(8) Conaco, C.; Tsoulfas, P.; Sakarya, O.; Dolan, A.; Werren, J.; Kosik, K. S. Detection of prokaryotic genes in the Amphimedon queenslandica genome. *PloS One* **2016,** *11*, (3), e0151092.

(9) Papanicolaou, A.; Schetelig, M. F.; Arensburger, P.; Atkinson, P. W.; Benoit, J. B.; Bourtzis, K.; Castañera, P.; Cavanaugh, J. P.; Chao, H.; Childers, C. The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. *Genome Biology* **2016,** *17*, (1), 192.

(10) Lindsey, A. R.; Werren, J. H.; Richards, S.; Stouthamer, R. Comparative genomics of a parthenogenesis-inducing Wolbachia symbiont. *G3: Genes, Genomes, Genetics* **2016,** *6*, (7), 2113-2123.

(11) Alneberg, J.; Bjarnason, B. S.; de Bruijn, I.; Schirmer, M.; Quick, J.; Ijaz, U. Z.; Lahti, L.; Loman, N. J.; Andersson, A. F.; Quince, C. Binning metagenomic contigs by coverage and composition. *Nature Methods* **2014,** *11*, (11), 1144-1146.

(12) Langmead, B.; Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **2012,** *9*, (4), 357.

(13) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009,** *25*, (16), 2078-2079.

(14) Eren, A. M.; Esen, Ö. C.; Quince, C.; Vineis, J. H.; Morrison, H. G.; Sogin, M. L.; Delmont, T. O. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **2015,** *3*, e1319.

(15) Hyatt, D.; Chen, G.-L.; LoCascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **2010,** *11*, (1), 119.

(16) Rinke, C.; Schwientek, P.; Sczyrba, A.; Ivanova, N. N.; Anderson, I. J.; Cheng, J.-F.; Darling, A.; Malfatti, S.; Swan, B. K.; Gies, E. A. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **2013,** *499*, (7459), 431.

(17) Campbell, J. H.; O'Donoghue, P.; Campbell, A. G.; Schwientek, P.; Sczyrba, A.; Woyke, T.; Söll, D.; Podar, M. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proceedings of the National Academy of Sciences* **2013,** *110*, (14), 5540-5545.
(18) Yoon, B.-J. Hidden Markov models and their applications in biological sequence analysis. *Current Genomics* **2009,** *10*, (6), 402-415.
(19) Kim, D.; Song, L.; Breitwieser, F. P.; Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research* **2016,** *26*, (12), 1721-1729.
(20) Noble, P. A.; Citek, R. W.; Ogunseitan, O. A. Tetranucleotide frequencies in microbial genomes. *Electrophoresis* **1998,** *19*, (4), 528-535.
(21) Yoshida, S.; Hiraga, K.; Takehana, T.; Taniguchi, I.; Yamaji, H.; Maeda, Y.; Toyohara, K.; Miyamoto, K.; Kimura, Y.; Oda, K. A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* **2016,** *351*, (6278), 1196-1199.
(22) Bombelli, P.; Howe, C. J.; Bertocchini, F. Polyethylene bio-degradation by caterpillars of the wax moth *Galleria mellonella*. *Current Biology 27*, (8), R292-R293.

## S3. Genome methylation and microRNAs

Shuai Chen,
OmicSoft Corporation, Cary, NC, USA

Karl Glastad
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Michael Goodisman
School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA

Maria Sepúlveda
Forestry and Natural Resources, Purdue University, West Lafayette, IN, USA

Correspondence to: Maria Sepúlveda, mssepulv@purdue.edu

### Introduction

DNA methylation is an epigenetics mechanism by which a methyl group ($CH_3$) covalently binds to DNA causing changes in gene expression. The majority of DNA methylation in animals occurs in the CpG dinucleotide context.[1] Methylated cytosines tend to mutate to thymines over evolutionary time. Therefore, if a genomic feature is methylated, then the observed frequency of CpG dinucleotides should be lower than the expected based upon the frequency of C and G nucleotides (i.e., CpG o/e will be less than 1.0). However, if a genomic feature is unmethylated then the observed frequency of CpG dinucleotides should equal the expected (i.e., CpG o/e will equal ~1.0).[2] The objective of our analyses was to analyze patterns of CpG dinucleotide depletion in the *H. azteca* genome in order to investigate putative patters of DNA methylation.

MicroRNAs (miRNAs) are a family of short non-coding RNAs (~22 nt in length) that play critical roles in post-translational gene regulation.[3] Many miRNAs families are conserved during evolution.[4] miRNA genes are usually transcribed into long primary miRNAs (pri-miRNAs, ~1000 nt) and then processed by nuclear protein Drosha and DGCR8 into hairpin precursor-miRNA (pre-miRNAs , ~ 80 nt) with mature miRNA located on one arm of the hairpin. After exportation to the cytoplasm, pre-miRNA hairpins are cleaved by RNase III enzyme Dicer to release the mature miRNAs, which guide the RNA-induced silencing complexes to suppress their target mRNAs' expression.[3] Recent research has revealed that miRNAs play key roles in aquatic crustaceans' response to environmental stressors (e.g. hypoxia and cadmium exposure),[5, 6] which makes miRNAs promising biomarkers for future aquatic toxicological research. We predicted *H. azteca* miRNAs based on sequence homology and hairpin structure identification.

## Methods

Normalized CpG dinucleotide content (CpG o/e) was calculated for genomic features as:

$$CpG\ o/e = \frac{length^2}{length} \times \frac{CpG\ count}{C\ count\ \times\ G\ count}$$

Features with CpG o/e or GpC o/e values of 0 or greater than 6 were excluded from analyses. Short features with length < 80bp were also excluded. For an estimate of genomic (non-coding) background we calculated CpG o/e for 1kb windows that did not fall within 2kb of an annotated gene model.

For metaplots, CpG o/e was calculated for sliding windows (window=200, step=20) across all genes with a length >= 3kb, as well as 1.5kb up and down-stream of these genes. Windows were dropped if CpG o/e content was zero for a given window or if > 50% of the window's bases were composed of masked bases. Genes were divided into two categories (high and low CpG o/e) based up whether they fell above or below the mean CpG o/e of all included genes. Mixture distributions were calculated using mixtools,[7] modeling each feature type's CpG o/e as a 2 component distribution.

*H*. *azteca* miRNAs were predicted according to method described by Chen et al.[8] Briefly, all animal mature miRNAs (miRBase Release 21) were searched against *H*. *azteca* genome by BLAST (e-value < -4)[9] for potential miRNA coding sites. The 50 nt flanking sequences surrounded the potential miRNA coding sites were fetched as candidate pre-miRNAs and folded by RNA-Fold[10] for hairpin structure identification.

## Results and Discussion

Our analyses show that *H*. *azteca* possesses strong indications of genomic DNA methylation, including CpG depletion of a subset of genes (Figure S3.1), and presence of the key DNA methyltransferase enzymes, DNMT1 and DNMT3 (Table S3.1). Furthermore, genes with lower levels of CpG o/e (putatively methylated) display a strong positional bias in CpG depletion (Figure S3.1 c, black line) with 5' regions of these genes being considerably depleted of CpGs. In contrast, genes with higher CpG o/e display no such positional bias (Figure S3.1 c, grey line). Several insects where DNA methylation has been empirically profiled at the single-base level possess such patterns of DNA methylation,[11] suggesting *H*. *azteca* likely has similar patterns of

DNA methylation as most insects which possess empirical measurement of single-base-resolution DNA methylation (but see: Glastad et al.[12]).

A total of 1,261 candidate miRNA coding sites were identified by BLAST. After hairpin structure identification, we predicted 148 *H. azteca* miRNAs which include several highly conserved miRNAs (e.g. miR-9 and let-7 family) (see Supplemental Table S6.1). A number of Cd-responsive miRNAs in *D. pulex* (miR-210, miR-71 and miR-252) have also been predicted in *H. azteca*, suggesting a conserved role of these miRNAs. This number of predicted miRNAs is comparable to what has been reported for other arthropods (Figure S3.2). Future research such as miRNA sequencing and miRNA expression profiles are required to validate our predictions.

## Table S3.1: Summary of epigenetic-related genes identified from *Hyalella azteca*.

| Gene Name | Gene ID | E-Value (Log) | Scaffold # | Similarity (%) | Length | Mismatches | Gaps | Query start position | Query end position | Scaffold start position | Scaffold stop position | Bit Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Activating Transcription Factor 2** | E9H1K5_DAPPU_ATF2 | -3.22 | 65 | 93 | 30 | 2 | 0 | 561 | 590 | 1916193 | 1916164 | 45 |
| | | -3.22 | 730 | 93 | 30 | 2 | 0 | 561 | 590 | 252302 | 252331 | 45 |
| **DNA Methyltransferase 1** | gi\|19263094\|gb\|AF483203.1\|DMNT1\| | -34.17 | 276 | 81 | 125 | 21 | 4 | 3890 | 4012 | 1447841 | 1447719 | 91 |
| | | -6.61 | 485 | 90 | 43 | 3 | 2 | 3092 | 3132 | 94892 | 94933 | 52 |
| | | -4.71 | 2227 | 87 | 45 | 4 | 2 | 3089 | 3132 | 34521 | 34478 | 49 |
| | | -3.41 | 874 | 89 | 38 | 2 | 2 | 3096 | 3132 | 111884 | 111920 | 47 |
| **DNA Methyltransferase 3** | NA | -86.04 | 324 | 47 | 311 | 154 | 2 | 368 | 689 | 38997 | 35941 | 103 |
| | LOC108673254 | -182.33 | 324 | 49 | 279 | 131 | 2 | 409 | 677 | 1648 | 2484 | 275 |
| **DGCR8** | E9GJ63_DAPPU_DGCR8 | -5.81 | 198 | 78 | 79 | 13 | 4 | 1261 | 1337 | 653841 | 653917 | 49 |
| **Dicer** | E9H7E4_DAPPU_Dicer | -23.23 | 247 | 78 | 134 | 23 | 7 | 4017 | 4147 | 391163 | 391293 | 76 |
| | | -7.13 | 352 | 89 | 44 | 3 | 2 | 3240 | 3282 | 103307 | 103349 | 53 |
| | | -5.23 | 1573 | 88 | 44 | 4 | 2 | 3240 | 3282 | 70127 | 70169 | 50 |
| | | -4.61 | 1351 | 80 | 71 | 8 | 5 | 3240 | 3307 | 12523 | 12590 | 49 |
| | | -4.28 | 990 | 85 | 49 | 6 | 2 | 3240 | 3287 | 370103 | 370055 | 49 |
| | | -3.54 | 105 | 86 | 47 | 4 | 3 | 3241 | 3286 | 182400 | 182355 | 47 |
| | | -3.32 | 67 | 86 | 44 | 4 | 2 | 3240 | 3282 | 1125442 | 1125400 | 47 |
| | | -3.32 | 151 | 84 | 51 | 6 | 3 | 3240 | 3289 | 1073176 | 1073225 | 47 |
| | | -3.32 | 169 | 86 | 45 | 4 | 3 | 3239 | 3282 | 1217958 | 1217915 | 47 |
| | | -3.32 | 341 | 86 | 43 | 6 | 0 | 3240 | 3282 | 237111 | 237153 | 47 |
| | | -3.32 | 354 | 86 | 43 | 6 | 0 | 3240 | 3282 | 976591 | 976549 | 47 |
| | | -3.32 | 441 | 86 | 44 | 4 | 2 | 3240 | 3282 | 376566 | 376608 | 47 |
| | | -3.32 | 470 | 86 | 44 | 4 | 2 | 3240 | 3282 | 344648 | 344606 | 47 |
| | | -3.32 | 1003 | 85 | 47 | 5 | 2 | 3240 | 3285 | 189871 | 189826 | 47 |
| | | -3.32 | 1050 | 85 | 47 | 5 | 2 | 3237 | 3282 | 165262 | 165217 | 47 |
| | | -3.32 | 1202 | 87 | 43 | 4 | 2 | 3241 | 3282 | 99255 | 99296 | 47 |
| | | -3.32 | 7572 | 88 | 41 | 3 | 2 | 3240 | 3279 | 1102 | 1063 | 47 |
| | | -3.32 | 8146 | 83 | 56 | 6 | 4 | 3240 | 3293 | 67 | 14 | 47 |
| **Dosha** | E9G2K5_DAPPU_Dosha | -24.80 | 546 | 79 | 125 | 21 | 6 | 1714 | 1834 | 148064 | 147944 | 77 |
| | | -22.53 | 2169 | 80 | 116 | 19 | 6 | 1535 | 1647 | 17819 | 17931 | 74 |
| | | -2.67 | 7 | 93 | 30 | 2 | 0 | 244 | 273 | 1987120 | 1987091 | 45 |
| | | -2.67 | 1202 | 93 | 30 | 2 | 0 | 244 | 273 | 80999 | 81028 | 45 |
| **Elongator Complex Protein 3** | E9FSS1_DAPPU_ELP3 | -32.35 | 15 | 81 | 120 | 18 | 6 | 960 | 1076 | 596349 | 596233 | 87 |
| **Gcn5-related N-Acetyltransferases** | E9H234_DAPPU_GCN6 | -35.46 | 686 | 81 | 137 | 21 | 7 | 1268 | 1400 | 521743 | 521875 | 92 |
| | | -31.75 | 1422 | 81 | 128 | 20 | 6 | 1264 | 1388 | 37975 | 38100 | 87 |
| | | -18.42 | 14 | 93 | 45 | 3 | 0 | 1915 | 1959 | 2970163 | 2970207 | 68 |
| | | -13.12 | 1491 | 86 | 56 | 8 | 0 | 1393 | 1448 | 100810 | 100755 | 60 |

**Table S3.1. Continued.**

| Gene Name | Gene ID | E-Value (Log) | Scaffold # | Similarity (%) | Length | Mismatches | Gaps | Query start position | Query end position | Scaffold start position | Scaffold stop position | Bit Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Histone Acetyltransferase 1** | E9FX74_DAPPU_HAT1 | -27.05 | 27 | 80 | 120 | 18 | 6 | 335 | 450 | 4087436 | 4087321 | 79 |
| | | -25.69 | 2 | 80 | 118 | 14 | 10 | 199 | 311 | 1578414 | 1578302 | 77 |
| | | -23.13 | 87 | 78 | 124 | 19 | 7 | 346 | 465 | 363144 | 363025 | 73 |
| **Histone Acetyltransferase KAT5** | G6CZK5_DANPL_Tip60 | -18.64 | 2 | 82 | 82 | 13 | 2 | 1002 | 1082 | 1578382 | 1578302 | 68 |
| | | -17.33 | 76 | 77 | 131 | 25 | 7 | 1181 | 1308 | 1563851 | 1563724 | 66 |
| | | -12.92 | 27 | 77 | 107 | 22 | 3 | 1078 | 1182 | 4089896 | 4089792 | 59 |
| | | -7.13 | 126 | 83 | 59 | 7 | 3 | 255 | 310 | 1614809 | 1614867 | 51 |
| **Histone Deacetylase 1** | E9GZX2_DAPPU_HisDe | -112.40 | 777 | 78 | 366 | 73 | 10 | 259 | 619 | 354285 | 354644 | 203 |
| | | -4.51 | 65 | 89 | 37 | 4 | 0 | 1158 | 1194 | 1880842 | 1880878 | 47 |
| | | -3.24 | 51 | 89 | 36 | 4 | 0 | 448 | 483 | 2010116 | 2010151 | 45 |
| **MYST Histone Acetyltransferase** | B7Q9K5_IXOSC_MYSTHis | -31.54 | 27 | 76 | 173 | 33 | 8 | 694 | 862 | 4101395 | 4101227 | 86 |
| | | -31.40 | 2 | 77 | 167 | 31 | 7 | 775 | 937 | 1577813 | 1577651 | 85 |
| | | -3.44 | 8 | 76 | 85 | 18 | 2 | 717 | 800 | 363271 | 363188 | 45 |
| **P300/CBP-Associated Factor** | E9H234_DAPPU_PCAF | -23.73 | 1422 | 80 | 122 | 20 | 6 | 1229 | 1347 | 37996 | 38114 | 75 |
| | | -20.49 | 686 | 82 | 97 | 15 | 4 | 901 | 995 | 518854 | 518948 | 71 |
| | | -18.42 | 14 | 93 | 45 | 3 | 0 | 1915 | 1959 | 2970163 | 2970207 | 68 |
| | | -13.12 | 1491 | 86 | 56 | 8 | 0 | 1393 | 1448 | 100810 | 100755 | 60 |

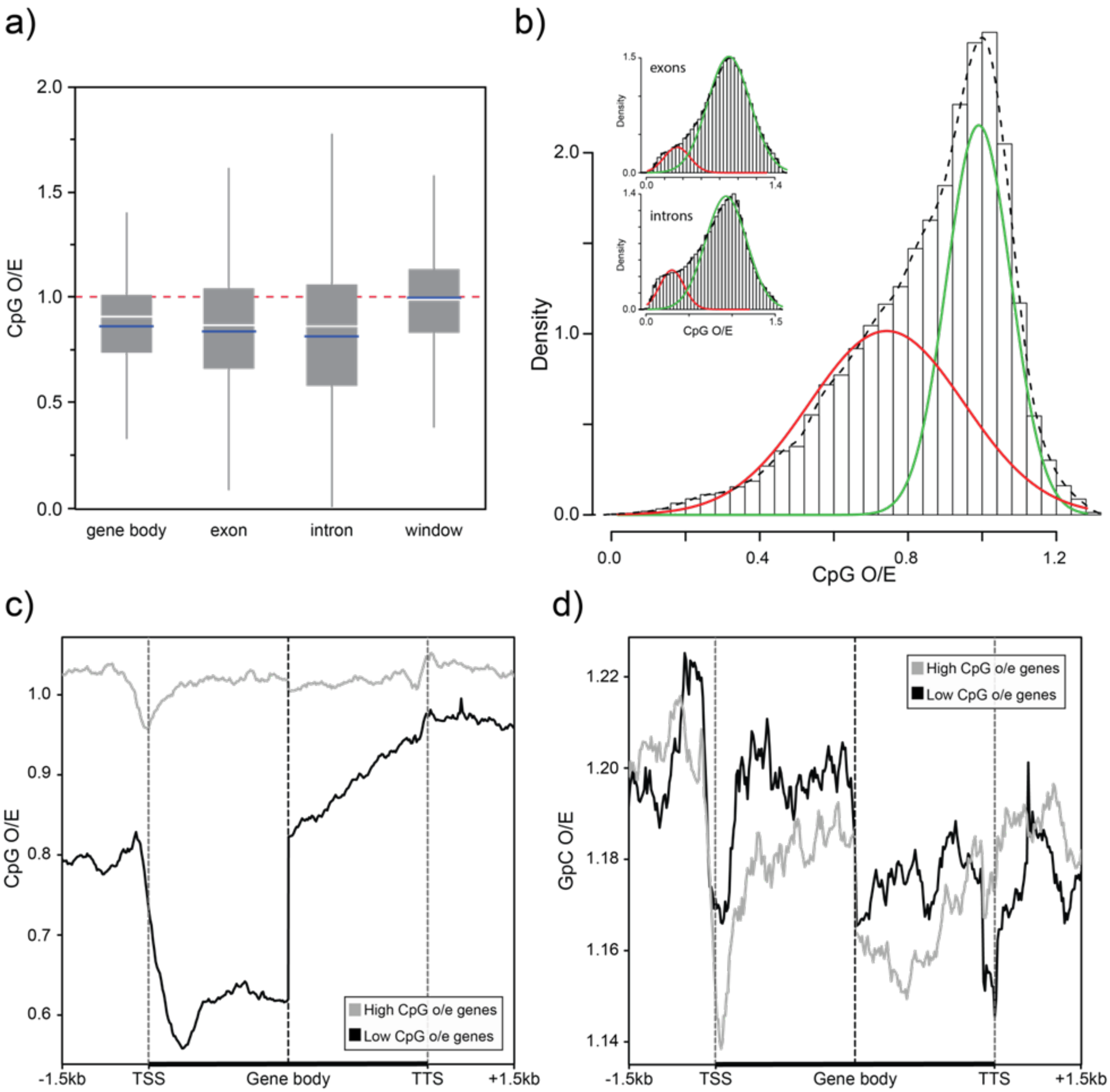



**Figure S3.1: CpG depletion indicates DNA methylation in *H*yalella azteca.** a) CpG o/e for gene bodies (full gene frame), exons, introns, and non-genic genomic windows, showing that genic features depart from the null expectation of CpG o/e == 0 (as shown by genomic windows; blue lines: mean). b) density plot of CpG o/e values for gene bodies with bimodal fit approximating putatively methylated (red) and unmethylated (green) fractions of genes (inset: the same for exons and introns). c-d) Metaplots of CpG o/e and GpC o/e for the first and last 2kb of gene bodies, as well as 1.5kb up- and downstream of genes; genes were split into either "high" or "low" CpG o/e categories for plotting, in order to illustrate strong positional bias of CpG depletion in low CpG o/e genes indicative of 5'-proximal DNA methylation.

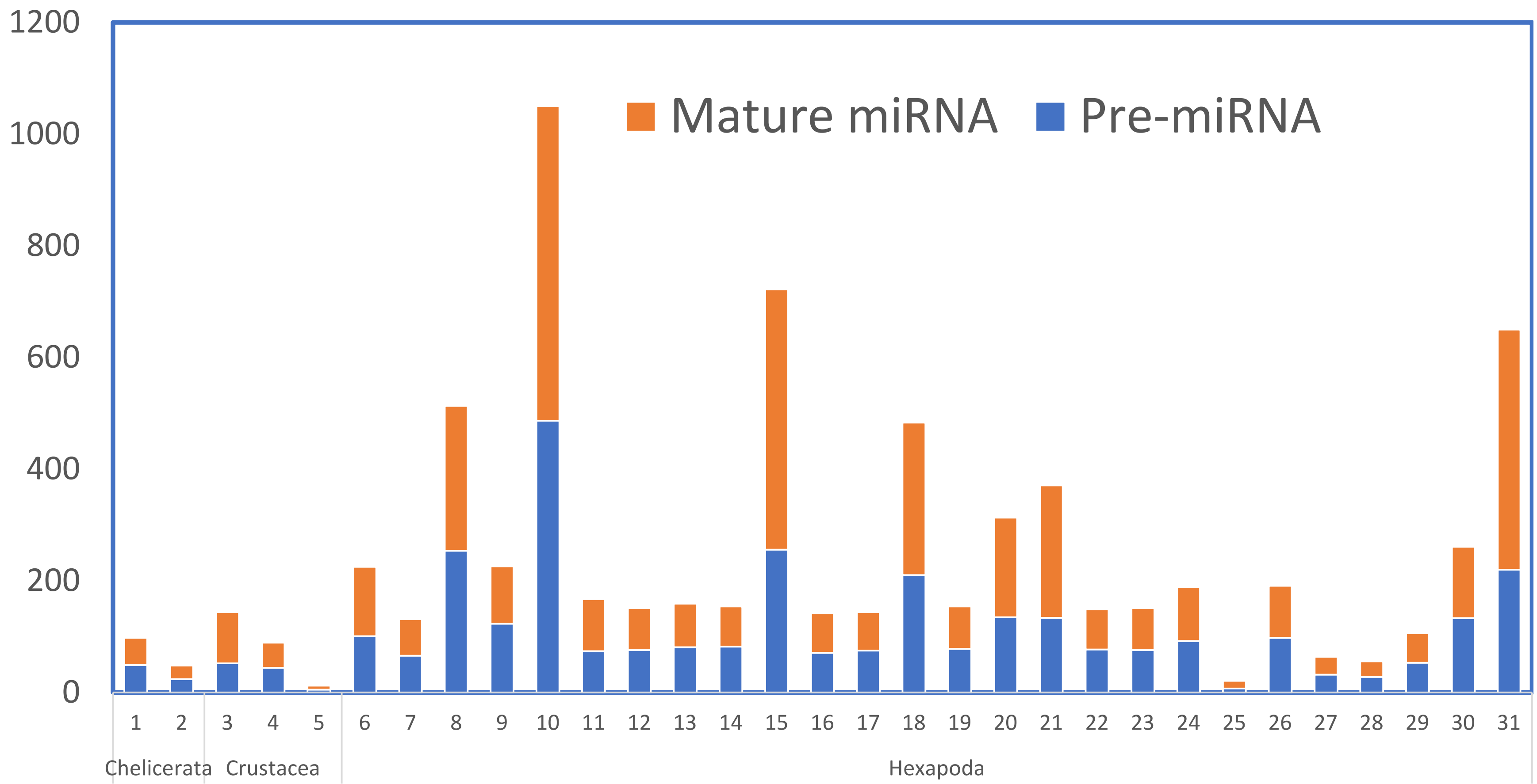



**Figure S3.2. Total number of mature and pre-miRNAs reported from Arthropods.** (source: www.mirbase.org; searched October 2017). List of species: 1. *Ixodes scapularis* 2. *Rhipicephalus microplus* 3. *Tetranychus urticae* 4. *Daphnia pulex* 5. *Marsupenaeus japonicas* 6. *Aedes aegypti* 7. *Anopheles gambiae* 8. *Apis mellifera* 9. *Acyrthosiphon pisum* 10. *Bombyx mori* 11. *Culex quinquefasciatus* 12. *Drosophila ananassae* 13. *D. erecta* 14. *D. erecta* 15. *D. grimshawi* 16. *D. melanogaster* 17. *D. mojavensis* 18. *D. persimilis* 19. *D. pseudoobscura* 20. *D. sechellia* 21. *D. simulans* 22. D. *virilis* 23. *D. willistoni* 24. *D. yakuba* 25. *Heliconius melpomene* 26. *Locusta migratoria* 27. *Manduca sexta* 28. *Nasonia giraulti* 29. N. *longicornis* 30. *N. vitripennis* 31. *Plutella xylostella* 32. *Tribolium casta*

## S3. References:


(1)  Glastad, K.; Hunt, B.; Yi, S.; Goodisman, M. DNA methylation in insects: on the brink of the epigenomic era. *Insect Molecular Biology* **2011,** *20*, (5), 553-565.

(2)  Elango, N.; Kim, S.-H.; Vigoda, E.; Soojin, V. Y.; Program, N. C. S. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. *PLoS Computational Biology* **2008,** *4*, (2), e1000015.

(3)  Choudhuri, S. Small noncoding RNAs: biogenesis, function, and emerging significance in toxicology. *Journal of Biochemical and Molecular Toxicology* **2010,** *24*, (3), 195-216.

(4)  Wheeler, B. M.; Heimberg, A. M.; Moy, V. N.; Sperling, E. A.; Holstein, T. W.; Heber, S.; Peterson, K. J. The deep evolution of metazoan microRNAs. *Evolution & Development* **2009,** *11*, (1), 50-68.

(5)  Chen, S.; McKinney, G. J.; Nichols, K. M.; Colbourne, J. K.; Sepúlveda, M. S. Novel cadmium responsive microRNAs in Daphnia pulex. *Environmental Science & Technology* **2015,** *49*, (24), 14605-14613.

(6)  Chen, S.; Nichols, K. M.; Poynton, H. C.; Sepúlveda, M. S. MicroRNAs are involved in cadmium tolerance in Daphnia pulex. *Aquatic Toxicology* **2016,** *175*, 241-248.

(7)  Benaglia, T.; Chauveau, D.; Hunter, D.; Young, D. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software* **2009,** *32*, (6), 1-29.

(8)  Chen, S.; McKinney, G. J.; Nichols, K. M.; Sepúlveda, M. S. In silico prediction and in vivo validation of Daphnia pulex microRNAs. *PloS One* **2014,** *9*, (1), e83708.

(9)  Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **1990,** *215*, (3), 403-410.

(10)  Gruber, A. R.; Lorenz, R.; Bernhart, S. H.; Neuböck, R.; Hofacker, I. L. The vienna RNA websuite. *Nucleic Acids Research* **2008,** *36*, (suppl_2), W70-W74.

(11)  Hunt, B. G.; Glastad, K. M.; Yi, S. V.; Goodisman, M. A. The function of intragenic DNA methylation: insights from insect epigenomes. In Oxford University Press: 2013.

(12)  Glastad, K. M.; Gokhale, K.; Liebig, J.; Goodisman, M. A. The caste-and sex-specific DNA methylome of the termite Zootermopsis nevadensis. *Scientific Reports* **2016,** *6*, 37110.

## S4.1. Chemoreceptors

Hugh M. Robertson
Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL U.S.A.

### Introduction

In non-insect arthropods there are two major families of chemoreceptors, the Gustatory Receptor (GR) family that forms the ancient base of the insect chemoreceptor superfamily of seven-transmembrane ligand-gated ion channels, which includes the insect-specific Odorant Receptor (OR) family [1-4], and the Ionotropic Receptor (IR) family that is a variant lineage of the otherwise highly conserved and widespread ionotropic glutamate receptors in animals and beyond.[4-7] These proteins are expressed in the chemosensory neurons of chemosensilla on several arthropod body parts like antennae, mouthparts, legs, wings, and genitalia, and mediate the specificity and sensitivity of arthropod chemoreception, that is, taste and smell.[8] While these gene families have been described from numerous insects and a few other arthropods, the only detailed description from a crustacean genome sequence is from *Daphnia pulex*.[6, 9] Eyun et al.[4] described some members of both families from several additional crustaceans, especially two copepods including *Eurytemora affinis*, the genome sequence of which they report. Other crustacean IR sequences are available from transcriptome studies of the American lobster *Homarus americanus*,[10] the Caribbean spiny lobster *Panulirus argus*,[11] and two hermit crabs *Pagurus bernhardus* and *Coenobita clypeatus*,[12, 13] but these are mostly too incomplete to usefully employ here or only represent additional versions of some of the most conserved IRs.

### Methods

The GR and IR families were manually annotated by pursuing TBLASTN output from searches of the genome scaffolds with the *D. pulex* and other arthropod proteins, both using the Apollo tool available with this genome project at the i5k Workspace and a local text editor as described previously (e.g. Penalva-Arana et al.[9]). All gene models were entered into the Apollo manual annotation as best possible given the fragmented nature of the genome assembly (see details below), however there are a few minor changes to models that were made after freezing of the Apollo manual annotations. All protein translations including repaired models and pseudogenes (neither available from the Apollo browser) are provided in Supplemental File S6.2. Iterative searches with E values up to 1000 were used to search exhaustively for additional family members, with visual recognition of relatively conserved domains like the TYhhhhhQF domain in the seventh transmembrane domain of the GRs.[3] Repairs were attempted on all partial gene models using a combination of raw genome and RNAseq reads from the SRA database at NCBI (an extreme example is provided by Ir8a, a highly conserved protein encoded by 19 exons, two of which are not assembled in the draft genome assembly, while the remainder are in four misordered contigs in Scaffold250). Pseudogenes were translated as best possible including stop codons and accommodating frameshifts and indels, and to be included in the dataset had to encode at least half of an intact related chemoreceptor. The *E. affinis* genes were annotated as above using the Apollo browser for that species at the i5k Workspace, while the *D. pulex* IR genes were built manually in TEXTWRANGLER and are not available from Apollo, but the encoded proteins are provided in Supplemental File S.5.2.

Proteins were aligned for each family for these three crustaceans and selected additional family members from other species for relevant comparisons using default settings in CLUSTALX v2.0.[14] The additional sequences came from *D. melanogaster*,[1, 5] the honey bee *Apis mellifera*,[15] the pea aphid *Aphis pisum*,[6] the human body louse *Pediculus humanus*,[6] the dampwood termite *Zootermopsis nevadensis*,[16] and the damselfly *Calopteryx splendens*.[17] Positions poorly represented in the alignments (generally the variable length N- and C-termini, plus a few internal positions in intra- or extra-cellular loops), were removed using TRIMAL v1.4,[18] using the "gappyout" option. Within the IR family three *H. azteca* sequences were particularly problematic for the alignments, specifically Ir200, 201, and 213, because of insertions between the first two transmembrane domains that are usually quite close to each other. These insertions were removed for the alignments based on alignments generated by PSI-BLASTP searches of the non-redundant proteins at NCBI, restricted to Arthropoda. Phylogenetic analysis was performed using maximum likelihood with default settings on the PHYML v3.0 server.[19] Tree figures were prepared using FIGTREE v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) and ADOBE ILLUSTRATOR.

**Results.** *Daphnia pulex* has 58 GRs of which two are pseudogenes,[9] however during this study the model for Gr58 was improved and a related pseudogene (Gr59P) was added to the compilation, for a total of 59 GRs with three being pseudogenic (an open intron was also introduced into Gr53, and the updated sequences of these three GRs are provided in Supplemental File S.5.2). This number is comparable to the 68 GRs in *Drosophila melanogaster*,[1] and many other arthropods including other non-insects like the centipede *Strigamia maritima* with 77 genes,[20] the tick *Ixodes scapularis* with 62,[21] and the predatory mite *Metaseiulus occidentalis* with 64,[22] although some insects have larger numbers of GRs. Manual annotation of the GR family in *H. azteca* revealed a considerably larger family of 155 genes, of which 41 are pseudogenes. In addition, two genes exhibit an unusual form of alternative splicing common to GR genes in arthropods in which separate, and often quite different, first long exons are spliced into three shared short exons encoding the somewhat conserved C-terminal region, yielding three additional GR proteins (Gr65a-c and 74a/b). This manual annotation effort was made complex by the fragmented nature of the genome assembly, however, so 65 of the apparently intact genes were fractured in some way (for example, with exons missing in gaps or off ends of scaffolds), of which 17 were partially or fully repaired using raw genomic and/or RNAseq reads, while one gene model spans two scaffolds. Some of these might not be intact genes, but in addition there were many fragments of genes too short to include in this analysis that might actually represent intact genes, so these two categories probably balance out and a total of 114 functional GRs likely reflects the coding capacity of this genome. This is approximately twice the number in the *D. pulex* genome. In addition, the GR genes of *E. affinis* were annotated, increasing them from 9 reported in Eyun et al.[4] to 67 genes, six of which are pseudogenes. These *E. affinis* GR genes have several unusual properties, including most being singletons in different scaffolds, with many of them occurring within introns of larger genes, suggesting that there has been a high level of genome flux or gene movement in this copepod genome. In addition, as initially discovered with the IRs discussed below,[23] several genes have GA intron donors.

Phylogenetic analysis of these crustacean GRs (Figure S4.1.1) reveals that *H. azteca* and *D. pulex* each have two clear members of the conserved subfamily of sugar receptors (HaztGr1/2 and DpulGr55/56), indicated both by their confident clustering with representative sugar receptors from other insects and by their having a glutamic acid (E) after the TY in the conserved C-terminal TYhhhhhQF domain.[24] These pairwise duplications occurred independently in the two crustaceans, implying that the basal condition in crustaceans might have been a single sugar receptor (although it is also possible that DpuGr57 is a sugar receptor as it clusters near them and has the TYE motif). While sugar receptors in insects are generally considered to function as dimers, and most insects with sugar receptors have more than one, recently the basal insect lineage of the damselfly *Calopteryx splendens* was found to have only one sugar receptor,[17] so it is possible that sugar receptors functioned as a monomer in basal arthropods. It is somewhat surprising that *E. affinis* does not have a member of this subfamily. These are the only crustacean GRs with convincing relationships with insect GRs with known functional roles. Thus none of these crustacean GRs cluster confidently with the DmelGr43a fructose receptor, which is also reasonably well-conserved in insects, or a subset of the best-known and most-conserved bitter taste receptors in *Drosophila* (Figure S4.1.1). The phylogenetic tree reveals four expansions specific to *D. pulex* (in addition to DpulGr57), with both long and short terminal branches suggesting a continuous process of gene duplication. The tree also reveals four major expansions in *H. azteca* (as well as the divergent HaztGr154 and Gr155), but with an unusual pattern of very recent duplications within each of these expansions. Closer examination reveals that these duplications involved "segmental duplications" of several sets of neighboring GRs. While segmental duplications in draft genome projects can sometimes be artifacts of the assembly, these appear to be real because the depth of raw genome reads in the Short Read Archive at NCBI is approximately twice that of singleton genes. Because the genes were named for their locations within scaffolds, genes with quite separate names are close neighbors in the tree. These segmentally-duplicated GRs also contain the vast majority of the 41 pseudogenes, suggesting that a recent expansion of the GR repertoire has been partially relegated to the evolutionary trash bin. Finally, the *E. affinis* GRs form a single lineage-specific clade.

The IR family was newly recognized[5] at the time of the *D. pulex* chemoreceptor analysis[9] so was not included therein. It also consists of around 60 genes in *D. melanogaster*[6] and is of comparable size in many other arthropods, although it shows major expansions in some, most greatly the dampwood termite *Zootermopsis nevadensis* with 150 IRs.[16] Unlike the GRs, the IR family has several members that have considerable conservation across insects and even other arthropods, and those are usually named for their *Drosophila* orthologs. In *H. azteca* these are Ir8a, 25a, 76b, and 93a. Ir8a and 25a are well known as co-receptors that function along with other IRs,[7, 25] while Ir76b is another co-receptor involved in sensing salt, amines, and amino acids,[26-30] and Ir93a has been identified as a temperature and humidity sensor in flies.[31-33] Another 114 more divergent IRs were identified in the *H. azteca* genome, for a total of 118 IRs. Following the example begun with the termite[16] and continued with the predatory mite genome,[22] these more divergent IRs are named in a series from 101-213, which avoids confusion with the *Drosophila* IRs, which were named for their cytological location and hence only go up to Ir100a. Only two of these *H. azteca* IRs are pseudogenes, but like the GRs, 65 were partial in some fashion in the assembly, of which 24 were partially or fully repaired, while one gene spans two scaffolds. Again, some of these partial gene models might not represent intact genes, while some of the many short fragments of genes might in fact represent intact genes, so a total of 116 functional IRs is a reasonable estimate.

Croset et al.[6] included 85 IRs from *D. pulex* in their compilation of the IR family from arthropods, however their proteins were largely based on the annotated proteins that were produced by automated gene modeling on a genome-wide scale, and chemoreceptors are notoriously recalcitrant to such automated annotation as they are generally rapidly evolving with usually low expression levels. Manual annotation of the IR family in this genome led to repair of the Ir93a model and recognition of Ir76b (Ir304 in Croset et al.[6]), but the Ir8a gene remains absent, and identified 151 divergent genes, of which Croset et al.[6] had 82 (although many were partial models). These 151 genes were re-named in a similar way to the *H. azteca* IRs, from DpulIr101-251 (with agreement from R. Benton, personal communication), and their corresponding numbers from Croset et al.[6] are provided along with their protein sequences in Supplemental File S.5.2. 26 of these are pseudogenes, while 12 were incomplete in the genome assembly, of which 7 were repaired, so the total functional IR count in *D. pulex* is 128.

Eyun et al.[4] reported the presence of Ir8a, 25a, 76b, and 93a, along with a possible ortholog of Ir21a, and three divergent IRs from *E. affinis*. Having recognized that this genome has an unusually high frequency of non-canonical GA and GG intron donors,[23] it was possible to complete the gene models for all of these IRs, and add 14 more. For consistency with the *H. azteca* and *D. pulex* IRs, the three divergent IRs from Eyun et al.[4] were renamed Ir101-103, and the 14 new ones are Ir104-117 (all 22 Eaff IR proteins are provided in Supplemental File S.5.2). This total of 22 IRs is relatively low, but extensive efforts to identify additional IRs were unsuccessful, although it must be noted that these *E. affinis* IRs are extremely divergent and are encoded by many short exons, making them difficult to find in TBLASTN searches of a genome sequence.

Phylogenetic analysis of these crustacean IRs (Figure S4.1.2) reveals that in addition to the highly conserved Ir25a and Ir8a (missing from *D. pulex*) lineages, the Ir93a lineage is reasonably well conserved, but the crustacean Ir76b proteins are rather divergent, perhaps explaining why Croset et al.[6] did not recognize their Ir304 as being the Ir76b ortholog (it was also recognized by Eyun et al.[4]). Beyond these four lineages that retain the names of their *Drosophila* orthologs, there are no other convincing orthologous relationships even between the IRs from these three crustaceans, let alone to other arthropod IRs, as shown by inclusion in the tree analysis of representatives of the fairly well-conserved Ir21a, 40a, and 68a lineages found as far basally within the insects as *C. splendens*[17] and involved in temperature and humidity sensing.[31-34] The protein named Ir21a by Eyun et al.[4] clusters weakly with Ir21a from insects, along with the Ir40a and Ir68a lineages, but this clade also includes multiple *H. azteca* and *D. pulex* receptors. As with the GRs, most of the rest of these crustacean IRs form large, and largely species-specific, subfamilies. The *H. azteca* IRs form several large subfamilies in the tree, with two *D. pulex* IR lineages (DpulIr101, and Ir102-106) clustering with them, as do most of the *E. affinis* proteins. The remainder of the *D. pulex* IRs (DpulIr107-151) form a discrete lineage.

**Discussion.** This analysis of the chemoreceptor genes in these three crustaceans reveals that, in addition to the presence of two candidate GR sugar receptors and the conserved Ir8a (missing from *D. pulex*), 25a, 76b, and 93a lineages that are implicated in perception of salt, amines, amino acids, and temperature in insects, they contain large expanded subfamilies of GRs and IRs, generally distinctive to each crustacean. *H. azteca* has about twice the number of GRs as *D. pulex* and *E. affinis* although many of the most recent gene duplicates are pseudogenes, while *H. azteca* and *D. pulex* have comparable numbers of divergent IRs, but *E. affinis* has relatively few of them. It is hard to speculate about the roles of these expanded GR and IR families in the chemosensory capabilities and chemical ecology of these crustaceans, however in insects the vast majority of the divergent IRs are involved in taste,[35, 36] so presumably these and most GRs mediate perception of diverse chemical cues about food quality and other environmentally relevant chemicals.

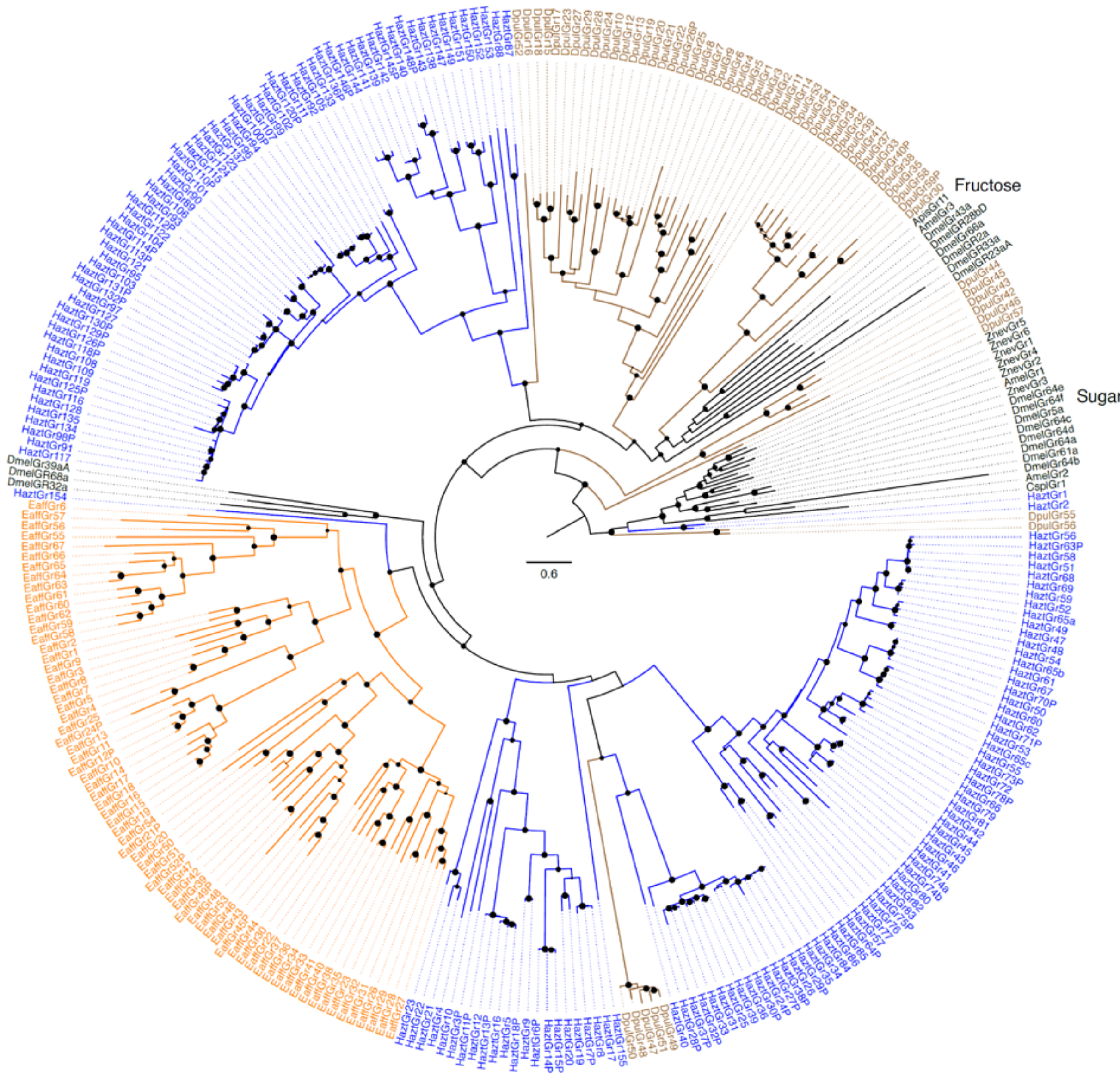



**Figure S4.1.1.** Phylogenetic relationships of the Gustatory Receptor (GR) family of *Hyalella azteca*, *Daphnia pulex,* and *Eurytemora affinis*. The sugar receptors including HaztGr1/2 and DpulGr55/56 were declared the outgroup to root the tree. This sugar lineage, as well as the DmelGr43a fructose receptor lineage, are indicated outside the circle of protein names. Branch support from the approximate likelihood-ratio test (aLRT) is shown as filled circles ranging from 0-1. The scale bar is substitutions per site. The *H. azteca*, *D. pulex,* and *E. affinis* names are in blue, brown, and orange, respectively, as are branches leading to them to emphasize discrete lineages. GRs from other species are in black. Suffix P after a name indicates a pseudogene.

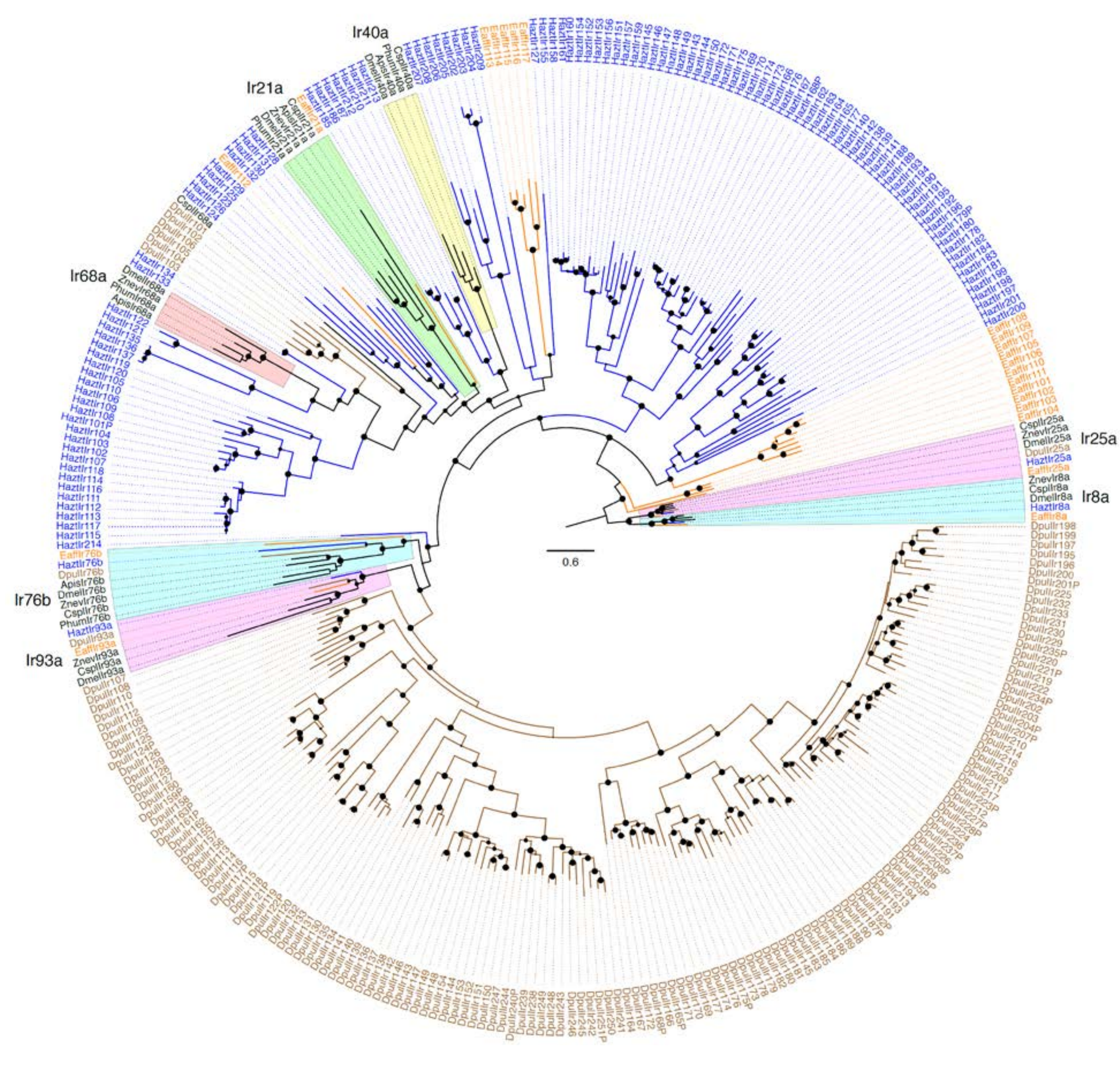



**Figure S4.1.2**. Phylogenetic relationships of the Ionotropic Receptor (IR) family of *Hyalella azteca*, *Daphnia pulex*, and *Eurytemora affinis.* The Ir8a and 25a receptors were declared the outgroup to root the tree, as they are most similar to the ionotropic glutamate receptors from which the IRs evolved. Major lineages are shaded and indicated with their names from *D. melanogaster* outside the circle of protein names. Other details as in Figure S4.1.1 legend.

**S4.1. References:**


(1) Robertson, H. M.; Warr, C. G.; Carlson, J. R. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences* **2003,** *100*, (suppl 2), 14537-14542.

(2) Saina, M.; Busengdal, H.; Sinigaglia, C.; Petrone, L.; Oliveri, P.; Rentzsch, F.; Benton, R. A cnidarian homologue of an insect gustatory receptor functions in developmental body patterning. *Nature Communications* **2015,** *6: 6243*, 1-12.

(3) Robertson, H. M. The insect chemoreceptor superfamily is ancient in animals. *Chemical Senses* **2015,** *40*, (9), 609-614.

(4) Eyun, S.-i.; Young Soh, H.; Posavi, M.; Munro, J. B.; Hughes, D. S.; Murali, S. C.; Qu, J.; Dugan, S.; Lee, S. L.; Chao, H. Evolutionary history of chemosensory-related gene families across the Arthropoda. *Molecular Biology and Evolution* **2017**, msx147.

(5) Benton, R.; Vannice, K. S.; Gomez-Diaz, C.; Vosshall, L. B. Variant ionotropic glutamate receptors as chemosensory receptors in Drosophila. *Cell* **2009,** *136*, (1), 149-162.

(6) Croset, V.; Rytz, R.; Cummins, S. F.; Budd, A.; Brawand, D.; Kaessmann, H.; Gibson, T. J.; Benton, R. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genetics* **2010,** *6*, (8), e1001064.

(7) Rytz, R.; Croset, V.; Benton, R. Ionotropic receptors (IRs): chemosensory ionotropic glutamate receptors in Drosophila and beyond. *Insect Biochemistry and Molecular Biology* **2013,** *43*, (9), 888-897.

(8) Joseph, R. M.; Carlson, J. R. Drosophila chemoreceptors: a molecular interface between the chemical world and the brain. *Trends in Genetics* **2015,** *31*, (12), 683-695.

(9) Peñalva-Arana, D. C.; Lynch, M.; Robertson, H. M. The chemoreceptor genes of the waterflea Daphnia pulex: many Grs but no Ors. *BMC Evolutionary Biology* **2009,** *9*, (1), 79.

(10) Hollins, B.; Hardin, D.; Gimelbrant, A. A.; McClintock, T. S. Olfactory-enriched transcripts are cell-specific markers in the lobster olfactory organ. *Journal of Comparative Neurology* **2003,** *455*, (1), 125-138.

(11) Corey, E. A.; Bobkov, Y.; Ukhanov, K.; Ache, B. W. Ionotropic crustacean olfactory receptors. *PloS One* **2013,** *8*, (4), e60551.

(12) Groh, K. C.; Vogel, H.; Stensmyr, M. C.; Grosse-Wilde, E.; Hansson, B. S. The hermit crab's nose—antennal transcriptomics. *Frontiers in Neuroscience* **2013,** *7*.

(13) Groh-Lunow, K. C.; Getahun, M. N.; Grosse-Wilde, E.; Hansson, B. S. Expression of ionotropic receptors in terrestrial hermit crab's olfactory sensory neurons. *Frontiers in Cellular Neuroscience* **2014,** *8*.

(14) Larkin, M. A.; Blackshields, G.; Brown, N.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007,** *23*, (21), 2947-2948.

(15) Robertson, H. M.; Wanner, K. W. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. *Genome Research* **2006,** *16*, (11), 1395-1403.

(16) Terrapon, N.; Li, C.; Robertson, H. M.; Ji, L.; Meng, X.; Booth, W.; Chen, Z.; Childers, C. P.; Glastad, K. M.; Gokhale, K. Molecular traces of alternative social organization in a termite genome. *Nature Communications* **2014,** *5*, 3636.

(17) Ioannidis, P.; Simao, F. A.; Waterhouse, R. M.; Manni, M.; Seppey, M.; Robertson, H. M.; Misof, B.; Niehuis, O.; Zdobnov, E. M. Genomic Features of the Damselfly *Calopteryx splendens* Representing a Sister Clade to Most Insect Orders. *Genome Biology and Evolution* **2017,** *9*, (2), 415-430.

(18) Capella-Gutiérrez, S.; Silla-Martínez, J. M.; Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **2009,** *25*, (15), 1972-1973.

(19) Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **2010,** *59*, (3), 307-321.
(20) Chipman, A. D.; Ferrier, D. E.; Brena, C.; Qu, J.; Hughes, D. S.; Schröder, R.; Torres-Oliva, M.; Znassi, N.; Jiang, H.; Almeida, F. C. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. *PLoS Biology* **2014,** *12*, (11), e1002005.
(21) Gulia-Nuss, M.; Nuss, A. B.; Meyer, J. M.; Sonenshine, D. E.; Roe, R. M.; Waterhouse, R. M.; Sattelle, D. B.; de La Fuente, J.; Ribeiro, J. M.; Megy, K. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. *Nature Communications* **2016,** *7*, 10507.
(22) Hoy, M. A.; Waterhouse, R. M.; Wu, K.; Estep, A. S.; Ioannidis, P.; Palmer, W. J.; Pomerantz, A. F.; Simão, F. A.; Thomas, J.; Jiggins, F. M. Genome sequencing of the phytoseiid predatory mite *Metaseiulus occidentalis* reveals completely atomized Hox genes and superdynamic intron evolution. *Genome Biology and Evolution* **2016,** *8*, (6), 1762-1775.
(23) Robertson, H. M. Noncanonical GA and GG 5′ Intron Donor Splice Sites Are Common in the Copepod Eurytemora affinis. *G3: Genes, Genomes, Genetics* **2017,** *7*, (12), 3967-3969.
(24) Kent, L. B.; Robertson, H. M. Evolution of the sugar receptors in insects. *BMC Evolutionary Biology* **2009,** *9*, (1), 41.
(25) Ai, M.; Blais, S.; Park, J.-Y.; Min, S.; Neubert, T. A.; Suh, G. S. Ionotropic glutamate receptors IR64a and IR8a form a functional odorant receptor complex in vivo in Drosophila. *Journal of Neuroscience* **2013,** *33*, (26), 10741-10749.
(26) Zhang, Y. V.; Ni, J.; Montell, C. The molecular basis for attractive salt-taste coding in Drosophila. *Science* **2013,** *340*, (6138), 1334-1338.
(27) Croset, V.; Schleyer, M.; Arguello, J. R.; Gerber, B.; Benton, R. A molecular and neuronal basis for amino acid sensing in the Drosophila larva. *Scientific Reports* **2016,** *6*, 34871.
(28) Hussain, A.; Zhang, M.; Üçpunar, H. K.; Svensson, T.; Quillery, E.; Gompel, N.; Ignell, R.; Kadow, I. C. G. Ionotropic chemosensory receptors mediate the taste and smell of polyamines. *PLoS Biology* **2016,** *14*, (5), e1002454.
(29) Ahn, J.-E.; Chen, Y.; Amrein, H. O. Molecular basis of fatty acid taste in Drosophila. *eLife* **2017,** *6*, e30115.
(30) Ganguly, A.; Pang, L.; Duong, V.-K.; Lee, A.; Schoniger, H.; Varady, E.; Dahanukar, A. A molecular and cellular context-dependent role for Ir76b in detection of amino acid taste. *Cell Reports* **2017,** *18*, (3), 737-750.
(31) Enjin, A.; Zaharieva, E. E.; Frank, D. D.; Mansourian, S.; Suh, G. S.; Gallio, M.; Stensmyr, M. C. Humidity sensing in Drosophila. *Current Biology* **2016,** *26*, (10), 1352-1358.
(32) Knecht, Z. A.; Silbering, A. F.; Ni, L.; Klein, M.; Budelli, G.; Bell, R.; Abuin, L.; Ferrer, A. J.; Samuel, A. D.; Benton, R. Distinct combinations of variant ionotropic glutamate receptors mediate thermosensation and hygrosensation in Drosophila. *eLife* **2016,** *5*, e17879.
(33) Knecht, Z. A.; Silbering, A. F.; Cruz, J.; Yang, L.; Croset, V.; Benton, R.; Garrity, P. A. Ionotropic Receptor-dependent moist and dry cells control hygrosensation in Drosophila. *eLife* **2017,** *6*, e26654.
(34) Ni, L.; Klein, M.; Svec, K. V.; Budelli, G.; Chang, E. C.; Ferrer, A. J.; Benton, R.; Samuel, A. D.; Garrity, P. A. The ionotropic receptors IR21a and IR25a mediate cool sensing in Drosophila. *eLife* **2016,** *5*, e13254.
(35) Koh, T.-W.; He, Z.; Gorur-Shandilya, S.; Menuz, K.; Larter, N. K.; Stewart, S.; Carlson, J. R. The Drosophila IR20a clade of ionotropic receptors are candidate taste and pheromone receptors. *Neuron* **2014,** *83*, (4), 850-865.
(36) Stewart, S.; Koh, T.-W.; Ghosh, A. C.; Carlson, J. R. Candidate ionotropic taste receptors in the Drosophila larva. *Proceedings of the National Academy of Sciences* **2015,** *112*, (14), 4195-4201.

## S4.2. Cuticle proteins

Andrew J. Rosendale and Joshua B. Benoit

Department of Biological Sciences, McMicken College of Arts and Sciences, University of Cincinnati, Cincinnati, OH 45221-0006

Correspondence to: benoitja@ucmail.uc.edu

**Introduction**

The arthropod cuticle provides protection and the mechanical and structural support that has aided the success of these organisms in diverse habitats.[1] The cuticles are primarily composed of chitin fibers embedded within a protein matrix[1] that consists of numerous cuticle proteins (reviewed in Willis et al.[2]). There is diversity in the number and type of cuticle proteins present among the arthropods with individual species containing a subset of approximately 12 families of cuticle protein.[3, 4] Families are generally classified by the presence of conserved motifs, with the CPR family, characterized by the "R&R Consensus" domain,[5] being the largest. Certain families, such as CPCFC and CPLCW are restricted to certain orders and/or even smaller taxonomic groups, whereas families including CPR, CPAP1 and CPAP3 are present in most arthropods.[3, 4] However, understanding of the diversity of cuticle proteins is currently evolving as more arthropod genomes are sequenced and expression patterns of cuticle protein genes are examined.

**Methods**

The official gene set for *Hyalella azteca* (maker annotation gene set version 5.3 obtained from the Baylor College of Medicine Human Genome Sequencing Center; BCM-HGSC) was searched (BLASTp[6]) using sequence motifs that are characteristic of several families of cuticle proteins.[2] Genes identified as potential cuticle proteins were analyzed with CutProtFam-Pred, a cuticular protein family prediction tool described in Ioannidou et al.,[3] to assign genes to gene families. To find the closest putative homolog to cuticle protein genes from *H. azteca*, genes were searched against the non-redundant (nr) or RefSeq protein sets downloaded from the National Center for Biotechnology Information (NCBI), or the official gene set for *Daphnia pulex* (nr), *Drosophila melanogaster* (Refseq), *Tribolium castaneum* (Refseq), *Acyrthosiphon pisum* (Refseq), *Bombyx mori* (Refseq), *Pediculus humanus corporis* (Refseq), *Apis mellifera* (Refseq), *Cimex lectularius* (maker v. 5.3, BCM-HGSC[7]), *Ixodes scapularis* (Refseq), *Strigamia maritime* (downloaded from the European Bioinformatics Institute; Chipman et al.[8]). The protein sequence with the lowest e-value was considered the closest putative homolog. Sequences of TWDL genes from *H. azteca* were aligned with MUSCLE[9] to create a phylogenetic tree.

**Results**

Ninety-two genes encoding for putative cuticle proteins were identified in *H. azteca* by searching the genome with sequence motifs characteristic of different cuticle protein families as established by Willis.[4] CutProtFam-Pred[3] was employed to assign these genes to one of five families (CPR, CPAP1, CPAP3, and TWDL; Table S4.2.1). The total number of cuticle protein genes identified in *H. azteca* (92) is much less than *D. pulex* (321), another crustacean, but is within the range of that observed in other arthropods (Table S4.2.2). As with most arthropods, the number of CPR genes (60), including RR-1 (soft cuticle), RR-2 (hard cuticle), and unclassifiable types, constituted the largest group of cuticle protein genes in the *Hyalella* genome. The number of genes in the protein families CPR, CPAP1, and CPAP3 were similar to the number in other arthropods (Willis[4]; Table S4.2.2). However, the 12 genes in the TWDL family were greater than the number found in most insect orders, and this family seems to be missing entirely from other non-insect arthropods (Table S4.2.2, Fig. S4.2.1).

| CPR[a] | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| RR-1 | RR-2 | Uncl | CPAP1 | CPAP3 | TWDL | Total |
| 10 | 16 | 34 | 13 | 7 | 12 | 92 |

**Table S4.2.1.** Number of genes identified as putative cuticle proteins per family in the genome of *Hyalella azteca*

[a]Sequences that scored above the assigned cutoffs for the RR-1 and RR-2 models were classified as the corresponding type, whereas sequences with scores below the assigned cutoffs but above 0 were characterized as “unclassified” (Uncl). For more information, see Ioannidou et al.[3]

| Subphylum | Class | Order | Species | CPR | CPAP1 | CPAP3 | CPCFC | CPF | CPLCA | CPLCG | CPLCP | TWDL | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chelicerata | Arachnida | Ixodida | *Ixodes scapularis* [a] | 103 | 15 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 121 |
| Myriapoda | Chilopoda | Geophilomorpha | *Strigamia maritime* [a] | 38 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| Crustacea | Branchiopoda | Cladocera | *Daphnia pulex* [b] | 289 | 20 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 321 |
| | Malacostraca | Amphipoda | *Hyalella azteca* [a] | 60 | 13 | 7 | 0 | 0 | 0 | 0 | 0 | 12 | 92 |
| Hexapoda | Insecta | Hymenoptera | *Apis mellifera* [b] | 38 | 15 | 7 | 0 | 4 | 0 | 0 | 0 | 2 | 66 |
| | | Diptera | *Drosophila melanogaster* [b] | 137 | 29 | 10 | 1 | 5 | 13 | 4 | 0 | 29 | 228 |
| | | Lepidoptera | *Bombyx mori* [b] | 144 | 13 | 6 | 1 | 1 | 2 | 0 | 0 | 4 | 171 |
| | | Coleoptera | *Tribolium castaneum* [b] | 110 | 13 | 7 | 2 | 5 | 1 | 1 | 0 | 3 | 142 |
| | | Hemiptera | *Cimex lectularius* [c] | 121 | 15 | 6 | 0 | 5 | 0 | 0 | 0 | 3 | 149 |
| | | Phthiraptera | *Pediculus humanus* [b] | 41 | 12 | 6 | 1 | 1 | 0 | 0 | 0 | 2 | 63 |
| | | Thysanoptera | *Frankliniella occidentalis* [a] | 64 | 14 | 6 | 2 | 3 | 0 | 0 | 2 | 10 | 101 |
| | | *Blattodea* | *Blattella germanica* [a] | 124 | 13 | 7 | 1 | 5 | 0 | 0 | 0 | 1 | 151 |
| | | Isoptera | *Zootermopsis nevadensis* [a] | 58 | 10 | 9 | 1 | 2 | 0 | 1 | 0 | 0 | 81 |
| | | Ephemeroptera | *Ephemera danica* [a] | 100 | 12 | 5 | 2 | 5 | 0 | 1 | 0 | 2 | 127 |
| | | Odonata | *Ladona fulva* [a] | 160 | 13 | 5 | 11 | 5 | 0 | 0 | 0 | 2 | 196 |

**Table S4.2.2.** Number of genes identified as putative cuticle proteins per family in the genomes of several arthropod groups.
[a] Cuticle protein numbers determined by analyzing gene sets with CutProtFam-Pred.[3]
[b] Cuticle protein numbers determined from Ioannidou et al.[3]
[c] Cuticle protein numbers determined from Benoit et al.[7]

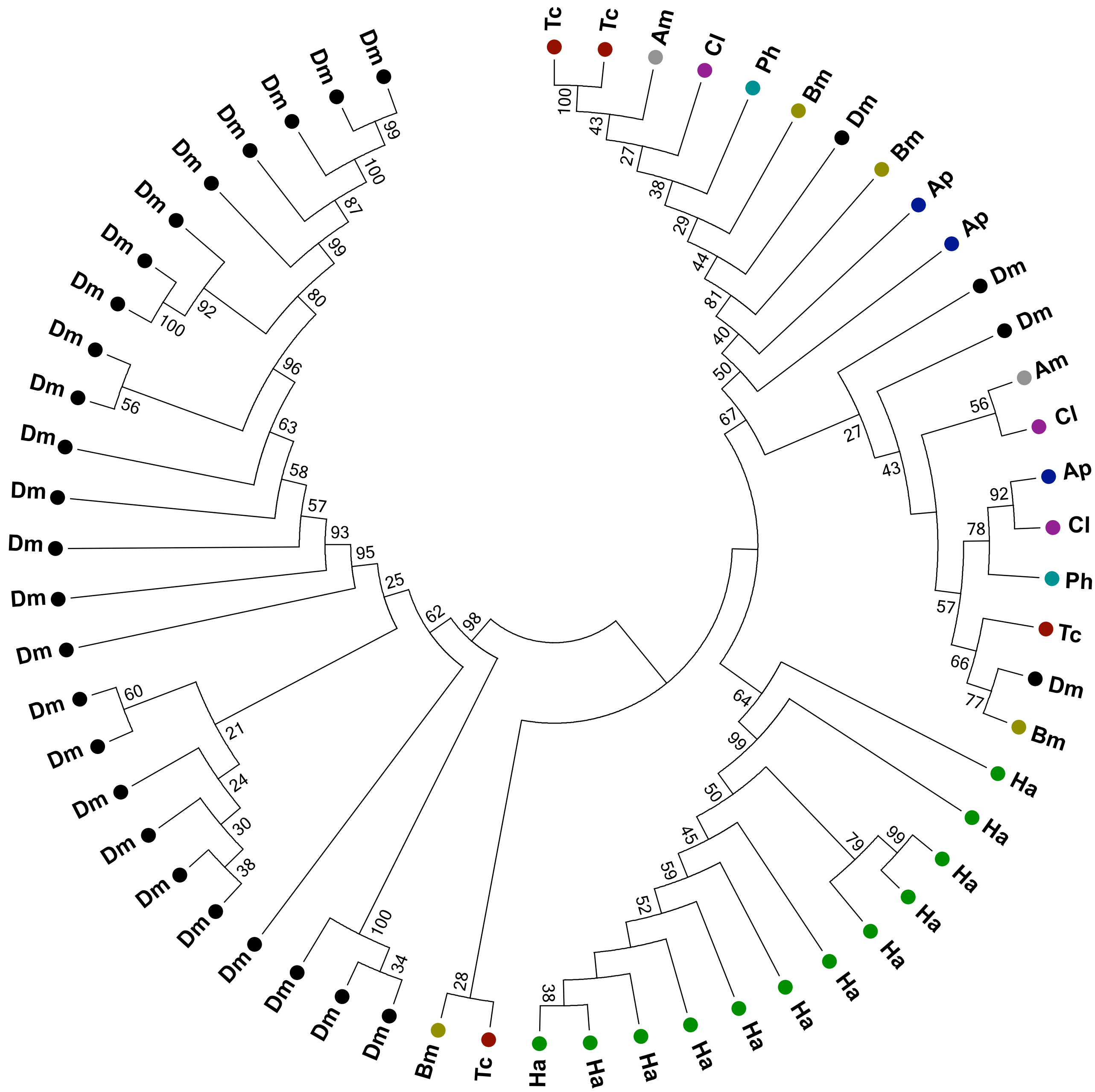


**Figure S4.2.1.** Phylogenetic tree demonstrating relationship of TWDL genes from *Hyalella azteca* (Ha), *Drosophila melanogaster* (Dm), *Tribolium castaneum* (Tc), *Apis mellifera* (Am), *Pediculus humanus corporis* (Ph), *Acyrthosiphon pisum* (Ap), *Bombyx mori* (Bm), *Cimex lectularius* (Cl). *H. azteca* showed a greater number of TWDL genes than other insects, with the notable exception of Dipterans such as *D. melanogaster.* Sequences were aligned with MUSCLE[9] and the tree was constructed using the neighbor-joining method in MEGA6 with Poisson correction and bootstrap replicates (10,000 replicates). Branch values indicate support following 10,000 bootstraps with values below 20% omitted.

**S4.2 References:**

(1) Neville, A. C. *Biology of the Arthropod Cuticle*. Springer: 1975; p 319-374.
(2) Willis, J.; Iconomidou, V.; Smith, R.; Hamodrakas, S. Comprehensive Insect Science. *Cuticular Proteins* **2005,** *4*, 79-109.
(3) Ioannidou, Z. S.; Theodoropoulou, M. C.; Papandreou, N. C.; Willis, J. H.; Hamodrakas, S. J. CutProtFam-Pred: detection and classification of putative structural cuticular proteins from sequence alone, based on profile hidden Markov models. *Insect Biochemistry and Molecular Biology* **2014,** *52*, 51-59.
(4) Willis, J. H. Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era. *Insect Biochemistry and Molecular Biology* **2010,** *40*, (3), 189-204.
(5) Rebers, J. E.; Riddiford, L. M. Structure and expression of a Manduca sexta larval cuticle gene homologous to Drosophila cuticle genes. *Journal of Molecular Biology* **1988,** *203*, (2), 411-423.
(6) McGinnis, S.; Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* **2004,** *32*, (suppl_2), W20-W25.
(7) Benoit, J. B.; Adelman, Z. N.; Reinhardt, K.; Dolan, A.; Poelchau, M.; Jennings, E. C.; Szuter, E. M.; Hagan, R. W.; Gujar, H.; Shukla, J. N. Unique features of a global human ectoparasite identified through sequencing of the bed bug genome. *Nature Communications* **2016,** *7*, 10165.
(8) Chipman, A. D.; Ferrier, D. E.; Brena, C.; Qu, J.; Hughes, D. S.; Schröder, R.; Torres-Oliva, M.; Znassi, N.; Jiang, H.; Almeida, F. C. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede Strigamia maritima. *PLoS Biology* **2014,** *12*, (11), e1002005.
(9) Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **2004,** *32*, (5), 1792-1797.

## S4.3. Cytochrome P450 genes

René Feyereisen
Department of Plant and Environmental Sciences, University of Copenhagen, Copenhagen, Denmark

Alexandra Mechler-Hickson and Carol Eunmi Lee
Center of Rapid Evolution (CORE), University of Wisconsin, Madison WI, 53706 USA

Correspondence to: carollee@wisc.edu

### Introduction

The cytochrome P450 superfamily of genes (P450 genes) is ubiquitous and diverse, as they have been found in all domains of life and are thought to have originated over 3 billion years ago.[1] P450 genes function in metabolizing a wide range of endogenous and exogenous compounds, including toxins, drugs, plant metabolites, and signaling molecules.[2-5] Originally named for an absorbance peak at 450 nm when bound to carbon monoxide, P450s act as monooxygenases, biotransforming various substrates by the reduction of atmospheric oxygen to water.[1,4] As of April 2016, over 35000 different CYP450 genes have been identified and named.[6] In humans, they are well-recognized for their role in drug metabolism, as more than 70% of clinically used drugs involve P*450* pathways.[7] Additionally, P450 proteins are vital for detoxification of many compounds in arthropods, such as caffeine,[8] DDT,[9] and crude oil.[10-12]

Despite functional importance in all taxa, cytochrome P450 genes have not been well characterized across the Arthropoda, especially in crustaceans. In recent years, however, sequencing efforts of crustaceans and other arthropods have opened the possibility of greatly expanding our knowledge of P450s. In particular, the i5k Arthropod Genomes Project has sequenced 35 arthropod genomes (https://www.hgsc.bcm.edu/i5k-pilot-project-summary), including that of the freshwater amphipod *Hyalella azteca*. This amphipod *H. azteca* is frequently used as a toxicological model system, including for response to aqueous pesticides, and as a biomonitor for heavy metals.[13, 14] Thus, the characterization of this vital superfamily of detoxification enzymes in the *H. azteca* genome would be of great use for future research.

### Methods and Materials

#### Gene Discovery, Curation and Nomenclature

P450 genes were queried in the *Hyalella azteca* genome with known *Daphnia pulex* P450 sequences, using blastp against the predicted gene models and tblastn against the *H azteca* scaffolds. The results were merged into a single list, revealing that the gene models represented only about half the CYPome. The new gene models were assembled with the help of RNA-seq reads to define exon/intron junctions. Gaps in the genome assembly prevented us from determining the exact number of P450 genes, although we classified all fragments containing recognizable P450 fragments as judged by the presence of key sequence motifs and by alignments with available arthropod P450 sequences.

The P450 sequences were named by the cytochrome P450 nomenclature committee (Dr. D.R. Nelson, University of Tennessee; see supplemental file S6.1, Table S6.2). Names were determined using evolutionary relationships based on expansive phylogenetic trees consisting of the total known complement of P450s.[6] While there are no strict percent identity requirements, generally P450s in the same family share at least 40% identity, while those in the same subfamily share 55% identity. Clans are a level designated by the nomenclature committee with generally lower percent identity among members, and are determined by clade groupings on the same phylogeny used for naming.[15] The CYPome size discussed below takes into account all complete genes, as well as fragments that contain the typical P450 motif surrounding the cysteine ligand to the heme. Therefore, a more complete genome assembly may reveal additional P450 genes.

**Phylogenetic Reconstruction of P450 sequences**

Selected complete P450 sequences were aligned using PRANK Probabilistic Alignment Kit.[16] Alignment was visually inspected. Phylogenetic reconstruction was performed using MrBayes with a mixed model of amino acid substitution.[17] Tree was visualized using FigTree v1.4.[18]

**Results and Discussion**

In the *H. azteca* genome, we found 70 genes or gene fragments that contained a typical P450 signature (FxxGxxxC), where C is the heme thiolate ligand. However, of these, only 27 were complete genes. Length of complete gene domains was around 500aa, in agreement with the typical P450 length. Two sequences were considered to represent pseudogenes because they contained at least one confirmed frame shift or stop codon.

The most notable difference between the P450 complement of crustaceans relative to hexapods (insects) was the expansion of the CYP2 clan P450s in Crustacea, while the other two clans are predominant in insects. The 70 P450 genes were classifiable into one of four recognized P450 clans, with the CYP2 clan (Fig. S4.3.1) being the largest, with 48 genes. Typical of expanded clades, we found several clusters of genes (and gene fragments) of the CYP2 clan. The term "cluster" here denotes a large number of P450 genes within a short genomic region. The largest such cluster comprised twelve CYP3213A genes on scaffold 51. Three more clusters of 6, 8 and 9 genes of the CYP3213A and CYP3214A subfamilies were found on different scaffolds. The CYP3 and CYP4 clans in *H. azteca* were represented by eight and seven genes, respectively. The fourth P450 clan is the mitochondrial P450 clan, with at least nine genes in *H. azteca*. We did not find any P450 of the CYP20 clan.  These enigmatic P450s are widely distributed in animals, but have been lost several times in the Pancrustacea. They are absent from most insects and from *Daphnia pulex*, but are found in copepods.[19]

The number of P450s found in *H. azteca* was greater to those found in other crustaceans, including the copepod *Tigriopus japonicus* (N = 52),[11] and the copepod *Paracyclopina nana* (N = 46),[12] but similar to that of *D. pulex* (N=75)[20] and somewhat fewer than those in hexapod taxa (including N = 106 in the mosquito *Anopheles gambiae*, N = 81 in the silkworm *Bombyx mori*).[21]

Only five genes that we discovered in the *H. azteca* genome had orthologs in other crustacean genomes. All five were related to the Halloween genes of *Drosophila* and are thought to encode ecdysteroid metabolizing enzymes. Two of these genes, CYP306D1 and CYP18H1 are clustered head-to-tail (N terminus of one gene to the C terminus of the other), whereas they are head-to-head (N terminus to N terminus) in insects[22] and in *Daphnia pulex*.[23] In insects, these genes encode the biosynthetic 25-hydroxylase and the peripheral molting hormone inactivating 26-hydroxylase/oxidase. The other orthologs are CYP302A1 (22-hydroxylase), CYP314A1 (20-hydroxylase) and CYP307A2 (involved in an early step of molting hormone biosynthesis). We did not find a CYP315A1 ortholog in the genome assembly or transcript collection of *H. azteca*, and we cannot be certain as to whether the gene is genuinely missing or simply absent from the current assembly. As this gene encodes the 2-hydroxylase, its bona fide absence would probably imply that the molting hormone of *Hyalella azteca* is 2-deoxy-20-hydroxyedysone. However, 20-hydroxyecdysone was tentatively identified in other amphipods,[24-27] making the absence of CYP315A1 unusual.

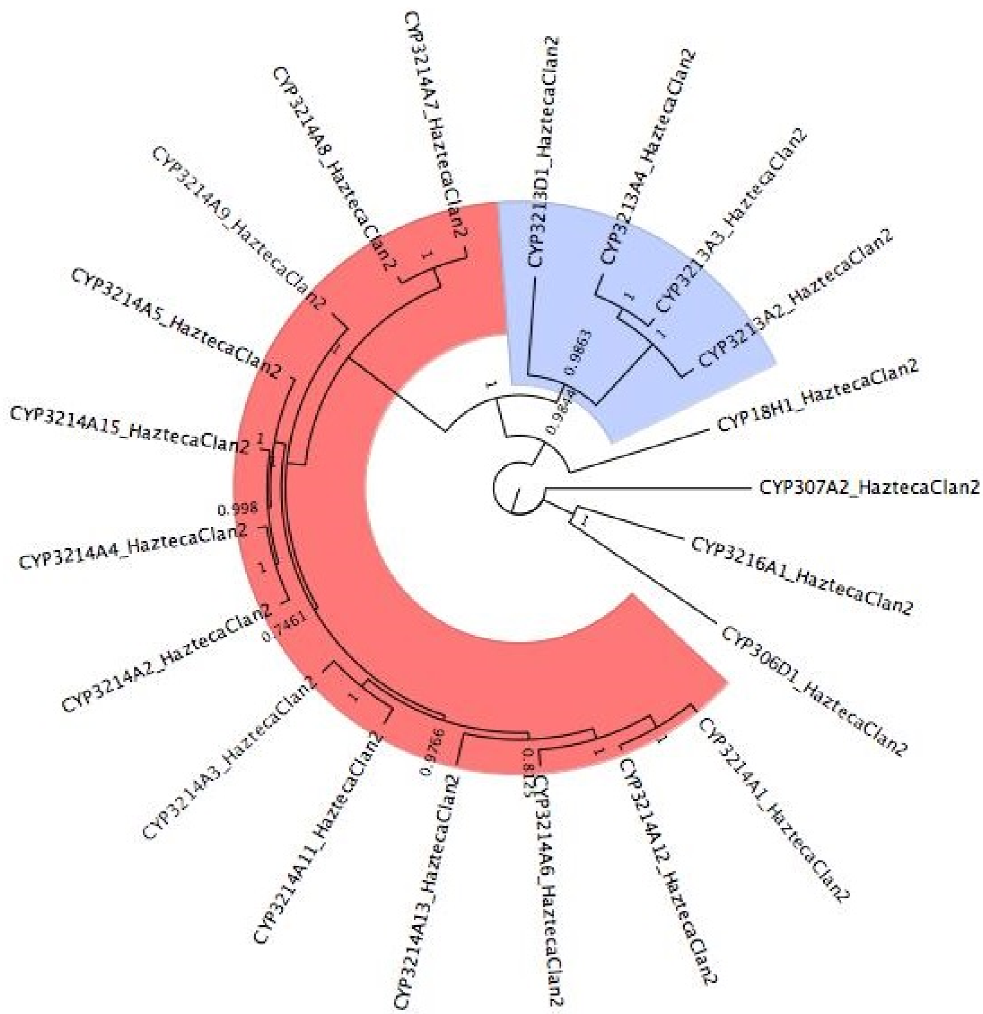



**Figure S4.3.1.** Phylogeny of Clan 2 of *Hyalella azteca* P450s with complete domains. Highlighted clades represent families of Clan 2 P450s that were found in clusters, typical of a gene family expansion. The red highlighted clade is family CYP3214, whereas the blue highlighted clade is family CYP3213. Values at the nodes are posterior probabilities. Alignment generated using PRANK. The phylogeny was generated using MrBayes and visualized using FigTree v1.4.

**S4.3 References:**

(1) Danielson, P. The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. *Current Drug metabolism* **2002,** *3*, (6), 561-597.
(2) Miller, W. L.; Chung, B.-c. The First Defect in Electron Transfer to Mitochondrial P450 Enzymes. In Oxford University Press: 2016.
(3) Morale, M. C.; L'Episcopo, F.; Tirolo, C.; Giaquinta, G.; Caniglia, S.; Testa, N.; Arcieri, P.; Serra, P.-A.; Lupo, G.; Alberghina, M. Loss of aromatase cytochrome P450 function as a risk factor for Parkinson's disease? *Brain Research Reviews* **2008,** *57*, (2), 431-443.
(4) Sigel, A.; Sigel, H.; Sigel, R. K. *The ubiquitous roles of cytochrome P450 proteins*. Wiley Onlie Library: 2007; Vol. 3.
(5) Wahlang, B.; Falkner, K. C.; Cave, M. C.; Prough, R. A. Role of Cytochrome P450 Monooxygenase in Carcinogen and Chemotherapeutic Drug Metabolism. *Advances in Pharmacology* **2015,** *74*, 1-33.
(6) Nelson, D. R. The cytochrome p450 homepage. *Human Genomics* **2009,** *4*, (1), 59.
(7) Zanger, U. M.; Schwab, M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics* **2013,** *138*, (1), 103-141.
(8) Coelho, A.; Fraichard, S.; Le Goff, G.; Faure, P.; Artur, Y.; Ferveur, J.-F.; Heydel, J.-M. Cytochrome P450-dependent metabolism of caffeine in Drosophila melanogaster. *PloS One* **2015,** *10*, (2), e0117328.
(9) Mitchell, S. N.; Stevenson, B. J.; Müller, P.; Wilding, C. S.; Egyir-Yawson, A.; Field, S. G.; Hemingway, J.; Paine, M. J. I.; Ranson, H.; Donnelly, M. J. Identification and validation of a gene causing cross-resistance between insecticide classes in *Anopheles gambiae* from Ghana. *Proceedings of the National Academy of Sciences* **2012,** *109*, (16), 6147-6152.
(10) Matsuo, A. Y.; Woodin, B. R.; Reddy, C. M.; Val, A. L.; Stegeman, J. J. Humic substances and crude oil induce cytochrome P450 1A expression in the Amazonian fish species Colossoma macropomum (Tambaqui). *Environmental Science & Technology* **2006,** *40*, (8), 2851-2858.
(11) Han, J.; Won, E.-J.; Hwang, D.-S.; Shin, K.-H.; Lee, Y. S.; Leung, K. M.-Y.; Lee, S.-J.; Lee, J.-S. Crude oil exposure results in oxidative stress-mediated dysfunctional development and reproduction in the copepod Tigriopus japonicus and modulates expression of cytochrome P450 (CYP) genes. *Aquatic Toxicology* **2014,** *152*, 308-317.
(12) Han, J.; Won, E.-J.; Kim, H.-S.; Nelson, D. R.; Lee, S.-J.; Park, H. G.; Lee, J.-S. Identification of the full 46 cytochrome P450 (CYP) complement and modulation of CYP expression in response to water-accommodated fractions of crude oil in the cyclopoid copepod Paracyclopina nana. *Environmental Science & Technology* **2015,** *49*, (11), 6982-6992.
(13) Weston, D. P.; Holmes, R. W.; You, J.; Lydy, M. J. Aquatic toxicity due to residential use of pyrethroid insecticides. *Environ Sci Technol* **2005,** *39*, (24), 9778-84.
(14) Couillard, Y.; Grapentine, L.; Borgmann, U.; Doyle, P.; Masson, S. The amphipod Hyalella azteca as a biomonitor in field deployment studies for metal mining. *Environmental Pollution* **2008,** *156*, (3), 1314-1324.
(15) Nelson, D. R. Cytochrome P450 Nomenclature, 2004. *Cytochrome P450 Protocols* **2006**, 1-10.
(16) Löytynoja, A.; Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **2008,** *320*, (5883), 1632-1635.
(17) Ronquist, F.; Teslenko, M.; Van Der Mark, P.; Ayres, D. L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M. A.; Huelsenbeck, J. P. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **2012,** *61*, (3), 539-542.
(18) Rambaut, A. FigTree v1. 4. *Molecular evolution, phylogenetics and epidemiology.* Edinburgh, UK: University of Edinburgh, Institute of Evolutionary Biology **2012**.

(19) Han, J.; Kim, D.-H.; Seo, J. S.; Kim, I.-C.; Nelson, D. R.; Puthumana, J.; Lee, J.-S. Assessing the identity and expression level of the cytochrome P450 20A1 (CYP20A1) gene in the BPA-, BDE-47, and WAF-exposed copepods *Tigriopus japonicus* and *Paracyclopina nana*. *Comp Biochem Physiol C Toxicol Pharmacol* **2017,** *193*, 42-49.
(20) Baldwin, W. S.; Marko, P. B.; Nelson, D. R. The cytochrome P450 (CYP) gene superfamily in *Daphnia pulex*. *BMC Genomics* **2009,** *10*, (1), 169.
(21) Feyereisen, R. Evolution of insect P450. In Portland Press Limited: 2006.
(22) Claudianos, C.; Ranson, H.; Johnson, R.; Biswas, S.; Schuler, M.; Berenbaum, M.; Feyereisen, R.; Oakeshott, J. G. A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Molecular Biology* **2006,** *15*, (5), 615-636.
(23) Rewitz, K. F.; Gilbert, L. I. Daphnia Halloween genes that encode cytochrome P450s mediating the synthesis of the arthropod molting hormone: evolutionary implications. *BMC Evolutionary Biology* **2008,** *8*, (1), 60.
(24) Blanchet, M.; Porcheron, P.; Dray, F. Étude des variations du taux des ecdysones au cours du cycle d'intermue chez le male d'Orchestia gammarella Pallas (Crustacé Amphipode) par dosage radio-immunologique. *CR Acad. Sc. Paris* **1976,** *283*, 651-654.
(25) Blanchet, M.; Porcheron, P.; Dray, F. Variations du taux des ecdystéroïdes au cours des cycles de mue et de vitellogenèse chez le Crustacé Amphiphode, Orchestia gammarellus. *International Journal of Invertebrate Reproduction* **1979,** *1*, (2), 133-139.
(26) Graf, F.; Delbecque, J. Ecdysteroid titers during the molt cycle of *Orchestia cavimana* (Crustacea, Amphipoda). *General and Comparative Endocrinology* **1987,** *65*, (1), 23-33.
(27) Block, D. S.; Bejarano, A. C.; Chandler, G. T. Ecdysteroid concentrations through various life-stages of the meiobenthic harpacticoid copepod, *Amphiascus tenuiremis* and the benthic estuarine amphipod, *Leptocheirus plumulosus*. *General and Comparative Endocrinology* **2003,** *132*, (1), 151-160.

## S4.4. Early Developmental genes – Maternal Effect genes, Gap genes, and Pair Rule Genes

Andrew G. Cridge
Laboratory for Evolution and Development, Department of Biochemistry, University of Otago, Dunedin, 9054, New Zealand

Correspondence to: andrew.cridge@otago.ac.nz

### Introduction

One of the main reasons for choosing to sequence the *Hyalella azteca* genome was due to its emerging status as a developmental model system. For this reason, it was of particular interest to analyse its early developmental gene complement and compare it to the extensively studied model organisms *Drosophila melanogaster* and other sequenced crustaceans (e.g. *Daphnia pulex*).

Early developmental genes are responsible for determining anterior/posterior (A/P) and the dorsal/ventral (D/V) axis of the embryo. The establishment of the axes enables the embryo to undergo segmentation. Segmentation, or the subdivision of the developing embryo into serially homologous units, is one of the hallmarks of arthropods development. Arthropod segmentation is best understood in the fly *Drosophila melanogaster*. *D. melanogaster* differs from most arthropods in that all segments are formed from the early blastoderm and segments are formed simultaneously (so-called long-germ developmental mode). In most other arthropods only the anterior segments are formed in a similar way to *D. melanogaster* with the posterior sections added one at a time or in pairs of two from cell material derived from a posterior growth zone (so-called short-germ developmental mode). Crustaceans show a range of short to long germ development. However, segmentation mechanisms are not universally conserved, and only little is known about the genetic patterning of the anterior segments, hence the need to study early development genes in *H. azteca*.

### Methods

The choice of early developmental genes to annotate was informed by GO term annotations in *D. melanogaster* and *D. pulex*. Protein sequences for developmental genes for *D. melanogaster* and *D. pulex* were obtained from http://flybase.org/[1] and http://wfleabase.org/[2] respectively. Contig sequences were searched for homology to the selected protein sequences using tbastn. Gene models (HAZTv0.5.3-models or augustusmasked) which aligned with the regions of highest homology identified by tbastn search were chosen for further analysis. RNAseq mapped reads were compared with the gene models to determine the transcribed regions. The transcribed regions were used to determine the protein sequences of the gene. Theses protein sequences were used in reciprocal blast searches (blastx NCBI) to confirm the homology of the orthologs. Gene models were edited to resolve conflicts between RNAseq, blastx and homology data. Development gene protein sequences that could not be mapped to individual sequence contigs were subsequently mapped to the *H. azteca* RNA-seq redundant assembly using CLC (tblastn), to verify the presence or absence of the gene.

**Results and Discussion**

In total, ~30 genes that are known, in other anthropoda, to be involved in developmental processes were annotated. Including those involved in early pattern (terminal-patterning and head patterning) and the segmentation cascade.

Early patterning genes (e.g. *caudal*, *hunchbach*) appear conserved relative to what is known from other insects (Table S4.4.1). There is, as expected, no *bicoid* orthologue.[3] Other genes know from *D. melanogaster* but not found in other insects, such as *swallow* are also not found in the *H. azteca* genome. A notable absence from the genome was the early patterning gene *nanos*. However, a mRNA sequence homologue to *nanos* was identified in the RNA-seq reads from adult *H. azteca*, and we thus conclude that the absence of *nanos* from the genome results from the incomplete coverage of sequencing, rather than a real absence from the genome. This conclusion is supported by the presence of the downstream targets of *nanos* such as *hunchback*[4] in the genome.

In *D. melanogaster,* one of the early patterning events controls the specification of the anterior and posterior terminal, via a process known as terminal-patterning. This process is controlled by the spatially restricted activation of the receptor tyrosine kinase *torso*,[5] by the presumptive ligand *trunk*.[6] The genome analysis of *H. azteca* did not recover a copy of the receptor *torso* or ligand *trunk* genes. Similarly, these sequences were also absent from the transcriptome. In the crustacean *D. pulex*, genes similar to *trunk* are found, but these are more similarity to Vertebrate and Lophotrochozoan *noggin* proteins.[2] However no representative of the *noggin* class of proteins was present in the *H. azteca* genome. Also, the torso-like protein, which appears to mediate the spatial restriction of terminal-patterning pathway in *D. melanogaster*,[7] that is present in *D. pulex*, was also absent in the *H. azteca* genome.

Gene orthologs for *hunchback*, *orthodenticle*, *collier*, *cap-n-collar* and *crocodile*, and the trunk gap gene *Krüppel* (*Kr*) the critical proteins involved in insect head patterning were all identified in the *H. azteca* genome (Table S4.4.1). Conserved expression of these genes in insects (reviewed by Rosenberg et al.[8]), myriapod[9] and crustacea suggest that the anterior segmentation system may be conserved in at least these three classes of arthropods. This finding implies that the anterior patterning mechanism already existed in the last common ancestor of this sub-phylum.

The pair-rule genes known from the *D. melanogaster* segmentation cascade were all found in the *H. azteca* genome (Table S4.4.1). Two isoforms of *sloppy-paired* gene (Sp-1 and Sp-4) were identified. However, several genes that are duplicated in the *D. melanogaster* genome were found in only one copy in *H. azteca* (e.g. *gooseberry/gooseberry-neuro*).

| Genes | Species | | |
|---|---|---|---|
| | *D. melanogaster* | *D. pulex* | *H. azteca* |
| *caudal* | ✓ | ✓ | ✓ |
| *hunchback* | ✓ | ✓ | ✓ |
| *nanos* | ✓ | ✓ | ✓ |
| *bicod* | ✓ | ✗ | ✗ |
| *swallow* | ✓ | ✗ | ✗ |
| *trunk* | ✓ | ✗ | ✗ |
| *torso* | ✓ | ✗ | ✗ |
| *torso-like* | ✓ | ✓ | ✗ |
| *noggin/noggin-like* | ✗ | ✓ | ✗ |
| *orthodenticle* | ✓ | ✓ | ✓ |
| *buttonhead* | ✓ | ✗ | ✗ |
| *collier* | ✓ | ✓ | ✓ |
| *cap-n-collar* | ✓ | ✓ | ✓ |
| *crocodile* | ✓ | **?** | ✓ |
| *Krüppel* | ✓ | ✓ | ✓ |
| *huckebein* | ✓ | ✓ | ✓ |
| *even-skipped* | ✓ | ✓ | ✓ |
| *odd-paired* | ✓ | ✓ | ✓ |
| *odd-skipped* | ✓ | ✓ | ✓ |
| *paired* | ✓ | ✓ | ✓ |
| *sloppy-paired* | ✓ | **?** | ✓ x2 |
| *runt/runt-like* | ✓ | ✓ | ✓ |
| *lozenge* | ✓ | **?** | ✓ |
| *gooseberry* | ✓ | **?** | ✓ |

**Table S4.4.1** Presence/absence of early patterning genes in the genomes of *Drosophila melanogaster*, *Daphnia pulex* and *Hyalella azteca*

**S4.4 References:**


(1) Attrill, H.; Falls, K.; Goodman, J. L.; Millburn, G. H.; Antonazzo, G.; Rey, A. J.; Marygold, S. J. FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. *Nucleic Acids Research.* **2016,** *44*, (D1), D786-792. .

(2) Colbourne, J. K.; Singan, V. R.; Gilbert, D. G. wFleaBase: the Daphnia genome database. *BMC Bioinformatics.* **2005,** *6*, 45.

(3) Stauber, M.; Jackle, H.; Schmidt-Ott, U. The anterior determinant bicoid of Drosophila is a derived Hox class 3 gene. *Proceedings of the National Academy of Sciences* **1999,** *96*, (7), 3786-3789.

(4) Wharton, R. P.; Struhl, G. RNA regulatory elements mediate control of Drosophila body pattern by the posterior morphogen nanos. *Cell.* **1991,** *67*, (5), 955-967.

(5) Sprenger, F.; Stevens, L. M.; Nusslein-Volhard, C. The Drosophila gene torso encodes a putative receptor tyrosine kinase. *Nature.* **1989,** *338*, (6215), 478-483.

(6) Casali, A.; Casanova, J. The spatial control of Torso RTK activation: a C-terminal fragment of the Trunk protein acts as a signal for Torso receptor in the Drosophila embryo. *Development.* **2001,** *128*, (9), 1709-1715.

(7) Savant-Bhonsale, S.; Montell, D. J. torso-like encodes the localized determinant of Drosophila terminal pattern formation. *Genes and Development.* **1993,** *7*, (12B), 2548-2555.

(8) Rosenberg, M. I.; Lynch, J. A.; Desplan, C. Heads and tails: evolution of antero-posterior patterning in insects. *Biochimica et Biophysica Acta* **2009,** *1789*, (4), 333-342.

(9) Janssen, R.; Budd, G. E.; Damen, W. G. Gene expression suggests conserved mechanisms patterning the heads of insects and myriapods. *Developmental Biology* **2011,** *357*, (1), 64-72.

## S4.5 Hox genes in the *Hyalella azteca* genome

Monica Munoz-Torres
Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Correspondence to: moni@phoenixbioinformatics.org

### Introduction

Hox genes encode transcription factors with a pivotal role in cell-fate determination and embryonic development of the animal body plan, determining the anterior-posterior axis of the bilaterian body.[1] These transcription factors contain domains composed of sixty amino acids that bind DNA in a sequence-specific manner, known as the homeodomain. Mutations in these genes lead to transformations of body segments and organs known as homeotic mutations.

Hox genes are very conserved in sequence and expression across arthropods and other animals. The homeotic capability of the Hox genes is conserved among arthropods and vertebrates, which diverged more than 600 million years ago. Ten groups of orthology, presumably present in the ancestor of all present-day arthropods in the Early Cambrian, make up what we know today as the Hox Complex.[2] These ten genes are expressed in Hox-like patterns in chelicerates and myriapods. In the insects, however, the closest *Hox 3* homologs (*zerknullt* (*zen*), *zerknullt 2* (zen2) and *bicoid* (*bcd*)) and *fushi tarazu*, have novel developmental roles that do not include a Hox-like role in determining segmental identity.[3]

Multiple Hox clusters have been described for several vertebrates including mice, humans, and fish. In contrast, single clusters have been identified in a number of invertebrates including amphioxus, sea urchins, and several insects like mosquitoes, beetles and locusts. In Drosophila, the complement of Hox genes is divided into two clusters, the Antennapedia Complex (ANT-C)[4] and the Bithorax Complex (BX-C),[5] separated by approximately 7.5Mb. This split is thought to be fairly recent in origin. Drosophila has eight genes with traditional Hox-like developmental function. The ANT-C contains genes required for proper development of the gnathal and thoracic segments (*labial* (*lab*), *proboscipedia* (*pb*), *Deformed* (*Dfd*), *Sex combs reduced* (*Scr*), and *Antennapedia* (*Antp*)), while the BX-C genes (*Ultrabithorax* (*Ubx*), *abdominal-A* (*abd-A*) and *Abdominal-B* (*Abd-B*)) are responsible for the development of the abdomen and telson portions of the insect body plan. Additionally, the Drosophila ANT-C contains the genes *zen*, *zen2*, *bcd* and *ftz*, all homologs of *Hox-3*, and thought to function as

Hox genes in less derived arthropods. Also, there are eight cuticle genes, five lysine tRNA genes, and amalgam (*ama*), which encodes a member of the immunoglobulin superfamily.

## Methods

The genes of the Hox cluster in the *Hyalella azteca* genome were manually curated using an instance of the Apollo Genome Annotation Editor[6] hosted at the i5k Workspace@NAL.[7] Candidate regions were located by querying the genome with known orthologs from closely related species and using a BLAST implementation from the i5k Workspace@NAL (https://i5k.nal.usda.gov/webapp/blast/).[8] Data available from RNAseq experiments conducted on tissues from a mix of juveniles and a mix of multi-aged individuals, as well as analysis on the percentage of GC content and masking of gaps in the assembly were used to inform decisions about every manually annotated gene. After all efforts were exhausted to identify the most accurate representation possible of the underlying biology for each gene model, finalized protein products were checked again against public databases at the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/).[9]

Metadata concerning the description of conserved domains was added using the Conserved Domain Database.[10] When available, primary literature and review articles were used to putatively assign molecular functions, participation in biological processes, and cellular localization for each of the annotated genes using PubMed identifiers from NCBI as well as term identifiers from the Gene Ontology Database.[11]

The length of the cluster, intergenic spaces, and gene sizes were estimated taking into account gaps present in the region. A combination of the intergenic space accounted for, as well as gene coding sequences (CDS) and intron sizes, where available, were used to estimate an approximate size of 0.78 Mb for the *H. azteca* Hox cluster. For the purpose of providing a reference, approximate expected sizes of regions not identified in the *H. azteca* genome (e.g. intervening space between *Ubx* and *abd-A* and between *abd-A* and *Abd-B*) were estimated using data from closely related species.

## Results and Discussion

In most of the insect genomes studied to date, besides the fruit fly, Hox genes are organized in a contiguous, single-copy cluster; e.g.: *Anopheles gambiae,*[12] *Tribolium castaneum*,[13] *Apis mellifera*,[14] and *Nasonia vitripennis*.[15] Little is known about the organization of this cluster in other arthropods. Research outside the insects includes the centipede *Strigamia maritima*, which contains all but one of the canonical arthropod Hox genes (it is missing *Hox3*); they are grouped together and all in the same transcriptional orientation.[16] Also, the mite *Tetranychus urticae*, which has a large gap between the genes *proboscipedia* and *Deformed*; has duplications of the genes *fushi tarazu* and *Antennapedia*, losses of *Hox3* and *abdominal-A*, and an inversion of *Abdominal-B*. [17] *abdominal-A* is also missing from the mite *Archegozetes longisetosus*,[18] in one species of sea spider,[19] and three species of barnacle.[20-22]

Exhaustive computational analyses indicate that the *Hyalella azteca* genome, as it is the case with most arthropods, contains a single copy of each Hox gene. However, only nine of the ten genes described in the arthropod Hox cluster were identified. After careful examination, a candidate gene for the gene *fushi tarazu* (*ftz*) was not found in either version of the assembly. Candidate regions identified as the possible ortholog of *ftz* were missing the LXXLL motif (necessary for the interaction of *ftz* with the gene *ftz transcription factor 1 ftz-f1*). We speculate that the absence of *ftz* is likely due to the numerous and extensive gaps found throughout the assembled genome, and not due to its loss in this lineage. Further experimental data will be needed in order to confirm or dismiss this assertion.

Despite aforementioned challenges imposed by the highly fragmented nature of the genome assembly, our findings suggest that, Hox genes in the *H. azteca* genome are grouped in a contiguous, compact cluster, and without interruptions in the direction of transcription (Figure S4.5.1). Fragmentation of the assembled genome also made it difficult to precisely estimate intergenic distances and gene sizes (Table S4.5.1). For instance, gaps spanning large regions overlapping *Antp* account for a particularly messy assembly in the region, causing the order of the exons to be inverted, but without there being experimental data in support of an apparent interruption in the direction of transcription. The error was noted and accounted for in generating the final, manually curated sequence for *Antp*.

Our analyses indicate that the *H. azteca* Hox cluster, at approx. 0.78 Mb in size, is comparable to that of the scorpion *Mesobuthus martensii*,[23] about double the size of the Hox cluster in another crustacean, the water flea (*Daphnia pulex*),[24] larger than in the myriapod *Strigamia maritima*,[16] and smaller than the cluster in most reported insect species [Pace, 2016]. Intergenic regions of the *H. azteca* Hox cluster contain three microRNAs in conserved positions with respect to many other arthropod species analyzed thus far [see Pace[3] for a review]. *miR-iab-4* is located towards the 5' end of the cluster, between *abd-A* and *Abd-B*; *mir-10* was found between *Dfd* and *Scr*, and *miR-993* is located in its conserved position between *Hox3-A* and *Dfd*. Intergenic regions, miRNAs included, are suspected to be of crucial importance to the regulation of expression of at least some of these homeotic genes.[25]

| Gene | Scaffold | Coordinates | Length (Kb) | CDS Length (aa) | Distance to next gene (Kb) |
|---|---|---|---|---|---|
| ***Abd-B*** | 413 | 550357-602146 | 36 | 370 | n/a |
| ***abd-A*** | | | n/a | 146 | n/a |
| exon1 | 8029 | 1-160 | 0.16 | | |
| exon2 | 1112 | 61659-61937 | 0.27 | | |
| ***Ubx*** | | | n/a | 307 | 101 |
| exon1 | 4427 | 12087-12576 | 0.49 | | |
| exon2 | 652 | 686291-686722 | 0.43 | | |
| ***Antp*** | | | n/a | 290 | 81 |
| exon1 | 652 | 451834-452451 | 0.62 | | |
| exon2 | 652 | 475689-475940 | 0.25 | | |
| ***ftz*** | | | not found | | |
| ***Scr*** | 652 | 238333-297044 | 38 | 342 | 67 |
| ***Dfd*** | | | n/a | 362 | 94 |
| exon1 | 652 | 97143-97617 | 0.48 | | |
| exon2 | 38 | 19578-20188 | 0.61 | | |
| ***Hox3-A*** | 38 | 214983-216344 | 1.3 | 453 | 37 |
| ***pb*** | 38 | 273863-368751 | 61 | 726 | 43 |
| ***lab*** | 38 | 453149-530831 | 19 | 371 | - |

**Table S4.5.1.** Location and size of the ten HOX genes found in the *H. azteca* genome assembly HAZT_1.0

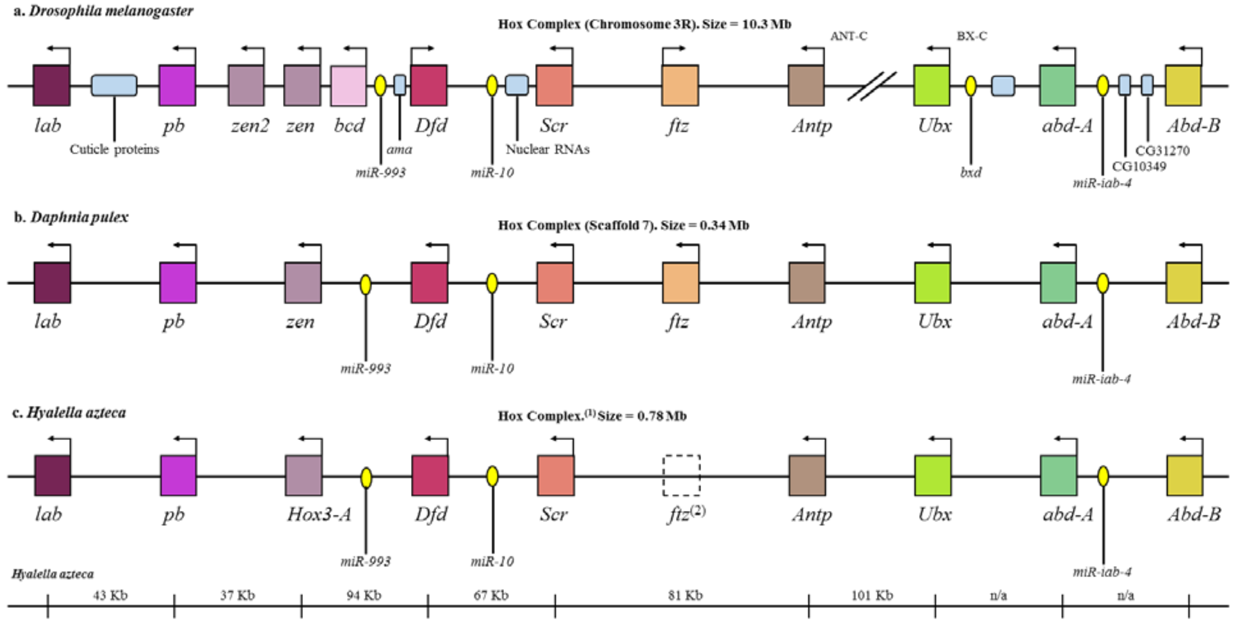



**Figure S4.5.1.** Arrangement of the *H. azteca* HOX gene complex and comparison with HOX gene complexes from *D. pulex* and *D. melanogaster*

## S4.5 References:


(1) Lemons, D.; McGinnis, W. Genomic evolution of Hox gene clusters. *Science* **2006,** *313*, (5795), 1918-1922.
(2) Hughes, C. L.; Kaufman, T. C. Hox genes and the evolution of the arthropod body plan. *Evolution & Development* **2002,** *4*, (6), 459-499.
(3) Pace, R. M.; Grbić, M.; Nagy, L. M. Composition and genomic organization of arthropod Hox clusters. *EvoDevo* **2016,** *7*, (1), 11.
(4) Kaufman, T. C.; Lewis, R.; Wakimoto, B. Cytogenetic analysis of chromosome 3 in Drosophila melanogaster: the homoeotic gene complex in polytene chromosome interval 84a-B. *Genetics* **1980,** *94*, (1), 115-133.
(5) Duncan, I. The bithorax complex. *Annual Review of Genetics* **1987,** *21*, (1), 285-319.
(6) Lee, E.; Helt, G. A.; Reese, J. T.; Munoz-Torres, M. C.; Childers, C. P.; Buels, R. M.; Stein, L.; Holmes, I. H.; Elsik, C. G.; Lewis, S. E. Web Apollo: a web-based genomic annotation editing platform. *Genome Biology* **2013,** *14*, (8), R93.
(7) Poelchau, M.; Childers, C.; Moore, G.; Tsavatapalli, V.; Evans, J.; Lee, C.-Y.; Lin, H.; Lin, J.-W.; Hackett, K. The i5k Workspace@ NAL—enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Research* **2014,** *43*, (D1), D714-D719.
(8) BLAST at the i5k Workspace@NAL. https://i5k.nal.usda.gov/webapp/blast/
(9) National Center for Biotechnology Information (NCBI), U.S. National Library of Medicine. https://www.ncbi.nlm.nih.gov/
(10) Marchler-Bauer, A.; Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Research* **2004,** *32*, (suppl_2), W327-W331.
(11) Gene Ontonly Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research* **2015,** *43*, (D1), D1049-D1056.
(12) Holt, R. A.; Subramanian, G. M.; Halpern, A.; Sutton, G. G.; Charlab, R.; Nusskern, D. R.; Wincker, P.; Clark, A. G.; Ribeiro, J. C.; Wides, R. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **2002,** *298*, (5591), 129-149.
(13) Tribolium Genome Sequencing Consortium. The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **2008**. *452*, 949–955.
(14) The Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* **2006,** *443*, (7114), 931.
(15) Werren, J. H.; Richards, S.; Desjardins, C. A.; Niehuis, O.; Gadau, J.; Colbourne, J. K.; Group, N. G. W. Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science* **2010,** *327*, (5963), 343-348.
(16) Chipman, A. D.; Ferrier, D. E.; Brena, C.; Qu, J.; Hughes, D. S.; Schröder, R.; Torres-Oliva, M.; Znassi, N.; Jiang, H.; Almeida, F. C. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. *PLoS Biology* **2014,** *12*, (11), e1002005.
(17) Grbić, M.; Van Leeuwen, T.; Clark, R. M.; Rombauts, S.; Rouzé, P.; Grbić, V.; Osborne, E. J.; Dermauw, W.; Ngoc, P. C. T.; Ortego, F. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* **2011,** *479*, (7374), 487.
(18) Cook, C. E.; Smith, M. L.; Telford, M. J.; Bastianello, A.; Akam, M. Hox genes and the phylogeny of the arthropods. *Current Biology* **2001,** *11*, (10), 759-763.
(19) Manuel, M.; Jager, M.; Murienne, J.; Clabaut, C.; Le Guyader, H. Hox genes in sea spiders (Pycnogonida) and the homology of arthropod head segments. *Development Genes and Evolution* **2006,** *216*, (7-8), 481-491.
(20) Gibert, J. M.; Mouchel-Vielh, E.; Quéinnec, E.; Deutsch, J. S. Barnacle duplicate engrailed genes: divergent expression patterns and evidence for a vestigial abdomen. *Evolution & Development* **2000,** *2*, (4), 194-202.

(21) Blin, M.; Rabet, N.; Deutsch, J. S.; Mouchel-Vielh, E. Possible implication of Hox genes Abdominal-B and abdominal-A in the specification of genital and abdominal segments in cirripedes. *Development Genes and Evolution* **2003,** *213*, (2), 90-96.
(22) Mouchel-Vielh, E.; Rigolot, C.; Gibert, J.-M.; Deutsch, J. S. Molecules and the Body Plan: TheHoxGenes of Cirripedes (Crustacea). *Molecular Phylogenetics and Evolution* **1998,** *9*, (3), 382-389.
(23) Di, Z.; Yu, Y.; Wu, Y.; Hao, P.; He, Y.; Zhao, H.; Li, Y.; Zhao, G.; Li, X.; Li, W. Genome-wide analysis of homeobox genes from Mesobuthus martensii reveals Hox gene duplication in scorpions. *Insect Biochemistry and Molecular Biology* **2015,** *61*, 25-33.
(24) Colbourne, J. K.; Pfrender, M. E.; Gilbert, D.; Thomas, W. K.; Tucker, A.; Oakley, T. H.; Tokishita, S.; Aerts, A.; Arnold, G. J.; Basu, M. K.; Bauer, D. J.; Caceres, C. E.; Carmel, L.; Casola, C.; Choi, J. H.; Detter, J. C.; Dong, Q.; Dusheyko, S.; Eads, B. D.; Frohlich, T.; Geiler-Samerotte, K. A.; Gerlach, D.; Hatcher, P.; Jogdeo, S.; Krijgsveld, J.; Kriventseva, E. V.; Kultz, D.; Laforsch, C.; Lindquist, E.; Lopez, J.; Manak, J. R.; Muller, J.; Pangilinan, J.; Patwardhan, R. P.; Pitluck, S.; Pritham, E. J.; Rechtsteiner, A.; Rho, M.; Rogozin, I. B.; Sakarya, O.; Salamov, A.; Schaack, S.; Shapiro, H.; Shiga, Y.; Skalitzky, C.; Smith, Z.; Souvorov, A.; Sung, W.; Tang, Z.; Tsuchiya, D.; Tu, H.; Vos, H.; Wang, M.; Wolf, Y. I.; Yamagata, H.; Yamada, T.; Ye, Y.; Shaw, J. R.; Andrews, J.; Crease, T. J.; Tang, H.; Lucas, S. M.; Robertson, H. M.; Bork, P.; Koonin, E. V.; Zdobnov, E. M.; Grigoriev, I. V.; Lynch, M.; Boore, J. L. The ecoresponsive genome of *Daphnia pulex*. *Science* **2011,** *331*, (6017), 555-561.
(25) Yekta, S.; Tabin, C. J.; Bartel, D. P. MicroRNAs in the Hox network: an apparent link to posterior prevalence. *Nature Reviews. Genetics* **2008,** *9*, (10), 789.

## S4.6. Glutathione peroxidases

Peter Bain
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Urrbrae, Australia.

Correspondence to: peter.bain.0@gmail.com

### Introduction

Members of the glutathione peroxidase (GPx) family of antioxidant enzymes catalyse the reduction of hydrogen peroxide or organic hydroperoxides using glutathione as the electron donor, thus contributing to the maintenance of cellular redox balance and protecting against the effects of oxidative stress. In mammals, 8 GPx subfamilies have been described that differ in substrate specificity or biological function.[1-3] Mammalian GPx1-4 and GPx6 are selenoproteins containing selenocysteine residues at the active site; selenocysteine (Sec) insertion sequence (SECIS) elements present in the 3'-untranslated region (3'-UTR) of the mRNA direct the read-through of an internal opal stop codon (UGA), resulting in the recruitment of Sec to the nascent polypeptide chain.[4] GPx genes are present in organisms from all kingdoms of life, and although glutathione-dependent selenoenzymes are the canonical forms, the nomenclature has been expanded to include Cys-containing variants that use thioredoxin as the reducing substrate (e.g. *Drosophila melanogaster Dm*GPx[5]). Although the specific functions have not been well studied, arthropods are also known to express Sec-containing GPx proteins.[1] While these diverge somewhat from the mammalian subfamilies, a recent phylogenetic analysis suggests that arthropod GPx homologues characterised to date form a clade with mammalian GPx3.[6]

### Methods

Candidate GPx genes were identified in the *Hyalella azteca* draft genome assembly by conducting similarity searches using selected malacostraca GPx sequences as queries. Initially, four *Daphnia pulex* glutathione peroxidases from the UniProtKB database (E9HDF9_DAPPU, E9I0W4_DAPPU, E9FVL7_DAPPU, and E9FVL8_DAPPU) were used tblastn queries to search for similar sequences in the *H. azteca* genome via the i5k BLAST portal (https://i5k.nal.usda.gov/webapp/blast). The default settings were used for all searches. Additional searches were carried out in nucleotide space (blastn) using *Metapenaeus ensis* ovary GPx cDNA (GenBank EU399681) and *Macrobrachium nipponense* GPx1 cDNA (GenBank HQ651155) as queries.

For putative Sec-containing GPx genes, the 3’-UTR was analysed to identify candidate SECIS elements using SECISearch3, as implemented in the Seblastian web tool (http://seblastian.crg.es/).[7]

**Results and Discussion**

In total, four putative GPx genes were identified in the draft assembly (Table S4.6.1). One of these, a GPx3-like gene on scaffold 135, contains a putative Sec codon in exon 1 and a SECIS element in the 3’-UTR (Figure S4.6.1). Note that an additional (first) exon, which contains the putative Sec codon, was manually added to the gene model. The deduced protein matches well with GPx3 sequences from other crustaceans (as well as some sequences annotated as GPx6), including the position of the conserved Sec residue (Figure S4.6.2). There is strong read coverage support for the exons in the revised model.

At this stage, there is little evidence that the putative GPx7 homologue (scaffold 54; encoded by the HaztTmpA003459-RA model in the draft genome) is a functional gene. The 1$^{st}$ and 2$^{nd}$ exons in the model do not appear code for GPx, and there is a gap upstream from the 3$^{rd}$ exon. Comparison of the deduced protein to other GPx7 sequences in public databases, including one from *Daphnia magna* (GenBank acc. KZS21126.1), shows that the NH2-terminal end does not match with known sequences. Extending the coding sequence by marking the TGA as a read-through (Sec codon) results in a longer coding sequence that matches well with database sequences, but two factors suggest that this is an unlikely scenario: 1. there is no evidence of a SECIS element in the downstream sequence, and 2. none of the matching database proteins contain Sec at this position – instead there is a highly conserved alanine. Therefore, it must be concluded that the first exon/s, and probably the promoter region, are missing from the assembly.

| Putative subfamily | Scaffold | Location | Completeness | Best match for deduced protein |
|---|---|---|---|---|
| GPx3 (or 6) | 135 | 1104852 - 1107448 | Complete (after manual edits) | Glutathione peroxidase 3 [*Penaeus monodon*] gb\|ALM09356.1 |
| GPx7 | 54 | 765335-765855 | Partial (missing first exon/s) | Glutathione peroxidase 7 [*Daphnia magna*] gb\|KZS21126.1 |
| Ambiguous (poss. PH-GPx) | 458 | 399767-402009 | Partial (missing exons upstream and downstream, many gaps in the region) | Phospholipid-hydroperoxide glutathione peroxidase [*Scylla paramamosain*] gb\|AIW42687.1 |
| Bacterial GPx | 3 | 5985433-5985918 | Complete (but no read support; entire scaffold looks bacterially-derived) | Glutathione peroxidase [*Curvibacter gracilis*] ref\|WP_027474993.1 |

**Table S4.6.1.** Summary of locations and quality of the gene models encoding putative GPx-family enzymes.

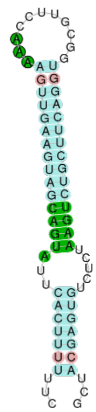


**Figure S4.6.1**. Secondary structure of the putative SECIS element in the 3'-UTR of *H. azteca* GPx3.

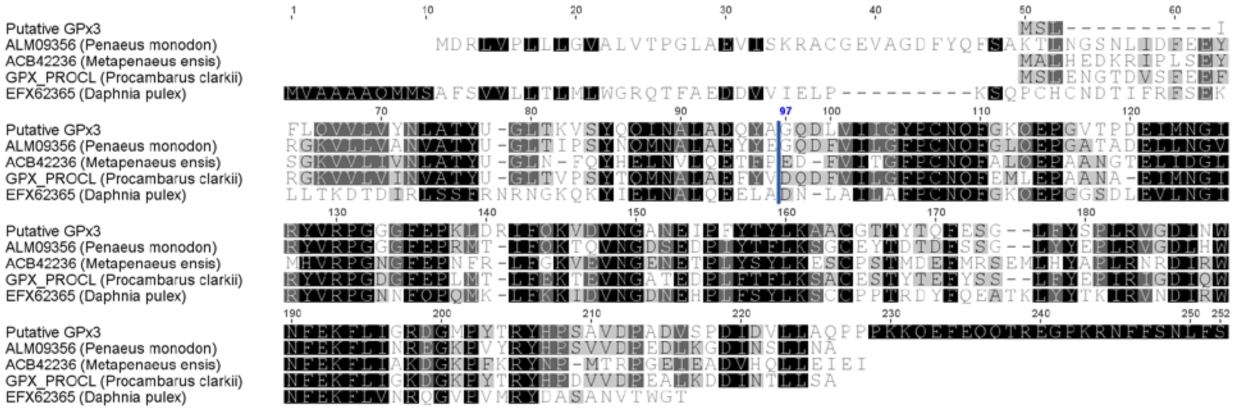



**Figure S4.6.2.** Alignment of *H. Azteca* putative GPx3 with GPx proteins from selected crustaceans.

**S4.6. References:**

(1) Margis, R.; Dunand, C.; Teixeira, F. K.; Margis-Pinheiro, M. Glutathione peroxidase family – an evolutionary overview. *FEBS Journal* **2008,** *275*, (15), 3959-3970.
(2) Flohé, L. Glutathione Peroxidases. In *Selenoproteins and Mimics*, Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; 10.1007/978-3-642-22236-8_1pp 1-25.
(3) Arthur, J. R. The glutathione peroxidases. *Cellular and Molecular Life Sciences : CMLS* **2000,** *57*, (13-14), 1825-35.
(4) Allmang, C.; Wurth, L.; Krol, A. The selenium to selenoprotein pathway in eukaryotes: More molecular partners than anticipated. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2009,** *1790*, (11), 1415-1423.
(5) Maiorino, M.; Ursini, F.; Bosello, V.; Toppo, S.; Tosatto, S. C. E.; Mauri, P.; Becker, K.; Roveri, A.; Bulato, C.; Benazzi, L.; De Palma, A.; Flohé, L. The Thioredoxin Specificity of Drosophila GPx: A Paradigm for a Peroxiredoxin-like Mechanism of many Glutathione Peroxidases. *Journal of Molecular Biology* **2007,** *365*, (4), 1033-1046.
(6) Dias, F. A.; Gandara, A. C. P.; Perdomo, H. D.; Gonçalves, R. S.; Oliveira, C. R.; Oliveira, R. L. L.; Citelli, M.; Polycarpo, C. R.; Santesmasses, D.; Mariotti, M.; Guigó, R.; Braz, G. R.; Missirlis, F.; Oliveira, P. L. Identification of a selenium-dependent glutathione peroxidase in the blood-sucking insect Rhodnius prolixus. *Insect Biochemistry and Molecular Biology* **2016,** *69*, 105-114.
(7) Mariotti, M.; Lobanov, A. V.; Guigo, R.; Gladyshev, V. N. SECISearch3 and Seblastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Research* **2013,** *41*, (15), e149.

## S4.7. Glutathione S-Transferases

Austin Manny
Department of Microbiology & Cell Science, University of Florida, Gainesville, FL, USA

Correspondence to: austin.manny@gmail.com

### Introduction

Glutathione S-transferases (GSTs) are a ubiquitous family of enzymes found in almost all aerobic organisms.[1] These GSTs function to detoxify compounds that pose a threat to the organism, from endogenous molecules like superoxides to xenobiotics. GSTs catalyze the transfer of the sulfhydryl group (-SH) from glutathione to an electrophilic species, making the target compound more stable and thus deactivating it. For this reason, GSTs are important in the study of pesticide resistance in arthropods and non-target species like *Hyalella azteca*.

Glutathione S-transferases can be broken down into three main groups: cytosolic, mitochondrial, and microsomal. The primary group found in insects and amphipods is the cytosolic GSTs; several microsomal GSTs (MGSTs) have also been found. Mitochondrial GSTs have thus far been virtually absent from most arthropod and amphipod genomes.[2]

The cytosolic GSTs can be further broken down into several classes: Delta, Epsilon, Sigma, Theta, Omega, and Zeta. Delta and Epsilon classes are found almost exclusively in insects. Meanwhile, the Sigma class is found in a diverse range of eukaryote.[3] The remainder of insect GSTs not found in the Delta and Epsilon classes belong to Zeta, Theta, and Omega classes.[4]

The structure of GSTs is conserved to a fairly high degree. The cytosolic GSTs tend to be 200-250 amino acids in length and have 2-4 exons. Microsomal GSTs are usually shorter, at approximately 150 amino acids with 1-2 introns.[2]

Because of their potent detoxifying capacity, Glutathione S-Transferases play an important role in cycling reactive compounds through the environment. GSTs have also been implicated in the emergence of resistance to multiple classes of pesticide including DDT, pyrethroid, and organophosphate pesticides. This resistance can be due to increased expression of GSTs across all classes, or in some cases it is the result of favorable mutations in specific GST classes. This resistance to organophosphate has been directly linked to Glutathione S-Transferases in many organisms.[1] These findings, combined with the historic use of *H. azteca* for toxicological assessment, make a strong case for using comparative genomics to further understand the scope of GSTs in *Hyalella azteca.*

**Methods**

Shi *et al.*[2] characterized all Glutathione S-transferases in the red flour beetle, *Tribolium castaneum*. That analysis found 36 cytosolic GSTs and 5 microsomal GSTs. All 41 GST sequences from *T. castaneum* were searched in the *H. azteca* genome. Successful hits were subsequently annotated with a name corresponding to the closest gene in *T. castaneum.*

The amino acid sequence from each *T. castaneum* GST was queried in the i5k@NAL workspace using the tBLASTn algorithm.[5] A list of likely matches in *H. azteca* was then returned and analyzed with respect to percent identity, length, and e-value. The hit with the best holistic combination of these criteria was selected for annotation with the analogous name from *T. castaneum*. If multiple genes had the same tBLASTn hit, the scores were assessed and the found gene was reevaluated for annotation. Annotation was performed with JBrowse. Additional evidence for gene annotation was provided by JBrowse Hazt_v4.5.3 and augustus_masked gene models as well as transcriptomics data, previously loaded into JBrowse.

**Results and Discussion**

In total, 41 Glutathione S-transferases were queried in the *H. azteca* genome. Of the 34 cytosolic Glutathione S-transferases queried, 19 definitive cytosolic GSTs were found, primarily from the Delta and Epsilon classes, though there was a large number from the Sigma class. Of the five microsomal GSTs queried, only one returned a match strong enough to yield an annotation.

*<u>Delta and Epsilon Glutathione S-Transferases</u>*

The *T. castaneum* contains 3 known Delta GSTs and 19 known Epsilon GSTs. These amino acid sequences were queried in the *Hyalella azteca* genome via the i5k workspace. All three Delta GSTs were found in the *H. azteca* genome with strong sequence fidelity. Six of the Epsilon GSTs were identified in the *H. azteca* genome. Thus, nearly half of all the discovered GSTs in *H. azteca* belong to these two related classes. This finding is consistent with studies of GST genes in other organisms. This is of extreme importance because these two classes are found almost exclusively in arthropods. The wide abundance of ortholog GSTs in *H. azteca* underscores the utility of studying this organism in modeling intricate environmental processes such as the complex outcomes of insecticide utilization.

*Sigma Glutathione S-Transferases*

Seven Sigma GSTs were searched for in the *H. azteca* genome, with six successful hits in *H. azteca.* This makes Sigma the second largest individual class of Glutathione S-Transferases in *H. azteca* behind Epsilon. It also proves Sigma to be the largest sect of non-insect GSTs in *H. azteca*.

*Omega Glutathione S-Transferases*

*T. castaneum* possesses 3 Omega GSTs, all of which were queried and successfully matched with orthologs in the *H. azteca* genome.

*Theta Glutathione S-Transferases*

One Theta GST was searched for and ultimately found in the *H. azteca* genome.

*Microsomal Glutathione S-Transferases*

Five microsomal GSTs (MGSTs) were looked for in the *H. azteca* genome. Interestingly, only one MGST was found in the *H. azteca* genome. This rate of success was markedly lower than it was for cytosolic GSTs. However, it is worth noting that in many arthropods and crustaceans alike, there are indeed many more cytosolic GSTs than MGSTs.[6]

| Gene Name | Scaffold | Starting Position | Ending Position | Strand |
|---|---|---|---|---|
| Glutathione S-Transferase Delta 1 | Scaffold54 | 139408 | 142468 | + |
| Glutathione S-Transferase Delta 2 | Scaffold82 | 1903123 | 1909464 | - |
| Glutathione S-Transferase Delta 3 | Scaffold1079 | 211454 | 215908 | + |
| Glutathione S-Transferase Epsilon 1 | Scaffold1079 | 224175 | 225272 | + |
| Glutathione S-Transferase Epsilon 3 | Scaffold54 | 1033797 | 1050169 | + |
| Glutathione S-Transferase Epsilon 4 | Scaffold82 | 1945935 | 1948264 | + |
| Glutathione S-Transferase Epsilon 5 | Scaffold18 | 459432 | 459728 | + |
| Glutathione S-Transferase Epsilon 7 | Scaffold18 | 438884 | 441255 | + |
| Glutathione S-Transferase Epsilon 12 | Scaffold46 | 1290313 | 1300768 | - |
| Glutathione S-Transferase Sigma 1 | Scaffold481 | 823591 | 828714 | + |
| Glutathione S-Transferase Sigma 2 | Scaffold1410 | 22041 | 24515 | + |
| Glutathione S-Transferase Sigma 3 | Scaffold899 | 610149 | 616882 | - |
| Glutathione S-Transferase Sigma 4 | Scaffold1756 | 10105 | 12540 | - |
| Glutathione S-Transferase Sigma 5 | Scaffold481 | 790530 | 791865 | + |
| Glutathione S-Transferase Sigma 7 | Scaffold85 | 1934869 | 1936746 | + |
| Glutathione S-Transferase Omega 1 | Scaffold276 | 1373722 | 1377287 | + |
| Glutathione S-Transferase Omega 2 | Scaffold276 | 1384338 | 1386686 | + |
| Glutathione S-Transferase Omega 3 | Scaffold120 | 344355 | 346761 | - |
| Glutathione S-Transferase Theta 1 | Scaffold51 | 2705356 | 2708456 | - |
| Microsomal Glutathione S-Transferase 1 | Scaffold235 | 910366 | 914077 | + |

**Table S4.7.1:** Glutathione S-Transferases in *H. azteca.* List of GST genes found in *H. azteca*, with name and genome location coordinates. These 20 genes provide essential detoxifying capability to *H. azteca*, and can serve as a strong basis for better understanding multifaceted environmental phenomena*.*

**S4.7 References:**

(1) Enayati, A. A.; Ranson, H.; Hemingway, J. Insect glutathione transferases and insecticide resistance. *Insect Molecular Biology* **2005,** *14*, (1), 3-8.
(2) Shi, H.; Pei, L.; Gu, S.; Zhu, S.; Wang, Y.; Zhang, Y.; Li, B. Glutathione S-transferase (GST) genes in the red flour beetle, *Tribolium castaneum*, and comparative analysis with five additional insects. *Genomics* **2012,** *100*, (5), 327-335.
(3) Agianian, B.; Tucker, P. A.; Schouten, A.; Leonard, K.; Bullard, B.; Gros, P. Structure of a Drosophila sigma class glutathione S-transferase reveals a novel active site topography suited for lipid peroxidation products. *Journal of Molecular Biology* **2003,** *326*, (1), 151-165.
(4) Ranson, H.; Claudianos, C.; Ortelli, F.; Abgrall, C.; Hemingway, J.; Sharakhova, M. V.; Unger, M. F.; Collins, F. H.; Feyereisen, R. Evolution of supergene families associated with insecticide resistance. *Science* **2002,** *298*, (5591), 179-181.
(5) BLAST at the i5k Workspace@NAL. https://i5k.nal.usda.gov/webapp/blast/
(6) Roncalli, V.; Cieslak, M. C.; Passamaneck, Y.; Christie, A. E.; Lenz, P. H. Glutathione S-transferase (GST) gene diversity in the crustacean *Calanus finmarchicus*–contributors to cellular detoxification. *PloS One* **2015,** *10*, (5), e0123322.

## S4.8. Heat Shock Proteins

Helen C. Poynton
School for the Environment, University of Massachusetts Boston, Boston MA 02067

Correspondence to: helen.poynton@umb.edu

### Introduction

The heat shock protein (HSP) molecular chaperones are highly conserved proteins that facilitate in the refolding of denatured proteins following stress, including thermal stress, but also metals and other toxicants, oxidative stress, and dehydration.[1] Although their discovery was related to their induction by thermal stress,[2] these proteins are actually regulated by the build-up of unfolded proteins allowing them to respond to a diversity of cellular stresses.[3]

HSPs have been divided into several families based on their molecular weight. Of the different families, HSP70, HSP90 and HSP60 play a major role in protein refolding while HSP40 DNAJ protein is a co-factor to HSP70 and delivers nonnative proteins to HSP70.[3] In vertebrate systems, HSP90 also plays a role in the regulation of nuclear receptors, binding to and sequestering them in the cytosol.[4] HSPs recognize unfolded proteins by the presence of exposed hydrophobic residues on the outside of the protein. This ability to recognize and bind exposed hydrophobic residues provides them with the flexibility to interact with and refold a diversity of proteins.[3]

The conservation of the heat shock response across prokaryotic and eukaryotic organisms has made them an appealing study system across diverse fields in biology.[5] HSPs have been the subject of phylogenetic and in evolutionary studies of stress resistance.[5, 6] For example, differences in constitutive expression and induciblity of HSPs across closely related species has provided insight into survival in extreme environments (e.g. desert ants[7]), speciation and niche selection in amphipods,[8] and supports a model of diversifying selection in Antarctic krill.[9] Their conservation and their high induction in response to environmental contaminants initially created much excitement as they were attractive candidates for biomarkers in ecotoxicology.[10] However, their lack of specificity and the difficulty in linking their response to adverse effects has limited their widespread application to environmental monitoring.[11] Despite these limitations, their induction signals the onset of the stress response and warrants further investigations in important ecological species. Here we identify and characterize the sequences of the major HSPs in *H. azteca*, and provide a foundation for their study in this important evolutionary and ecotoxicology model.

## Methods

To identify potential heat shock proteins, differentially expressed contigs (described in supplemental methods) with similarity to *hsp*70 (10 contigs), *hsp*90 (10 contigs), *hsp*60 (2 contigs), and *hsp*40 (3 contigs) were aligned to the genome using blastn,[12] Areas of the genome with alignments were investigated for gene models and RNAseq evidence and exon boundaries were modified if necessary to match RNAseq evidence. In addition, Hsps from *Euphausia superba*[9] and *Homarus americanus*[13] were also tblasted against the *H. azteca* genome to find potential sequences that may not be induced by Cd, Zn, or pesticide treatment.

Coding sequences from predicted gene models were searched for similarity to HSPs in other organisms using blastp against the NCBI protein database. The best blast hit for each predicted model was then blasted back against the genome using tblast to ensure the gene models were complete. In cases were the gene models were not complete or gaps in the scaffold assembly truncated the gene model, these sequences investigated further in the Hazt_assembly 2.0.

To classify HSP70 sequences into the Arthropod gene families recently described by Baringou et al.,[13] we aligned the sequences to the *Procambarus clarkii* HSP70 A2 (KU613184) using clustal omega and identified particular motifs within the proteins that characterized them as Arthropod Group A, Group B, or Group C (shown in Table 2).

## Results and Discussion

*Identification of HSPs in Hyalella azteca*

Several homologous *hsp* sequences were identified in the *H. azteca* genome assembly 1.0 and were annotated on the Apollo web browser (Table S4.8.1). We identified the highest number of gene copies for *hsp70*, with five genes forming a gene cluster on scaffold 277, corresponding to scaffold 288 in assembly 2.0 (see Figure S4.8.1). A sixth gene also found within this cluster had a frame shift resulting in a premature stop codon and was therefore classified as a pseudogene. Three additional *hsp70* genes were found on other scaffolds. All eight *hsp70* are highly similar (75-94% similar at the nucleotide level) and characteristic of other arthropod *hsp70* genes, they contained no introns.[13]

Blast searches to assembly 1.0 revealed three other potential *hsp70* clusters on scaffold 108 (2 gene models) scaffolds 401 (8 genes models) and scaffold 503 (4 gene models). In particular, the gene cluster on scaffold 401 contained six nearly complete models with high similarity to HSP70 and containing the HSP70 signature patterns (I-V) described by Karin and Brocchieri.[14] However, further inspection of these sequences revealed that they lacked the highly conserved critical linker region as well as motifs common to other arthropod HSP70s (see XP_0180017811 in Table S4.8.2). These sequences are therefore classified as “HSP70-like.” *Hsp70* sequences found on scaffolds 108 and 503 were incomplete and fragmented. Although identifying these sequences in assembly 2.0 did help to remove gaps within the gene models, they did not produce complete coding sequences and are likely pseudogenes.

The overall number of *hsp*s described here likely represents the full set of *hsp* genes in *H. azteca.* The number of *hsp*70 (8 genes), *hsp*90 (3 genes), *hsp*40 (3 genes), and *hsp*60 (1 gene) is well within the expected number of these genes found in throughout Arthropoda.[7] Six to twelve *hsp*70 genes have been identified in insects and the five gene cluster found on scaffold 277 is similar to gene clusters identified in *D. melanogaster* [15] and *Aedes aegypti*.[16] Seven out of the eight HSP70 proteins have the C-terminal cytosolic tag “EEVD,” while HSP70A5b2 contains a stop codon that reduces the C-terminus by 21 amino acids and truncates the peptide which would otherwise contain “GGMP” and “EEVD” motifs. Functional analysis is needed to determine the effect of this deletion on protein translocation and function, but RNAseq data and expression analysis from our Cd study suggest that the gene is expressed and induced under stress.

*Classification of Hsp70 genes*

In agreement with Baringou et al.,[13] the HSP70s described here cannot be easily divided into inducible and cognate forms based on sequence characteristics. Of the eight full length *hsp*70 genes found in the *H. azteca* genome, seven of these were induced 6-13 fold following Cd exposure (Table S4.8.1) with *hsp*70A4 as the sole exception. We instead, decided to compare our eight sequences to sequence motifs described by Baringou et al.[13] and classify the *H. azteca* HSP70s according to their framework (Table S4.8.2). According to these motifs and the classification methods described, all *H. azteca* sequences belong to Group A, which agrees with Baringou et al.[13] finding that all amphipod HSP70s characterized to date are Group A proteins. One gene contained slightly different motif characteristics and was grouped with A4 proteins, while the remaining sequences were grouped together in A5.

The *hsp70* genes appear to be evolving rapidly. We see an expansion in the A5 family and loss of genes on scaffold 108 and 503. Gene duplication within a species appears to be a common mechanism in *hsp70* gene evolution and has been described in *Drosophila melanogaster* and mosquitos.[15, 16] The redundancy in *hsp* genes may be related to tissue specific or stress response expression patterns.[2]

*Induction of HSPs by Cd exposure*

Members of all four families of large heat shock proteins were induced by Cd exposure (Table S4.8.1). Interestingly, although all genes showed some level of induction, the induction level varied between family members, most notably for the HSP40 DNAJ family where expression levels varied from 2 fold to 13 fold. Of the heat shock 70 family, all family members showed high levels of induction with the exception of HSP70A4, which is the only member within the A4 subfamily. Investigation of *cis* regulatory regions, particularly patterns of upstream heat shock elements (HSEs), in these *hsp* genes may help to explain expression differences in response to stress.[17]

Heavy metal exposure is a known inducer of HSP gene expression across taxa.[11] More specifically, Cd has been shown to induce the expression of multiple heat shock proteins in different arthropods including *Daphnia magna*,[18, 19] collembolan,[20] chronomids,[21] and moths;[22] although patterns differ across species. In *H. azteca* HSP60 and HSP70 protein induction was examined using polyclonal antibodies following exposure to Cd and pesticides.[23] Their results shown that concentrations as low as 0.5 μg/L of Cd are sufficient to increase protein levels of these genes. At 5 μg/L, the concentration used in the present study, protein levels of HSP70 and HSP60 increased 10-fold and 3-fold respectively, showing a high level of consistency with the increased mRNA expression seen in the present study.

The induction of HSPs by Cd is mediated through glutathione depletion and oxidative stress, where increased oxidation of native proteins contributes to misfolding, likely triggering the heat shock response.[24, 25] This model is also supported by the upregulation of glutathione S-transferase genes (HAZT011637 and HAZT011815) and several genes involved in protein degradation including ubiquitin-conjugating enzyme (HAZT002980) and cathepsin (HAZT011233) (see Table S4.1). Extensive evidence has shown that Cd reduces glutathione levels, increases the concentration of reactive oxygen species, and leads to protein, lipid, and

DNA oxidation.[26] However, because Cd is not a direct oxidant, the mechanisms leading to antioxidant/oxidant imbalance is not clear.[26] Glutathione is suspected to play an important role, and Cd has been shown to bind to and become sequestered by glutathione.[25]

## Conclusions

This analysis aimed to identify and classify the major heat shock proteins in the genome of *H. azteca.* Although previous studies have investigated changes in protein expression of HSPs in amphipods,[23] and many HSPs have been cloned in decapods, shrimp, and amphipods, to our knowledge, this is the first full characterization of heat shock proteins in the genome of a Malacostracan species. Given the importance of this species for environmental monitoring, the present identification and description of HSPs will aid in the further characterization of the stress response and how specific HSPs respond to pollution. In the present study, we found recent gene duplications for both *hsp90* and *hsp70* found both within gene clusters and across scaffolds. While similar recent duplications and gene arrangements have been documented in other species, the *H. azteca* species complex provides an exciting opportunity to study the evolution of HSPs and the corresponding heat shock response across closely related species inhabiting a variety of aquatic habitats.

| Gene name/ symbol | Fold induction by Cd | Scaffold Location:start site (HAZT 1.0) | Scaffold Location:start site (HAZT 2.0) | OGS gene ID | NCBI predicted protein ID |
|---|---|---|---|---|---|
| *hsp40A-1* | **1.9** | scaffold338:817915 | scaffold1600:89484 | HAZT007650 | XP_018153871 |
| *hsp40B-4* | **13.0** | scaffold233:116402 | scaffold1357:59622 | HAZT006327 | XP_018152140 |
| *hsp40C-3* | **2.0** | scaffold258:174854 | scaffold491:161102 | HAZT006641 | XP_018165231 |
| *hsp60* | **2.6** | scaffold385:452842 | scaffold592:179650 | HAZT008078 | XP_018167156 |
| *hsp70A4* | unknown | scaffold277:119001 | scaffold288:98121 | HAZT006856 | XP_018015331 |
| *hsp70A5b1* | **6.5** | scaffold72:76105 | scaffold1272:55811 | HAZT012276 | XP_018006997 |
| *hsp70A5b2* | **6.9** | scaffold277:198422 | scaffold288:135970 | HAZT006860 | XP_018015329 |
| *hsp70A5b3* | **7.7** | scaffold277:187530 | scaffold288:125239 | HAZT006859 | XP_018015327 |
| *hsp70A5b4* | **7.1** | scaffold277:187530 | scaffold288:115903 | HAZT006858 | XP_018015326 |
| *hsp70A5a1* | **8.8** | scaffold42:324876 | scaffold1872:22790 | HAZT001905 | XP_018011061 |
| *hsp70A5a2* | **12.7** | scaffold277:204948 | scaffold288:145691 | HAZT006861 | XP_018015328 |
| *hsp70A5a3* | **10.6** | scaffold1878:27104 | scaffold3410:25425 | HAZT011866 | XP_018017214 |
| *hsp90A-1* | **5.0** | scaffold199:954622 | scaffold595:132235 | HAZT005785 | XP_018167196 |
| *hsp90A-2* | **4.6** | scaffold199:961409 | scaffold1271:9457 | HAZT005786 | XP_018167194 |
| *hsp90B* | **10.8** | scaffold2202:39003 | scaffold595:137575 | HAZT011952 | XP_018151502 |

**Table S4.8.1: Heat Shock Proteins identified within the *Hyalella azteca* genome**. High molecular weight heat shock proteins were identified through tblast searches with HSPs from related species and blastn searches of differentially expression transcripts following Cd exposure. HSP70 sequences were named based on the classification of Baringou et al.[13] Location of each gene on the Hazt 1.0 and Hazt 2.0 assemblies are provided along with the official gene ID and corresponding protein ID from the automated annotation in NCBI.

| | NBD | | | | Linker | | SBD | CTD | | | | | Localization Tag |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | motif 1 | motif 2 | motif 3 | motif 4 | motif 5 | motif 6 | motif 7 | motif 8 | motif 9 | motif 10 | motif 11 | motif 12 | |
| | 113 | 187-189 | 213//214 | 359 | 382-387 | 388-393 | 543 | 552-557 | 561-575 | 599-602 | 605 | 610 - 630 | 630-643 |
| reference (KU613184) | Y | VGG | - | K | KSEAVQ | DLLLLDV | Y | EDEKFK | SSTDRSKILDA | QICN | I | GGAP (x2)<br>GGFP (x1) | EEVD |
| HSP70A4 | T | VGG | - | K | KSEAVQ | DLLLLDV | Y | EDDKVK | SEEDRKKIMEACDEA | KVCT | I | GGMP (x1) | EEVD |
| HSP70A5b1 | S | GTG | - | K | KSEAVQ | DLLLLDV | Y | EDDKLK | PEEELKKALGACSKA | NICS | I | GGMP (x1) | ERVD |
| HSP70A5b2 | S | GTG | - | K | KSEAVQ | DLLLLDV | Y | EDDKLK | SEEERKKALDACSEA | KICS | I | none | none |
| HSP70A5b3 | S | GTG | - | K | KSEAVQ | DLLLLDV | Y | EDDKLK | PEEERKKALDACSEA | KICS | I | GGMP (x3)<br>GGVP (x1) | EEVD |
| HSP70A5b4 | S | GTG | - | K | KSEAVQ | DLLLLDV | Y | EDDKLK | PEEDRKKALDACSEA | KICS | I | GGMP (x3) | EEVD |
| HSP70A5a1 | S | GSSAG | - | K | KSEAVQ | DLLLLDV | Y | EDDKVK | SENDRKKALDA | KVCT | I | GGMP (x1) | EEVD |
| HSP70A5a2 | S | GSTAG | - | K | KSEAVQ | DLLLLDV | Y | EDDKLK | PENDREKALNA | KVCA | I | GGMP (x1) | EEVD |
| HSP70A5a3 | S | GSTMG | - | K | KSEAVQ | DLLLLDV | Y | EDEKLK | EDDRKKALDACNDA | QVCA | I | none | EEVD |
| HSP70 like (XP_018017811) | S | LK | | Q | DKT--IE | NIKFVDV | L | SEQRDQ | EEDELDKFMNV | EVKE | L | none | none |

**Table S4.8.2: Alignment and classification of *H. azteca* HSP70 proteins based on 13 protein motifs found in Arthropods.** All eight *H. azteca* HSP70 genes were aligned to the reference HSP70A gene from *Procambarus clarkii* (KU613684) and corresponding protein motifs were identified in these genes. A representative HSP70-like sequence from scaffold 401 (XP_0180017811) was also aligned to *P. clarkia* HSP70A to determine which motifs were present in this related sequence.

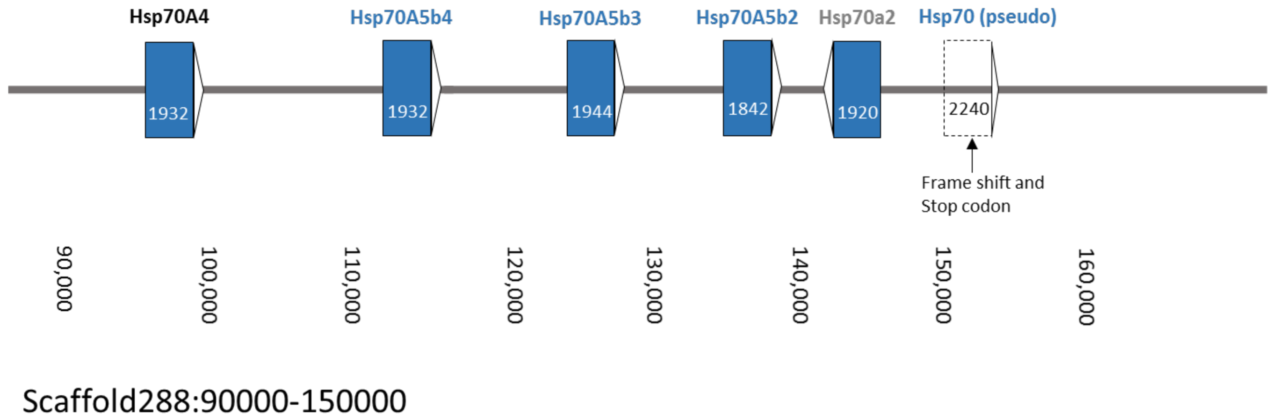



**Figure S4.8.1: The structure and arrangement of *hsp70* genes on Scaffold 288 of the Hazt 2.0 assembly.** Coding regions of *hsp70* genes identified on scaffold 277 of Hazt 1.0 assembly were blasted against the Hazt 2.0 assembly and resulted in the identification of a gene cluster on scaffold 288. Annotation of the genes on this assembly was able to close gaps within sequences found in the Hazt 1.0 assembly and identify the sixth gene as a likely pseudogene.

**S4.8. References:**


(1) Zhao, L.; Jones, W. Expression of heat shock protein genes in insect stress responses. *Invertebrate Surviv J* **2012,** *90*, 93-101.
(2) Daugaard, M.; Rohde, M.; Jäättelä, M. The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions. *FEBS letters* **2007,** *581*, (19), 3702-3710.
(3) Richter, K.; Haslbeck, M.; Buchner, J. The heat shock response: life on the verge of death. *Molecular Cell* **2010,** *40*, (2), 253-266.
(4) Lindquist, S.; Craig, E. The heat-shock proteins. *Annual Review of Genetics* **1988,** *22*, (1), 631-677.
(5) Feder, M. E.; Hofmann, G. E. Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. *Annual Review of Physiology* **1999,** *61*, (1), 243-282.
(6) Sørensen, J. G. Application of heat shock protein expression for detecting natural adaptation and exposure to stress in natural populations. *Current Zoology* **2010,** *56*, (6), 703-713.
(7) Nguyen, A. D.; Gotelli, N. J.; Cahan, S. H. The evolution of heat shock protein sequences, cis-regulatory elements, and expression profiles in the eusocial Hymenoptera. *BMC Evolutionary Biology* **2016,** *16*, (1), 15.
(8) Bedulina, D.; Evgen'Ev, M.; Timofeyev, M.; Protopopova, M.; Garbuz, D.; Pavlichenko, V.; Luckenbach, T.; Shatilina, Z.; Axenov-Gribanov, D.; Gurkov, A. Expression patterns and organization of the hsp70 genes correlate with thermotolerance in two congener endemic amphipod species (*Eulimnogammarus cyaneus* and *E. verrucosus*) from Lake Baikal. *Molecular Ecology* **2013,** *22*, (5), 1416-1430.
(9) Cascella, K.; Jollivet, D.; Papot, C.; Léger, N.; Corre, E.; Ravaux, J.; Clark, M. S.; Toullec, J.-Y. Diversification, evolution and sub-functionalization of 70kDa heat-shock proteins in two sister species of Antarctic krill: differences in thermal habitats, responses and implications under climate change. *PLoS One* **2015,** *10*, (4), e0121642.
(10) Mukhopadhyay, I.; Nazir, A.; Saxena, D.; Chowdhuri, D. K. Heat shock response: hsp70 in environmental monitoring. *Journal of Biochemical and Molecular Toxicology* **2003,** *17*, (5), 249-254.
(11) Lewis, S.; Handy, R.; Cordi, B.; Billinghurst, Z.; Depledge, M. Stress proteins (HSP's): methods of detection and their use as an environmental biomarker. *Ecotoxicology* **1999,** *8*, (5), 351-368.
(12) BLAST at the i5k Workspace@NAL. https://i5k.nal.usda.gov/webapp/blast/
(13) Baringou, S.; Rouault, J.-D.; Koken, M.; Hardivillier, Y.; Hurtado, L.; Leignel, V. Diversity of cytosolic HSP70 Heat Shock Protein from decapods and their phylogenetic placement within Arthropoda. *Gene* **2016,** *591*, (1), 97-107.
(14) Karlin, S.; Brocchieri, L. Heat shock protein 70 family: multiple sequence comparisons, function, and evolution. *Journal of Molecular Evolution* **1998,** *47*, (5), 565-577.
(15) Bettencourt, B. R.; Feder, M. E. Hsp70 duplication in the *Drosophila melanogaster* species group: how and when did two become five? *Molecular Biology and Evolution* **2001,** *18*, (7), 1272-1282.
(16) Gross, T. L.; Myles, K. M.; Adelman, Z. N. Identification and characterization of heat shock 70 genes in *Aedes aegypti* (Diptera: Culicidae). *Journal of Medical Entomology* **2014,** *46*, (3), 496-504.
(17) Åkerfelt, M.; Morimoto, R. I.; Sistonen, L. Heat shock factors: integrators of cell stress, development and lifespan. *Nature reviews. Molecular Cell Biology* **2010,** *11*, (8), 545.
(18) Connon, R.; Hooper, H. L.; Sibly, R. M.; Lim, F. L.; Heckmann, L. H.; Moore, D. J.; Watanabe, H.; Soetaert, A.; Cook, K.; Maund, S. J.; Hutchinson, T. H.; Moggs, J.; De Coen, W.;

Iguchi, T.; Callaghan, A. Linking molecular and population stress responses in *Daphnia magna* exposed to cadmium. *Environ Sci Technol* **2008,** *42*, (6), 2181-8.
(19) Poynton, H. C.; Lazorchak, J. M.; Impellitteri, C. A.; Smith, M. E.; Rogers, K.; Patra, M.; Hammer, K. A.; Allen, H. J.; Vulpe, C. D. Differential gene expression in *Daphnia magna* suggests distinct modes of action and bioavailability for ZnO nanoparticles and Zn ions. *Environ Sci Technol* **2011,** *45*, (2), 762-8.
(20) Nota, B.; Timmermans, M. J.; Franken, O.; Montagne-Wajer, K.; Mariën, J.; Boer, M. E. d.; Boer, T. E. d.; Ylstra, B.; Straalen, N. M. v.; Roelofs, D. Gene expression analysis of collembola in cadmium containing soil. *Environmental Science & Technology* **2008,** *42*, (21), 8152-8157.
(21) Planelló, R.; Martínez-Guitarte, J.; Morcillo, G. Effect of acute exposure to cadmium on the expression of heat-shock and hormone-nuclear receptor genes in the aquatic midge Chironomus riparius. *Science of the Total Environment* **2010,** *408*, (7), 1598-1603.
(22) Wang, H.; Li, K.; Zhu, J. Y.; Fang, Q.; Ye, G. Y. Cloning and expression pattern of heat shock protein genes from the endoparasitoid wasp, Pteromalus puparum in response to environmental stresses. *Archives of Insect Biochemistry and Physiology* **2012,** *79*, (4-5), 247-263.
(23) Werner, I.; Nagel, R. Stress proteins HSP60 and HSP70 in three species of amphipods exposed to cadmium, diazinon, dieldrin and fluoranthene. *Environ Toxicol Chem* **1997,** *16*, (11), 2393-2403.
(24) Abe, T.; Konishi, T.; Katoh, T.; Hirano, H.; Matsukuma, K.; Kashimura, M.; Higashi, K. Induction of heat shock 70 mRNA by cadmium is mediated by glutathione suppressive and non-suppressive triggers. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1994,** *1201*, (1), 29-36.
(25) Gaubin, Y.; Vaissade, F.; Croute, F.; Beau, B.; Soleilhavoup, J.-P.; Murat, J.-C. Implication of free radicals and glutathione in the mechanism of cadmium-induced expression of stress proteins in the A549 human lung cell-line. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **2000,** *1495*, (1), 4-13.
(26) Bertin, G.; Averbeck, D. Cadmium: cellular effects, modifications of biomolecules, modulation of DNA repair and genotoxic consequences (a review). *Biochimie* **2006,** *88*, (11), 1549-1559.

## S4.9. Insecticide target genes: Acetylcholinesterase and Voltage-gated Sodium Channel

Kaley M. Major and Helen C. Poynton
School for the Environment, University of Massachusetts Boston, Boston MA

Monica C. Muñoz-Torres
Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA,

Correspondence to: helen.poynton@umb.edu

### Introduction

Aceytlcholinesterase (AChE, EC3.1.1.7) is the enzyme responsible for the termination of neurotransmission via the hydrolysis of acetylcholine to choline and acetate. It is an essential enzyme in both the central and peripheral nervous systems of animals, and is a member of the serine hydrolase family.[1] Most insects have multiple forms of AChE,[2] of which *ace-1* (also called *para-ace* by the original *Drosophila melanogaster* nomenclature[3]) is predominantly responsible for synaptic function.[4]

Aside from its essential biological function in insects, AChE serves as the target site for enzyme inhibitors including organophosphates and carbamates.[5] Chemicals in these pesticide classes inhibit acetylcholinesterase by covalently bonding to the serine in the active site, part of the catalytic triad,[6] causing an abundance of acetylcholine in synapses and eventually organism death.[7] However, active-site mutations in the *ace*-1 in insects have become common in conferring organophosphate resistance in target pests, although *ace*-2 mutations have sometimes been implicated as another source of resistance as well (for a review, see Fournier et al.[5]). Recently, wild populations of *H. azteca* have been identified in areas where sediments harbor up to five times the lethal concentration of an organophosphate (chlorpyrifos) for to laboratory-exposed organisms,[8] possibly indicating the development of organophosphate resistance mechanisms in some wild *H. azteca*. Identification of the *ace* gene(s) within the *H. azteca* genome would provide an opportunity to quantify target site changes that may be occurring in resistant *H. azteca* compared to sensitive ones. The selection for target-site mutations conferring organophosphate resistance in these nontarget organisms could have implications for both ecotoxicology and evolutionary biology.

The voltage-gated sodium channel (VGSC or Nav) evolved in basal metazoans and served the need for rapid signal transmittance in the evolving nervous system.[9] Its primary role within the peripheral and central nervous system is the creation of the action potential along the axon of neurons. The VGSC consists of four domains, each with six transmembrane segments. This structure forms a pore for passage of $Na^+$, and contains two gates that respond to the membrane depolarization and membrane potential.[10]

Its essential role in the nervous system has made the VGSC a particularly useful target for both naturally evolved and synthetic neurotoxins including tetrodotoxin (TTX), dichlorodiphenyltrichloroethane (DDT), pyrethrin, and the synthetic pyrethroid insecticides.[9] DDT and pyrethroids impede closing of the channel (i.e. gate functioning), causing hyperexcitability and repetitive firing. Eventually the nerve cells become exhausted, resulting in incapacitation of the organism.[10] However, many species have evolved resistance to these neurotoxins through mutations in the VGSC, with over 120 documented cases of insecticide resistance developing across arthropoda, mostly in pest species,[11] but also in *H. azteca.*[8] Although partial sequences of the *vgsc* are available through cloning efforts, the complete sequence of the *vgsc* will greatly enhance efforts to document cases of insecticide resistance in the non-target crustacean, *H. azteca* and potentially facilitate efforts to identify resistance in other related crustaceans.

**Methods**

We identified the *H. azteca ace*-1 gene by blasting (tblastn) the *ace*-1 and *ace*-2 protein sequences of a two crustaceans, *Tigriopus japonicus* (NCBI GenBank accession numbers AIU38228 and AIU38229), and *Lepeophtheirus salmonis* (AIY62313 and AIY62314), against the *H. azteca* genome.[12] These sequences each aligned best to *H. azteca* gene model HaztTmpM010125-RA, located on Scaffold409. This gene model originally predicted only two exons for the acetylcholinesterase *H. azteca* homolog, both on the minus DNA strand. However, the addition of RNA Seq data and the assembled transcriptome (described elsewhere) revealed support for an additional, untranslated exon upstream of the previously-predicted exons.

Once the gene model was manually annotated, we blasted it against the NCBI GenBank database and found it to be most similar to *ace*-1, with best hits to *Liposcelis bostrychophila, Tribolium castaneum, and Alphitobius diaperinus* with 59-60% similarity (E-value=0.0). To further the search for a possible *ace*-2 gene in *H. azteca*, the *ace*-2 genes of *T. japonicus* (AIU38229)*, T. castaneum* (ADU33190)*, and Daphnia magna* (JAN89707) were blasted against the assembled *H. azteca* transcriptome, with transcript TR51702 in *H. azteca* being the best match. Next, transcript TR51702 was blasted against the *H. azteca* genome. It aligned to a gene model on Scaffold 89, which was blasted against the NCBI GenBank database to reveal that it was instead a carboxylesterase with no homology to *ace*-2.

To identify the *H. azteca* voltage-gated sodium channel (*vgsc*), partial amino acid sequences of this gene (available in [8]) were aligned to the *H. azteca* genome in Apollo using BLAT. These sequences aligned to a single gene model, HaztTempA006232_RA located on scaffold 282. Blasting of this gene model against the NCBI nr database showed homology to arthropod voltage-gated sodium channels including *Cancer borealis* ABL10360.2, but also suggested that large portion of the upstream region of this gene was missing from the *H. azteca* gene model. Two upstream models (HaztTempA006233_RA and HaztTempA006242_RA) were then blasted against NCBI nr database which predicted that these models were also part of the *vgsc*. These three gene models were then merged, exon boundaries were modified to produce canonical splice sites, and exons were deleted or added based on RNAseq evidence.

**Results and Discussion**

*Acetylocholinesterase* - The annotated *H. azteca ace*-1 gene is composed of three exons on the minus (-) DNA and two coding regions that produce a 563 amino acid protein product (Figure 3.9.1). Although the HaztTmpM010125-RA gene model only predicted two exons, both RNA Seq and assembled transcriptome datasets support the presence of three exons with the inclusion of a 47-bp 5' untranslated region (UTR) exon. No viable translation start sites exist in the vicinity of the upstream, 47-bp exon, making it more likely to be a 5' UTR region than a translated portion of the protein product.

Although other crustaceans have two homologous *ace* genes, including crustacean relatives *D.magna, T. japonicus,* and *L. salmonis*, no evidence of an *ace*-2 gene exists for *H. azteca*. This conclusion is based on 1) the tblastn results that aligned both the *ace*-1 and *ace*-2 genes of the aforementioned crustaceans to only a single gene model (most similar to *ace*-1 in other organisms) in *H. azteca*, and 2) the transcriptome blast and subsequent NCBI GenBank blast that that identified the gene model for transcript TR51702 in *H. azteca* as a carboxylesterase instead of an *ace*-2 homolog.

The existence of only an *ace*-1 homolog in *H. azteca* is unique in that many other crustaceans seem to retain both the paralogous (*ace*-1) and the orthologous (*ace*-2) *ace* genes. Further, the presence of only a single *para-ace* suggests only one primary target site for organophosphate binding and toxicity in *H. azteca*, allowing for a more directed approach for the determination of the source of organophosphate resistance potentially conferred via target-site mutation.

*Voltage-gated sodium channel-* The coding region of the *vgsc* lies on scaffold 282 and spans over 40,000 bp and includes 26 exons (see Figure S4.9.2). It was initially divided into three separate gene models that were merged based on their homology to insect and crustacean *vgsc* genes. The full sequence of the *vgsc* provided here will greatly increase our ability to screen for resistance mutations in *H. azteca.* For example, previous sequencing efforts of the *vgsc* [8] only partially covered the 41 amino sites associated with pyrethroid resistance in insects.[11]

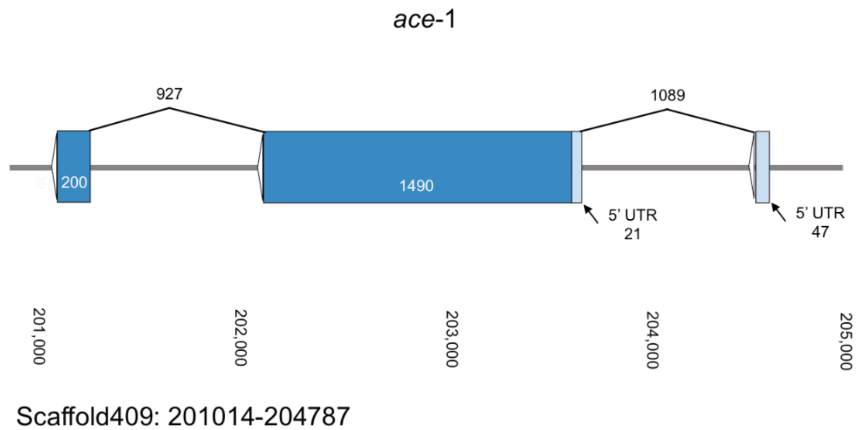



**Figure S4.9.1**: *Ace*-1 gene model based on HaztTmpM010125-RA with modifications supported by RNA Seq and assembled trancsriptome data. Boxes indicate exons. Dark blue portions of exons are coding regions and light blue boxes are 5' UTRs.

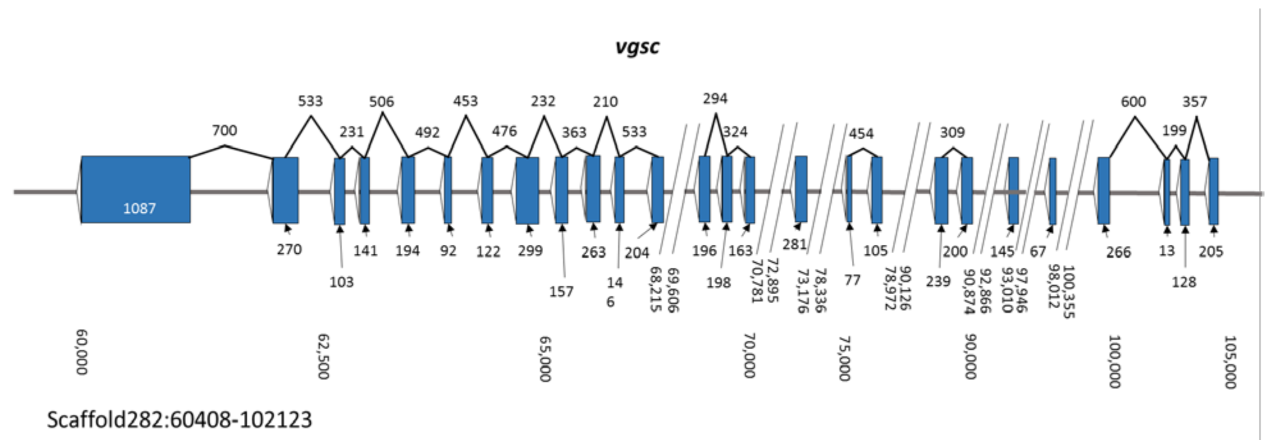



**Figure S4.9.2**: *Vgsc* gene model based on merging of HaztTempA006232, HaztTempA006233, and HaztTempA00624 with modifications supported by RNA Seq and assembled trancsriptome data. Dark blue boxes indicate exons.

### S4.9. References


(1) Sussman, J. L.; Harel, M.; Frolow, F.; Oefner, C.; Goldman, A.; Toker, L.; Silman, I. Atomic structure of acetylcholinesterase from *Torpedo californica*: A prototypic acetylcholine-binding protein. *Science* **1991,** *253*, 872-879.

(2) Lu, Y.; Pang, Y. P.; Park, Y.; Gao, X.; Yao, J.; Zhang, X.; Zhu, K. Y. Genome organization, phylogenies, expression patterns, and three-dimensional protein models of two acetylcholinesterase genes from the red flour beetle. *PLoS One* **2012,** *7*, (2), e32288.

(3) Hall, L.; Spierer, P. The Ace locus of *Drosophila melanogaster*: structural gene for acetylcholinesterase with an unusual 5'leader. *The EMBO journal* **1986,** *5*, (11), 2949.

(4) Kaur, K.; Bakke, M. J.; Nilsen, F.; Horsberg, T. E. Identification and Molecular Characterization of Two Acetylcholinesterases from the Salmon Louse, *Lepeophtheirus salmonis*. *PLoS One* **2015,** *10*, (5), e0125362.

(5) Fournier, D. Mutations of acetylcholinesterase which confer insecticide resistance in insect populations. *Chem Biol Interact* **2005,** *157-158*, 257-61.

(6) Quinn, D. M. Acetylcholinesterase: Enzyme structure, reaction dynamics, and virtual transition states. *Chem. Rev.* **1987,** *87*, (955-979).

(7) Aldridge, W. Some properties of specific cholinesterase with particular reference to the mechanism of inhibition by diethyl p-nitrophenyl thiophosphate (E 605) and analogues. *Biochemical Journal* **1950,** *46*, (4), 451.

(8) Weston, D. P.; Poynton, H. C.; Wellborn, G. A.; Lydy, M. J.; Blalock, B. J.; Sepulveda, M. S.; Colbourne, J. K. Multiple origins of pyrethroid insecticide resistance across the species complex of a nontarget aquatic crustacean, *Hyalella azteca*. *Proc Natl Acad Sci U S A* **2013,** *110*, (41), 16532-7.

(9) Zakon, H. H. Adaptive evolution of voltage-gated sodium channels: the first 800 million years. *Proc Natl Acad Sci U S A* **2012,** *109*, (Supplement 1), 10619-10625.

(10) Davies, T. G.; Field, L. M.; Usherwood, P. N.; Williamson, M. S. DDT, pyrethrins, pyrethroids and insect sodium channels. *IUBMB Life* **2007,** *59*, (3), 151-62.

(11) Dong, K.; Du, Y.; Rinkevich, F.; Nomura, Y.; Xu, P.; Wang, L.; Silver, K.; Zhorov, B. S. Molecular biology of insect sodium channels and pyrethroid resistance. *Insect Biochemistry and Molecular Biology* **2014,** *50*, 1-17.

(12) BLAST at the i5k Workspace@NAL. https://i5k.nal.usda.gov/webapp/blast/

# S4.10. Gene families of enzymes involved in ion transport: sodium/hydrogen antiporter (NHA), V-type proton ATPase (VHA), and carbonic anhydrase (CA)

Nicholas Mathers and Carol Eunmi Lee
Center of Rapid Evolution (CORE), University of Wisconsin Madison, Madison, WI 53706

Correspondence to carollee@wisc.edu

## Introduction

Ion transport proteins regulate ionic concentrations in cells and organelles through the passive or active exchange of ions across biological membranes. These transport proteins control the movement of substrates such as of inorganic ions, amino acids, nucleotides, sugars, metabolites, and pharmaceuticals.[1] A subset of membrane bound ion transporters and intracellular enzymes play integral roles in basic biological processes of maintaining cellular homeostasis and regulating epithelial transport of common ions such as $H^+$, $Na^+$, $K^+$, and $Cl^-$ in arthropods.[2-4] [5] Despite the fundamental importance of these proteins in all living cells, the specific functions of many of these ion transporters and the genes encoding them remain poorly characterized.

The goal of this study was to annotate and characterize some key proteins involved in ion transport and regulation in the genome of the amphipod *Hyalella azteca*. This amphipod has a broad geographic range and inhabits a wide range of habitat types (see introduction in main text). An increased physiological demand for ion uptake and regulation in *H. azteca* would be expected, given its wide distribution and ecological success in freshwater and brackish habitats, as well as habitats that vary in pH.

*H. azteca* has a broad geographical and ecological range. This small crustacean is a common and widely distributed freshwater amphipod in North America, with populations of the *Hyalella* species complex also found in Central and northern South America.[6-8] In addition to its broad geographic range, *H. azteca* is distributed across a wide range of habitat types. *H. azteca* has the ability to tolerate a relatively wide range of salinities, ranging from freshwater (0 PSU) to brackish waters (up to 15 PSU).[9] The species can also tolerate a wide water pH, ranging from an acidic threshold near pH 4.5[6] up to an alkaline pH of 9.3.[10]

Some ion transporters and enzymes hypothesized to be crucial for ion uptake from dilute environments (i.e., freshwater) and intracellular pH regulation include the proton pump V-type $H^+$ ATPase (VHA, ATP6), the sodium/hydrogen antiporter (NHA, SLC9B), and carbonic anhydrase (CA).[2-4] [5] Of these ion transporters, the proton pump V-type $H^+$ ATPase (VHA, ATP6) was first known to be involved in pH regulation,[11] and then later recognized to play a crucial role in energizing ion uptake into the cell.[3, 5, 12] The secondary transporter, "sodium/hydrogen antiporter" (NHA, SLC9B), has been far less studied in animals and its function is less clear, but it might in some cases cooperate with VHA to transport cations into the cell.[13-16] In addition, the intracellular carbonic anhydrases (CA) form a family of enzymes that catalyze the reaction of carbon dioxide and water to bicarbonate and protons (or *vice versa)*. Thus, carbonic anhydrases supply protons to V-type H+ ATPase to facilitate ion uptake into the cell.[3, 4, 16, 17]

The proton pump V-type $H^+$ ATPase (VHA, ATP6) is an evolutionarily conserved, but functionally dynamic molecular machine having a wide range of functions. This proton pump is found embedded in the membranes of cells as well as of many organelles, such as endosomes, lysosomes, and secretory vesicles.[18] The ability of VHA to actively translocate $H^+$ across the membranes of cells or organelles allows it to generate electrochemical $H^+$ gradients.[11] These proton gradients could then drive $H^+$-coupled substrate transport of common bioavailable cations ($Na^+$, $K^+$, $Li^+$) and regulate other protein activity that is dependent on pH, such as catalytic activity or protein-protein interactions.[19] VHA has been found to be involved in ion uptake in several arthropods, such as mosquito larvae and some crustaceans,[4, 14] by generating transmembrane potentials in epithelial ion regulatory cells. These transmembrane potentials can be used by secondary ion transporters or channels to deliver cations into the cell.[18]

VHA is a large, two domain protein complex (V1 and V0). The cytoplasmic domain (V1) is responsible for ATP hydrolysis and is composed of 8 subunits (A, B, C, D, E, F, G). The membrane-bound, proton conducting domain (V0) is composed of 5 subunits (a, c, c'', d, e). These 13 VHA subunits are ubiquitous in eukaryotes and is thought to be expressed in virtually every eukaryotic cell.[18] There are also two accessory subunits in some taxa but their distribution across the Metazoa is poorly understood. In *Drosophila melanogaster*, the best studied animal model for VHA, there are 33 genes that encode the 15 subunits (including accessories). Out of the 15 subunits described in *D. melanogaster*, only subunits B, C, E, G, and H have a single gene copy each, while the other subunits have between two and five paralogs.[20]

The sodium/hydrogen antiporters (NHA, SLC9B2, CPA2) are a subfamily of transmembrane ion transporters, which was only recently discovered in animal genomes in 2005 and first cloned for characterization in mosquito larvae in 2007.[21-23] NHAs are a highly divergent, but understudied, gene family. NHAs are essential for life in *Drosophila melanogaster*, but their actual ion specificity, stoichiometry, and functions throughout eukaryotic evolution are poorly understood.[15, 23, 24] A previous phylogenetic analysis distinguished NHAs from the more extensively studied electroneutral (i.e., $1Na^+/1H^+$) sodium/hydrogen exchangers.[21] NHAs form a distinct lineage of ion transporters, found in every fully sequenced metazoan genome, and sharing an ancient common ancestor with electrogenic (i.e., $2Na^+/1H^+$) prokaryotic and fungal transporters.[21] In both arthropods and mammals, evidence indicates that NHA is coupled to VHA as a secondary electrogenic transporter for ion uptake against concentration gradients. This coupling has been observed in insect Malpighian tubules and larval midguts, and also in human kidney cells and osteoclasts.[14-16, 22, 25] Two orthologs of NHA genes are present in Diptera (flies), but the number of NHA copies across Arthropods varies, leading to the absence of clear sequence orthology.

Metazoan carbonic anhydrases (CA) are ubiquitous zinc-containing metalloenzymes that catalyze the interconversion of carbon dioxide and bicarbonate: $H_2O + CO_2 \rightleftarrows HCO3^- + H^+$.[26] Two of the CA groups, alpha-CAs and beta-CAs, are present in metazoans. In *D. melanogaster*, alpha-CAs include both cytosolic and membrane-bound enzymes, while beta-CAs are thought to be localized in mitochondria. Because CAs catalyze the production of $H^+$, they can supply VHA with the protons required to generate a voltage gradient. Alpha-CAs have a high rate of enzymatic activity, making them important proteins in the regulation of pH and ion transport. This high catalytic rate has been observed in the midgut and Malpighian tubules of insects, where VHA and NHA also have high activity. [27, 28] The functions of specific alpha-CAs are unknown, but insect subfamilies have been identified in a previous phylogenetic analysis.[29] In insects, CA has experienced extensive gene duplication and loss within various lineages, leading to a large variation in gene family size among species.[29] However, the distribution of CA genes across the Arthropoda has not been fully explored. The identification and classification of CAs in non-hexapod arthropods, such as *H. azteca*, may help understand the evolution of the CA gene family and infer the subfunctionalization of CA enzymes by resolving gene subfamilies.

Thus, the goals of this study were (1) annotate the NHA, VHA, and CA genes in the *H. azteca* genome, (2) establish homology of these genes to previously characterized genes in *D. melanogaster*, and (3) determine the level of genetic divergence in these gene families within and among *H. azteca* and *D. melanogaster*. Annotations of these gene sequences will be among the first in crustaceans and will extend the capacity for future comparative and functional studies of these gene families with respect to evolution of novel aquatic habitats and toxicology.

Ion uptake via VHA, NHA, and CA might be a key physiological mechanism underlying adaptation to low salinities, as freshwater and brackish organisms have an increased requirement for ion uptake. These three gene families may have important functions in *H. azteca* that contribute to their wide ecological range and tolerance of fresh and brackish waters. In addition, ionic and pH regulation could have indirect effects on metal and inorganic ion toxicity (see Discussion). As *H. azteca* is an important model for toxicological studies, characterizing these transport functions is important for understanding its physiological response to toxins, as well as to natural variation in salinity and pH.

**Methods and Materials**

*Sequence identification and annotation*

Peptide sequences from the orthologs of each gene family in *Drosophila melanogaster* were used as query sequences to tBlastn search the *Hyalella azteca* genome assembly database (Table S4.10.1). Loci containing tBlastn hits with an E-value < 0.1 were annotated as candidates for classification into the query family. If any annotations could be generated that produced a reciprocal Blastp result of E-value < $10^{-5}$ with a known homolog in the NCBI non-redundant protein sequences (nr) database, they were retained as putative gene candidates. The manual annotation of primary gene models was guided using mapped transcriptome data and homology to curated genes in NCBI's GenBank. The NCBI Conserved Domain Database was used to determine the completeness of each sequence by comparing it to previously defined conserved protein domains for the corresponding protein family.[30] The nomenclature of VHA genes was designated according to homology the mammalian system. Due to the lack of clear orthology, NHA and CA genes were named arbitrarily in the order that they were identified.

*Sequence alignment and analysis*

Amino acid coding sequences of genes from *H. azteca* and *D. melanogaster* were aligned with the T-Coffee web server for multiple sequence alignments (tcoffee.crg.cat).[31] The PSI/TM-

Coffee algorithm was used for NHA sequences, as this algorithm was designed specifically for transmembrane proteins combining position specific iterative (PSI) BLAST search homology to a transmembrane protein library.[31] The PSI-Coffee algorithm, without the transmembrane extension, was used to align sequences of VHA subunits. The phylogeny-aware alignment algorithm PRANK was used to construct multiple sequence alignments of CAs.[32] Percentage identities, excluding gaps, were calculated from the multiple sequence alignments in Unipro UGENE.[33] The prediction of transmembrane protein helices in each amino acid sequence was performed on the TMHMM2.0 server.[34, 35]

## Results and Discussion

*Sodium/hydrogen antiporter (NHA) genes and their expansion in H. azteca*

Four complete and two partial sodium/hydrogen antiporter (NHA) genes were identified in *H. azteca.* Interestingly, all NHA search hits using tBlastn occurred within a single 260,000 bp region on scaffold 88, with all annotations occurring within a 162,000 bp region. HaztNHA1, HaztNHA3, and HaztNHA4 were found on the positive strand, while HaztNHA2 was localized on the negative strand.

The presence of four NHA genes in the *H. azteca* genome was unexpected, as only two NHA paralogs per genome had been found previously in animal genomes.[21] The four complete *H. azteca* NHAs formed a monophyletic clade in an arthropod-wide NHA phylogeny, suggesting they evolved after divergence of the main Arthropoda subphyla (Mathers & Lee, *In Prep*). These NHA gene duplications might be indicative of an adaptive function, as they had not been removed from the genome by purifying selection nor had they deteriorated through pseudogenization. The presence of unique mutations and divergent sequence identity between *H. azteca* NHA paralogs (Table S4.10.2) supported the existence of paralogs resulting from tandem gene duplications, rather than being artifacts of an assembly error. As more crustacean genomes become available, especially those within or directly outside of amphipods, it would be interesting to investigate whether these paralogs are found within *H. azteca* only or are more widespread among crustaceans.

There were numerous sequence fragments that shared homology with NHA, but had either missing or fragmented transcriptome data. As these NHA-like fragments did not resemble complete, functional genes, they were not annotated in the *H. azteca* genome. There were no non-NHA gene model predictions present on either strand of the genome within the approximately 100 kb region of complete, partial, and fragmented NHA sequences.

The four *H. azteca* NHA paralogs shared between 57% and 75% amino acid sequence identity to one another (excluding gaps) (Table S4.10.2, blue and green cells). In comparison, *H. azteca* NHAs were more distant from insect specific NHA orthologs represented by *Drosophila melanogaster* NHA1 and NHA2, sharing only between 26% and 41% amino acid identity (Table S4.10.2, yellow cells). All *H. azteca* NHAs were closer in sequence identity to *DmelNHA1* than to *DmelNHA2*.

The number of predicted transmembrane helices in *H. azteca* NHA amino acid sequences were 8, 7, 11, and 9 for *HaztNHA1*, *HaztNHA2*, *HaztNHA3*, and HaztNHA4, respectively. The *D. melanogaster* NHAs, *DmelNHA1* and *DmelNHA2*, both contained 12 predicted transmembrane helices. Amino acid sequences from all *H. azteca* and *D. melanogaster* NHA genes contained the conserved protein domains of the sodium/hydrogen exchanger superfamily (“Na_H_Exchanger”, accession: cl01133) and the NhaP-type sodium or potassium antiporter (NhaP, accession: COG0025) (highest E-value = $4.9^{-9}$).

*V-type $H^+$ ATPase subunits in the H. azteca genome*

We identified in the *H. azteca* genome 13 genes encoding each of the main subunits of the proton pump V-type H+ ATPase (VHA). All sequences were complete, except that of the V1G subunit (*Atp6v1g*). Evidence from mapped RNA-seq reads indicated the presence of a novel isoform of the V0d subunit (*Atp6v0d*), in which there was an additional exon relative to previously characterized arthropod *Atp6v0d* genes in BLASTp alignments. The sequence of this additional exon closely resembled the directly upstream exon, suggesting the presence of an intragenic exon duplication relative to other arthropod *Atp6v0d* genes.

There was substantial variation in the total number of VHA subunit genes among species and the level of sequence identity among subunits, despite generally high levels of sequence conservation. In contrast to VHA genes identified in *D. melanogaster*, genes of the two accessory subunits (*vhaAC45* and *vhaM8.9*) were missing from *H. azteca*. Genes for both of these accessory subunits were previously found in the human and mouse genomes, but only one of them (M8.9) was found in the nematode *C. elegans* and the yeast *Saccharomyces.*[20] This previous comparative analysis identified a wide range of VHA genes in the genomes of *D.*

*melanogaster* (33), human (24), mouse (24), *C. elegans* (19), *Arabidopsis* (28), and *Saccharomyces* (15). The high number of VHA genes identified in those organisms are in stark contrast to the only 13 VHA subunit genes present in *H. azteca*. The divergence of *H. azteca* VHA subunit genes ranged from 54% to 88% amino acid identity relative to their *D. melanogaster* homologs (Table S4.10.3). All VHA subunit genes were located on different scaffolds within the genome.

*Carbonic anhydrase genes in the H. azteca genome*

We found 12 alpha carbonic anhydrase genes in the *H. azteca* genome. However, we did not find the single beta carbonic anhydrase gene expected to be present in all invertebrates, but absent in vertebrates and chordates.[36] We did find alpha-like CAs of both catalytic and non-catalytic classes. The non-catalytic alpha-CAs, called "carbonic anhydrase related proteins" (CARPs), share a common ancestor with catalytic alpha-CAs prior to the evolution of the Arthropoda. They are genetically distinct and evolutionarily conserved, suggesting the conservation of some unknown function other than the known catalytic activity of alpha-CAs.[29]

We separated the 12 alpha-like CAs found in *H. azteca* into 8 putative catalytic alpha CAs and 4 inactive CARPs. There was one instance of an alpha-CA tandem gene duplication on scaffold 126 in the *H. azteca* genome (Hazt*CAH9* and *HaztCAH10*). *HaztCAH9* and *HaztCAH10* shared 88% amino acid identity, the most similar of any two carbonic anhydrase genes in either *H. azteca* or *D. melanogaster*. There were also two alpha-CAs (*HaztCAH1* and *HaztCAH2*) on scaffold 5, but they were not located in tandem. The remaining 8 CA genes were scattered throughout different scaffolds of the *H. azteca* genome.

The *H. azteca* genome contained 4 CARPs rather than the 2 found in *D. melanogaster*, but fewer normal alpha-CAs than in *D. melanogaster* (Figure S4.10.1). Both genomes had 7 alpha-CA genes when eliminating the alpha-CAs known to be *Drosophila* specific.[29] These genes could represent ancient alpha-CA subfamilies conserved across the Arthropoda or be more recently duplicated paralogs. Future phylogenetic inference is required to determine the orthologous relationships between these genes as the relationships could not be putatively determined due to high sequence divergence.

There was high divergence within both *H. azteca* and *D. melanogaster* among CARPs and among alpha-CAs*.* Amino acid sequence similarity among the *H. azteca* CARPs ranged between 27% and 51%, with an average similarity of 40%. The two D. melanogaster CARPs shared 54% similarity among themselves.

Alpha-CAs in the *H. azteca* genome were more conserved than those in the *D. melanogaster* genome. For *H. azteca* alpha-CAs, amino acid sequence similarity ranged between 36% and 88%, while that for *D. melanogaster* alpha-CAs ranged between 17% and 52%. The average similarity for *D. melanogaster* alpha-CAs was 27% whereas that among *H. azteca* alpha-CAs was 50%. The high similarity among *H. azteca* alpha-CAs suggested an expansion of alpha-CA genes after the divergence of insects. A more comprehensive phylogenetic analysis is required to determine the gene duplication and/or loss events of CAs across the major sublineages of the Arthropoda.

There was also high divergence in both CARPs and alpha-CAs between *H. azteca* and *D. melanogaster*. Interestingly, the uncharacterized CARPs were more conserved between these two species than putatively functional alpha-CAs, even though this gene family is considered to represent catalytically inactive carbonic anhydrases. The amino acid similarity shared between *H. azteca* and *D. melanogaster* CARPs ranged between 37% and 54%, with an average similarity of 45%. In normal alpha-CAs the average shared sequence similarity ranged between 21% and 38%, with an average of only 29%.

This was among first comprehensive efforts to annotate CAs in a non-hexapod pancrustacean. It was also the first time more than a single alpha carbonic anhydrase gene had been identified in a crustacean, with one alpha carbonic anhydrase each previously identified in the copepod *Caligus rogercresseyi* and the giant tiger prawn *Penaeus monodon*.[29] These results revealed that the alpha-CA gene family in arthropods is complex and warrants further study.

*The possible roles of VHA, NHA, and CA in response to toxicity*

In addition to adaptive function in freshwater, the functions of VHA, NHA, and CA in ion uptake and pH regulation may contribute to the complex effects of pH, ion concentrations, and salinity on metal toxicity. pH indirectly influences metal toxicity through the alteration of metal ion concentration and solubility, or by altering the electrochemical gradient at plasma membranes, thus affecting chemical speciation and bioavailability.[37] pH can directly affect toxicity via competition at the receptor responsible for metal ion uptake. However, the relationship of pH-dependent toxicity in freshwater organisms is complex and highly chemical and species specific, though a decrease in pH generally increases metal toxicity.[38] VHA and CA greatly contribute to the regulation of pH and the voltage at epithelial membranes that could alter the uptake or toxicity of metals.

Salinity and ion concentration of common inorganic ions also affect toxicity. In aquatic insects, high concentration of silver reduces $Na^+$ uptake, while the presence of copper increases $Na^+$ uptake.[39] In *H. azteca* and *D. magna*, an array of salts and inorganic ions, including sodium ($Na^+$) (a hypothesized substrate of NHA), alter the toxic effects of the metals Cd, Zn, and Cu.[37, 40, 41] The toxicity of insecticides is influenced by salinity, as studied using *H. azteca* and a cladocera crustacean *C. dubia*.[43] Chloride ion ($Cl^-$), another possible substrate of NHA, increases toxicity of sodium nitrate and sodium sulfate in one strain of *H. azteca*, but not in another.[42, 43] Thus, NHA may be a target of metal toxicity or alter transmembrane concentration gradients that influence toxicity via ion uptake, in concert with VHA and CA. Therefore, the aforementioned gene families might possibly perform functions closely intertwined with toxicity either directly (through transport of $Na^+$ or $Cl^-$ by NHA) or indirectly (through the regulation of pH and transmembrane electrical gradients by CA and VHA).

| *D. melanogaster* Gene (symbol) | NCBI Accession Number |
| --- | --- |
| **$Na^+/H^+$ antiporter (NHA)** | |
| $Na^+/H^+$ antiporter 1, isoform A (NHA1) | NP_723224.2 |
| $Na^+/H^+$ antiporter 2, isoform A (NHA2) | NP_732807.1 |
| **V-type H+ ATPase (VHA)** | |
| V-type H+ ATPase Complex V1 Subunit A, 68 kD subunit 2 (vha68-2, Atp6v1a) | NP_652004.2 |
| V-type H+ ATPase Complex V1 Subunit B, 55kD subunit (vha55, Atp6v1b) | NP_476908.1 |
| V-type H+ ATPase Complex V1 Subunit C, 44kD subunit (vha44, Atp6v1c) | NP_477266.1 |
| V-type H+ ATPase Complex V1 Subunit D, 36 kD subunit (vha36-1, Atp6v1d) | NP_570008.1 |
| V-type H+ ATPase Complex V1 Subunit E, 26 kD subunit (vha26, Atp6v1e) | NP_524237.1 |
| V-type H+ ATPase Complex V1 Subunit F, 14 kD subunit (vha14, Atp6v1f) | NP_476969.1 |
| V-type H+ ATPase Complex V1 Subunit G, 13 kD subunit (vha13, Atp6v1g) | NP_477437.1 |
| V-type H+ ATPase Complex V1 Subunit H, SFD subunit (vhaSFD, Atp6v1h) | NP_523585.2 |
| V-type H+ ATPase Complex V0 Subunit a, 100kD subunit 1 (vha100-1, Atp6v0a) | NP_733274.1 |
| V-type H+ ATPase Complex V0 Subunit c, 16kD subunit 1 (vha16-1, Atp6v0c) | NP_476801.1 |
| V-type H+ ATPase Complex V0 Subunit c'', PPA1 subunit 1 (vhaPPA1-1, Atp6v0c'') | NP_652010.1 |
| V-type H+ ATPase Complex V0 Subunit d, AC39 subunit 1 (vhaAC39-1, Atp6v0d) | NP_570080.1 |
| V-type H+ ATPase Complex V0 Subunit e. M9.7 subunit a (vhaM9.7-a, Atp6v0e) | NP_649327.2 |
| V-type H+ ATPase accessory subunit AC45 (vhaAC45, VAS1_Human) | NP_724770.1 |
| V-type H+ ATPase accessory subunit M8.9 (vhaM8.9, Atp6ap2) | NP_649876.1 |
| **Carbonic anhydrase (CA)** | |
| Alpha carbonic anhydrase 1 (CAH1) | NP_523561.1 |
| Alpha carbonic anhydrase 2, isoform A (CAH2) | NP_648555.1 |
| Beta carbonic anhydrase isoform A (CAHbeta) | NP_649849.1 |

**Table S4.10.1** Coding sequences of NHA, VHA, and CA of Drosophila melanogaster used as query sequences to identify homologs in the amphipod Hyalella azteca

| | DmelNHA1_A | DmelNHA2_A | HaztNHA1 | HaztNHA2 | HaztNHA3 | HaztNHA4 |
|---|---|---|---|---|---|---|
| DmelNHA1_A | 100% | 27% | 41% | 38% | 36% | 33% |
| DmelNHA2_A | 27% | 100% | 34% | 32% | 32% | 26% |
| HaztNHA1 | 41% | 34% | 100% | 71% | 73% | 75% |
| HaztNHA2 | 38% | 32% | 71% | 100% | 67% | 57% |
| HaztNHA3 | 36% | 32% | 73% | 67% | 100% | 61% |
| HaztNHA4 | 33% | 26% | 75% | 57% | 61% | 100% |

| Legend: | <30% | 30 – 50% | 50 – 70% | >70% |
|---|---|---|---|---|

**Table S4.10.1.** Distance matrix of percent amino acid identity, excluding gaps, obtained from a multiple sequence alignment of Drosophila melanogaster and Hyalella azteca sodium/hydrogen antiporters (NHA)

| VHA subunit genes in *H. azteca* | Percent identity to closest *D. melanogaster* homolog |
|---|---|
| Atp6v1a | 78% (vha68-2) |
| Atp6v1b | 88% (vha55) |
| Atp6v1c | 65% (vha44) |
| Atp6v1d | 66% (vha36-1) |
| Atp6v1e | 57% (vha26) |
| Atp6v1f | 72% (vha14) |
| Atp6v1g | 54% (vha13) |
| Atp6v1h | 61% (vhaSFD) |
| Atp6v0a | 67% (vha100-1) |
| Atp6v0c | 82% (vha16-1) |
| Atp6v0c’’ | 59% (vhaPPA1-1) |
| Atp6v0d | 77% (vhaAC39-1) |
| Atp6v0e | 54% (vhaM9.7a) |

**Table S4.10.2.** Percent amino acid sequence identity between H. azteca V-type H+ ATPase (VHA) subunits and Drosophila melanogaster homologs.

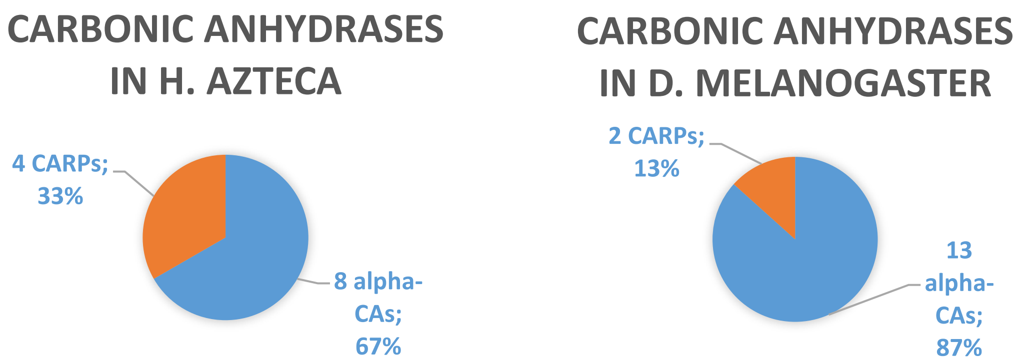



**Figure S4.10.1**. Relative distribution of CARPs and catalytic CAs in *H. azteca* and *D. melanogaster*

## S4.10. References:

(1) Hediger, M. A.; Clémençon, B.; Burrier, R. E.; Bruford, E. A. The ABCs of membrane transporters in health and disease (SLC series): introduction. *Molecular Aspects of Medicine* **2013,** *34*, (2), 95-107.
(2) Wieczorek, H.; Beyenbach, K. W.; Huss, M.; Vitavska, O. Vacuolar-type proton pumps in insect epithelia. *Journal of Experimental Biology* **2009,** *212*, (11), 1611-1619.
(3) Lee, C. E.; Kiergaard, M.; Gelembiuk, G. W.; Eads, B. D.; Posavi, M. Pumping ions: rapid parallel evolution of ionic regulation following habitat invasions. *Evolution* **2011,** *65*, (8), 2229-2244.
(4) McNamara, J. C.; Faria, S. C. Evolution of osmoregulatory patterns and gill ion transport mechanisms in the decapod Crustacea: a review. *Journal of Comparative Physiology B* **2012,** *182*, (8), 997-1014.
(5) Lee, C. E. Evolutionary mechanisms of habitat invasions, using the copepod Eurytemora affinis as a model system. *Evolutionary Applications* **2016,** *9*, (1), 248-270.
(6) Stephenson, M.; Mackie, G. Lake acidification as a limiting factor in the distribution of the freshwater amphipod Hyalella azteca. *Canadian Journal of Fisheries and Aquatic Sciences* **1986,** *43*, (2), 288-292.
(7) Smith, D. G. *Pennak's freshwater invertebrates of the United States: Porifera to Crustacea*. John Wiley & Sons: 2001.
(8) Gonzalez, E. R.; Watling, L. Redescription of *Hyalella azteca* from its type locality, Vera Cruz, Mexico (Amphipoda: Hyalellidae). *Journal of Crustacean Biology* **2002,** *22*, (1), 173-183.
(9) U.S. EPA. *Methods for measuring the toxicity and bioaccumulation of sediment-associated contaminants with freshwater invertebrates*; US Environmental Protection Agency: 2000.
(10) Lasier, P.; Winger, P.; Reinert, R. Toxicity of alkalinity to Hyalella azteca. *Bulletin of Environmental Contamination and Toxicology* **1997,** *59*, (5), 807-814.
(11) Wieczorek, H.; Putzenlechner, M.; Zeiske, W.; Klein, U. A vacuolar-type proton pump energizes K+/H+ antiport in an animal plasma membrane. *Journal of Biological Chemistry* **1991,** *266*, (23), 15340-15347.
(12) Wieczorek, H.; Brown, D.; Grinstein, S.; Ehrenfeld, J.; Harvey, W. R. Animal plasma membrane energization by proton-motive V-ATPases. *BioEssays* **1999,** *21*, (8), 637-648.
(13) Day, J. P.; Wan, S.; Allan, A. K.; Kean, L.; Davies, S. A.; Gray, J. V.; Dow, J. A. Identification of two partners from the bacterial Kef exchanger family for the apical plasma membrane V-ATPase of Metazoa. *Journal of Cell Science* **2008,** *121*, (15), 2612-2619.
(14) Harvey, W. R. Voltage coupling of primary H+ V-ATPases to secondary Na+-or K+-dependent transporters. *Journal of Experimental Biology* **2009,** *212*, (11), 1620-1629.
(15) Kondapalli, K. C.; Kallay, L. M.; Muszelik, M.; Rao, R. Unconventional chemiosmotic coupling of NHA2, a mammalian Na+/H+ antiporter, to a plasma membrane H+ gradient. *Journal of Biological Chemistry* **2012,** *287*, (43), 36239-36250.
(16) Xiang, M. A.; Linser, P. J.; Price, D. A.; Harvey, W. R. Localization of two Na+-or K+-H+ antiporters, AgNHA1 and AgNHA2, in Anopheles gambiae larval Malpighian tubules and the functional expression of AgNHA2 in yeast. *Journal of Insect Physiology* **2012,** *58*, (4), 570-579.
(17) Henry, R. P. The role of carbonic anhydrase in blood ion and acid-base regulation. *American Zoologist* **1984,** *24*, (1), 241-251.
(18) Beyenbach, K. W.; Wieczorek, H. The V-type H+ ATPase: molecular structure and function, physiological roles and regulation. *Journal of Experimental Biology* **2006,** *209*, (4), 577-589.
(19) Maxson, M. E.; Grinstein, S. The vacuolar-type H+-ATPase at a glance–more than a proton pump. In The Company of Biologists Ltd: 2014.

(20) Allan, A. K.; Du, J.; Davies, S. A.; Dow, J. A. Genome-wide survey of V-ATPase genes in Drosophila reveals a conserved renal phenotype for lethal alleles. *Physiological Genomics* **2005,** *22*, (2), 128-138.
(21) Brett, C. L.; Donowitz, M.; Rao, R. Evolutionary origins of eukaryotic sodium/proton exchangers. *American Journal of Physiology-Cell Physiology* **2005,** *288*, (2), C223-C239.
(22) Rheault, M. R.; Okech, B. A.; Keen, S. B.; Miller, M. M.; Meleshkevitch, E. A.; Linser, P. J.; Boudko, D. Y.; Harvey, W. R. Molecular cloning, phylogeny and localization of AgNHA1: the first Na+/H+ antiporter (NHA) from a metazoan, Anopheles gambiae. *Journal of Experimental Biology* **2007,** *210*, (21), 3848-3861.
(23) Donowitz, M.; Tse, C. M.; Fuster, D. SLC9/NHE gene family, a plasma membrane and organellar family of Na+/H+ exchangers. *Molecular Aspects of Medicine* **2013,** *34*, (2), 236-251.
(24) Chintapalli, V. R.; Kato, A.; Henderson, L.; Hirata, T.; Woods, D. J.; Overend, G.; Davies, S. A.; Romero, M. F.; Dow, J. A. Transport proteins NHA1 and NHA2 are essential for survival, but have distinct transport modalities. *Proceedings of the National Academy of Sciences* **2015,** *112*, (37), 11720-11725.
(25) Battaglino, R. A.; Pham, L.; Morse, L. R.; Vokes, M.; Sharma, A.; Odgren, P. R.; Yang, M.; Sasaki, H.; Stashenko, P. NHA-oc/NHA2: a mitochondrial cation–proton antiporter selectively expressed in osteoclasts. *Bone* **2008,** *42*, (1), 180-192.
(26) Syrjänen, L.; Tolvanen, M. E.; Hilvo, M.; Vullo, D.; Carta, F.; Supuran, C. T.; Parkkila, S. Characterization, bioinformatic analysis and dithiocarbamate inhibition studies of two new α-carbonic anhydrases, CAH1 and CAH2, from the fruit fly Drosophila melanogaster. *Bioorganic & Medicinal Chemistry* **2013,** *21*, (6), 1516-1521.
(27) Wessing, A.; Zierold, K.; Bertram, G. Carbonic anhydrase supports electrolyte transport in Drosophila Malpighian tubules. Evidence by X-ray microanalysis of cryosections. *Journal of Insect Physiology* **1997,** *43*, (1), 17-28.
(28) Shanbhag, S.; Tripathi, S. Epithelial ultrastructure and cellular mechanisms of acid and base transport in the Drosophila midgut. *Journal of Experimental Biology* **2009,** *212*, (11), 1731-1744.
(29) Ortutay, C. An evolutionary analysis of insect carbonic anhydrases. In *Advances in Medicine and Biology*, Bernhardt, L. E., Ed. Nova Science Publishers, Inc: Hauppauge, NY, 2010; Vol. 7, pp 145-168.
(30) Marchler-Bauer, A.; Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Research* **2004,** *32*, (suppl_2), W327-W331.
(31) Floden, E. W.; Tommaso, P. D.; Chatzou, M.; Magis, C.; Notredame, C.; Chang, J.-M. PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. *Nucleic Acids Research* **2016,** *44*, (W1), W339-W343.
(32) Löytynoja, A. Phylogeny-aware alignment with PRANK. *Multiple sequence alignment methods* **2014**, 155-170.
(33) Okonechnikov, Y. G., O.; Fursov, M.; UGENE team Unipro UGENE: a unified bioinformatics toolkit/ K. Okonechnikov, O. Golosova, M. Fursov. *Bioinformatics* **2012,** *28*, (8), 1166-1167.
(34) Sonnhammer, E. L.; Von Heijne, G.; Krogh, A. In *A hidden Markov model for predicting transmembrane helices in protein sequences*, Ismb, 1998; 1998; pp 175-182.
(35) Krogh, A.; Larsson, B.; Von Heijne, G.; Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* **2001,** *305*, (3), 567-580.
(36) Emameh, R. Z.; Barker, H.; Tolvanen, M. E.; Ortutay, C.; Parkkila, S. Bioinformatic analysis of beta carbonic anhydrase sequences from protozoans and metazoans. *Parasites & Vectors* **2014,** *7*, (1), 38.

(37) Jackson, B.; Lasier, P.; Miller, W.; Winger, P. Effects of calcium, magnesium, and sodium on alleviating cadmium toxicity to *Hyalella azteca*. *Bulletin of Environmental Contamination and Toxicology* **2000,** *64*, (2), 279-286.
(38) Wang, Z.; Meador, J. P.; Leung, K. M. Metal toxicity to freshwater organisms as a function of pH: A meta-analysis. *Chemosphere* **2016,** *144*, 1544-1552.
(39) Scheibener, S.; Richardi, V.; Buchwalter, D. Comparative sodium transport patterns provide clues for understanding salinity and metal responses in aquatic insects. *Aquatic Toxicology* **2016,** *171*, 20-29.
(40) de Schamphelaere, K. A.; Janssen, C. R. A biotic ligand model predicting acute copper toxicity for *Daphnia magna*: the effects of calcium, magnesium, sodium, potassium, and pH. *Environmental Science & Technology* **2002,** *36*, (1), 48-54.
(41) Tan, Q. G.; Wang, W. X. The influences of ambient and body calcium on cadmium and zinc accumulation in *Daphnia magna*. *Environ Toxicol Chem* **2008,** *27*, (7), 1605-1613.
(42) Soucek, D. J.; Mount, D. R.; Dickinson, A.; Hockett, J. R.; McEwen, A. R. Contrasting effects of chloride on growth, reproduction, and toxicant sensitivity in two genetically distinct strains of *Hyalella azteca*. *Environ Toxicol Chem* **2015,** *34*, (10), 2354-2362.
(43) Deanovic, L. A.; Markiewicz, D.; Stillway, M.; Fong, S.; Werner, I. Comparing the effectiveness of chronic water column tests with the crustaceans *Hyalella azteca* (order: Amphipoda) and *Ceriodaphnia dubia* (order: Cladocera) in detecting toxicity of current-use insecticides. *Environ Toxicol Chem* **2013,** *32*, (3), 707-712.

## S4.11. Metallothioneins

Helen C. Poynton
School for the Environment, University of Massachusetts Boston, Boston MA

Correspondence to helen.poynton@umb.edu

### Introduction

Metallothioneins (MTs) are a group of conserved metalloproteins with a high capacity for binding metal ions. These proteins are characterized by their low molecular weight (< 10 KDa), cysteine rich composition (often over 30%), lack of secondary structure in the absence of bound metal ions, and a two domain structure dictated by the bound ions. Despite their ubiquitous presence in animals, plants, fungi, and many prokaryotes, they vary considerably in their amino acid composition.[1, 2]

Metallothioneins were originally classified into three classes depending on the similarity to mammalian MTs.[3] Class I and II were expanded into fifteen families by Binz and Kagi (reviewed by Capdevila et al.[1]) based on phylogeny.

Assigning a specific biological function to the entire class of MTs has remained a challenge. However, MTs do share specific molecular functions including metal binding and redox activity. Capdevila et al.[1] argue that these molecular functions have allowed MTs to develop diverse biological functions in different organisms with likely roles in metal homeostasis, particularly of two important physiological metals, zinc and copper. The ability to bind metal ions; however, has also provided MTs with a role in detoxification, binding, and sequestration of toxic metals, particularly cadmium.[4] In addition to metal binding, transcript and protein levels are often highly induced by heavy metal exposure. This role is particularly relevant to ecotoxiocology where MTs have become indicators of heavy metal contamination. Given the importance of *Hyalella azteca* as sediment toxicity testing organism, identification of Metallothionein (*mt*) genes within the *H. azteca* genome would provide an opportunity to monitor metal contamination using *H. azteca* MTs.

### Methods

Due to the expectation that *mt* genes would be induced by Cd exposure, we explored the Cd gene expression data set (see supplementary methods; S1) for potential *mt* candidate genes. All 579 contig sequences which responded to Cd treatment were searched for homology to the blue crab, *Callinectes sapidus,* CdMT-I (AAF08964) using tbastn. The contigs with an E-value > 0.1 were sorted by level of induction and then mapped to the *H. azteca* genome. One contig, contig02661 aligned to scaffold 460 in three places with a partial alignment in another location. RNAseq mapped reads were compared with the alignment of contig02261 to determine the transcribed region. Protein sequences were obtained from the transcribed area and compared to other decapod MTs to determine possible sequence similarity.

To find the *mt* genes within the redundans assembly (assembly 2.0), the cDNA sequence of *mt* A was queried against Hazt_redundans_assembly_test_3_4_16.fa using blast2.3.0+ locally. A significant blast hit was found to several locations on scaffold 396 with e-values ranging from 0.0 to 0.001. cDNA sequences of the four Mtn genes were identified using the edit->find function in MS Word.

**Results and Discussion**

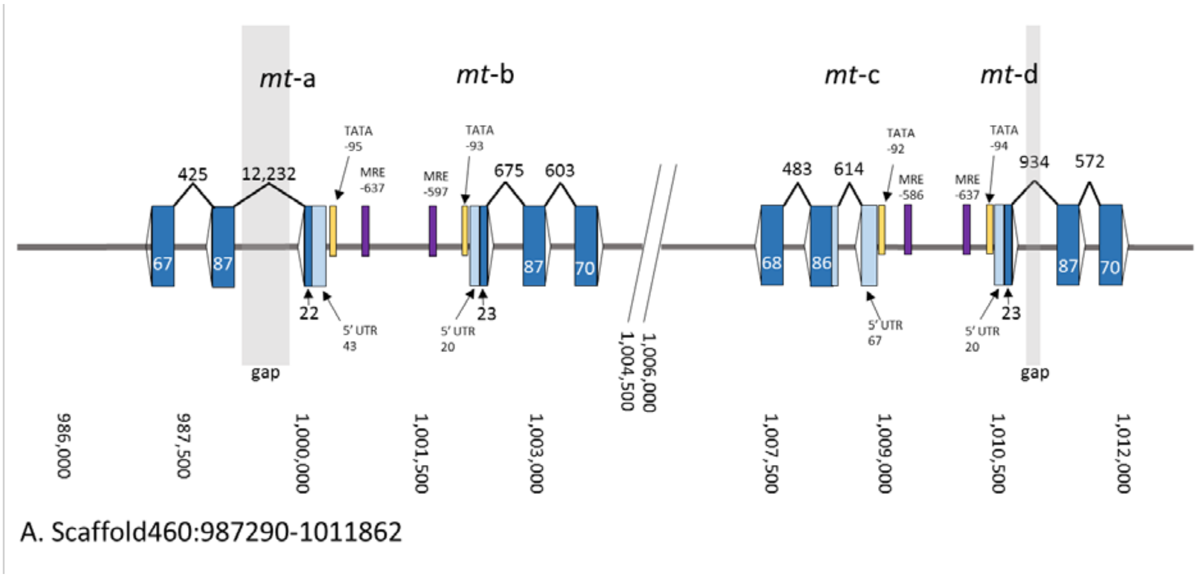
Four MT genes were identified by mapping Cd responsive contigs with homology to the *C. sapidus* CdMT-1 to the *H. azteca* genome. These four genes are arranged as repeats on scaffold 460 and each contains three exons, the typical gene structure of *mts* (see Figure S4.11.1). *Mt*-a, *mt*-b, and *mt*-d produce proteins with identical amino acid sequences, while *mt*-c is missing the downstream splice site on exon 1 and produces a truncated protein of 53 amino acids. Due to the similarity in the sequences of these four genes, it is not possible to determine if all four genes are actually transcribed or regulated differently based solely on the RNAseq mapped reads.

In the original all-paths assembly, there were gaps present within the introns of *mt* -a and *mt* -d producing long introns (Figure S4.11.1A). To determine if these long introns were an artifact of the assembly, we compared the genomic sequence of the *mt* gene cluster to the redundans assembly. The redundans assembly did remove the long gaps with these introns; however, it also revealed that a portion of *mt* A was missing including the start codon in exon 1. Unlike *mt* -c, there is no alternate start codon available, and it was concluded that this is likely a pseudogene.

To confirm that these three putative genes were *mt* genes, I compared the MT protein sequences to other MT proteins identified in Malacostraca. Within this class, MTs have been identified in several decapods, including the blue crab *C. sapidus*, the lobster *Homarus* *americanus*, and the crayfish *Pacifastacus leniusculus.* The *H. azteca* MTs showed homology and conserved placement of cysteine residues that matched closely with *H. americanus*, *P. leniusculus*, and the CdMT-I protein (Figure S4.11.2). In addition, with the exception of putative MT-c, these proteins contain 61 amino acids and 18 cysteine residues which is consistent with the pattern in other crustacean MTs described to date.

To determine if *H. azteca* MT is responsive to heavy metal exposure, I compared the expression level of contig02661 in the Cd gene expression data set and Zn gene expression data set. I found that *H. azteca* MT is highly induced by Cd exposure and is also induced about 2-fold from low levels of Zn exposure (Figure S4.11.3).

Three lines of evidence provide strong support that *mt* -b, and *mt* -d are genes which encode for the *H. azteca* MT protein. First, the gene structure with three exons and an upstream MRE is consistent with other crustacean MTs.[5] Second, the number of total amino acids, number of cysteine residues, and the alignment of *H. azteca* MT-b with other Malacostraca MTs provides very strong support that these newly identified genes are MTs. Finally, the gene expression pattern is also very characteristic of MTs with high induction of gene expression from Cd exposure and induction from Zn exposure.



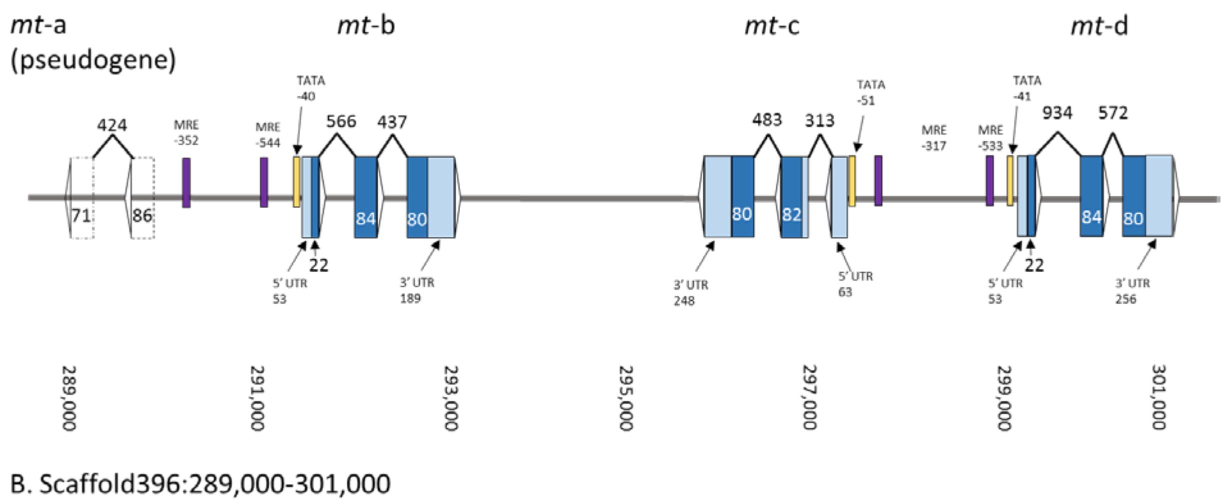




**Figure S4.11.1**: Gene models for duplicated *mt* genes within the *H. azteca* genome (A) all-paths genome assembly, (B) redundans genome assembly. Dark blue regions represent translated regions of the gene, while light blue regions represent 5' and 3' untranslated regions. Each *mt* paralog contained an upstream metal response element (MRE) shown in purple and TATA box shown in yellow. In the redundans assembly (B), all gaps are removed as well as a portion of *mt* A, leaving it without a start codon.

```
H. azteca                   MPKPCCQESCPCPEEVCSSKCTDCKDCDCRSKCDKCKESCGCATKEACAGNCTSPCSCCPK
C. sapidus       AAF08964   MPGPCCNDKCVCQEGGCKAGCQ-CTSCRCSP-CQKCTSGCKCATKEECSKTCTKPCSCCPK
H. americanus    CAC80859   MPGPCCKDKCECAEGGCKTGCK-CTSCRCAP-CEKCTSGCKCPSKDECAKTCSKPCKCCP
P. leniusculus   AAF07215   MPGPCCNDQCECAAGGCKTGCV-CTSCRCQP-CDKCTSGCKCPSKEECAKTCSKPCRCCP

conserved                   MP-PCC---C-C----C---C--C--C-C---C--C---C-C-CK--C---C--PC-CCP-
```

**Figure S4.11.2**: Amino acid alignment of *H. azteca* MT amino acid sequences with other Malacostraca MT proteins. Cysteine residues are shown in bold with conserved residues shaded in grey. The conserved sequence between all four species reveals the importance in the conservation of the position of the cysteine residues.

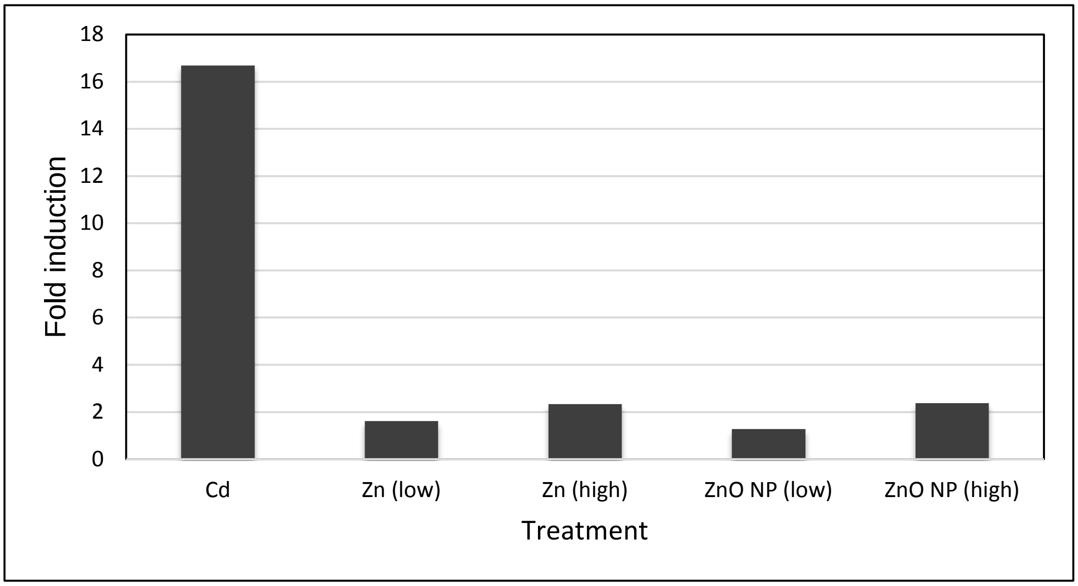



**Figure S4.11.3**: Gene expression pattern of contig02661 (representing the transcript for MT-a, MT-b, and MT-c) after 96-h exposure to cadmium, zinc, or zinc oxide nanoparticles. Bars represent the fold induction of MT in each treatment compared to a non-exposed control.  See supplementary methods (S1) for details on exposure concentrations and duration.

**S4.11. References:**

(1) Capdevila, M.; Bofill, R.; Palacios, O.; Atrian, S. State-of-the-art of metallothioneins at the beginning of the 21st century. *Coordination Chemistry Reviews* **2012,** *256*, (1), 46-62.
(2) Isani, G.; Carpenè, E. Metallothioneins, unconventional proteins from unconventional animals: a long journey from nematodes to mammals. *Biomolecules* **2014,** *4*, (2), 435-457.
(3) Fowler, B.; Hildebrand, C.; Kojima, Y.; Webb, M. Nomenclature of metallothionein. In *Metallothionein II*, Springer: 1987; pp 19-22.
(4) Amiard, J.-C.; Amiard-Triquet, C.; Barka, S.; Pellerin, J.; Rainbow, P. Metallothioneins in aquatic invertebrates: their role in metal detoxification and their use as biomarkers. *Aquatic Toxicology* **2006,** *76*, (2), 160-202.
(5) Shaw, J. R.; Colbourne, J. K.; Davey, J. C.; Glaholt, S. P.; Hampton, T. H.; Chen, C. Y.; Folt, C. L.; Hamilton, J. W. Gene response profiles for *Daphnia pulex* exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. *BMC Genomics* **2007,** *8*, (1), 477.

# S4.12. Nuclear Receptor Gene Family

Faith Lambert and Chris Vulpe
University of Florida
Center for Environmental and Human Toxicology, College of Veterinary Medicine, University of Florida, Gainesville, FL, USA

Correspondence to: fnlambert@ufl.edu.

Peter Bain
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Urrbrae, Australia.

Correspondence to peter.bain.0@gmail.com

## Introduction

Nuclear receptors (NRs) are transcription factors that regulate gene transcription in response to external and internal stimuli, causing downstream changes in animal physiology. Many NRs are activated by ligand binding, but a subset of these receptors (the orphan receptors) exists for which ligands are unknown. NRs are present throughout the animal kingdom as highly conserved and very successful signal mediators in animal physiology.[1] These receptors play an important role in controlling processes such as reproduction, growth, development, and behavior.[2, 3]

NRs are characterized by a five domain structure, which includes an N-terminal domain, a DNA binding domain (DBD), a hinge region, a ligand binding domain (LBD), and a C-terminal domain.[4] The N-terminal section contains transactivation domains which are necessary for inducing transcription. The DBD is a DNA sequence specific binding region that recognizes response elements in the DNA. It is the most highly conserved region between species and subclasses of nuclear receptors. The hinge region is variable among species and connects the DBD to the LBD. The LBD is also highly conserved between species. This domain is responsible for recognizing internal and external signals which activate the transcriptional response.[4] The nuclear receptors are classified into seven subfamilies (NR0-NR6). The groupings are based on their natural ligands and sequence similarity, particularly within the DBD and LBD domains.[5]

Due to their importance in physiology, the NRs have become of prominent interest in the field of ecotoxicology. Pollutants and toxicants in the environment, such as endocrine disrupting compounds, can potentially activate these receptors, causing physiological disruptions in humans and wildlife.[6]

Crustaceans are important components of aquatic ecosystems, and several model species exist for use in ecotoxicology studies.[7, 8] However, there is currently limited information on the sequences of many crustacean NRs. For this reason, the NRs of the recently sequenced *Hyalella azteca* genome, a commonly used model organism in ecotoxicology, were annotated.

**Methods**

*Daphnia pulex* is the species most closely related to *H. azteca* with a sequenced and annotated genome. Therefore, protein sequences of *D. pulex* nuclear receptors were used as a basis for identifying NR genes in the *H. azteca* genome. Additionally, *Drosophilia melanogaster* sequences were used to identify NRs in the *H. azteca* genome. *D. pulex* sequences were gathered from the JGI Genome Portal (http://genome.jgi.doe.gov/pages/search-for-genes.jsf?organism=Dappu1) and Fleabase (http://wfleabase.org/).[9] *D. melanogaster* sequences were obtained from Flybase (http://flybase.org/).[10] These sequences were used in tblastn searches against the assembled *H. azteca* genome. Areas of the *H. azteca* contigs that were highly homologous to *D. pulex* or *D. melanogaster* NR sequences were investigated further. NR genes were identified by the presence of either a highly conserved DBD, a highly conserved LBD, or both. Contig regions containing one or both of these conserved domains were investigated further. When a gene model (HAZTv0.5.3-models and/or augustusmasked) was present that aligned to the query used in tblastn, the model was used as a base for gene annotation. Gene models were altered when evidence from RNAseq mapped reads and homology to sequences of other species supported the change. If no model was present on the contig in a region of homology, a *de novo* model was generated using information either the blastx-Arthropoda programme or by using the homologous sequence as a guide. Further alteration of these *de novo* models was guided by resolving discrepancies between orthologs in other species and referencing RNAseq data. Coding sequences from the completed annotations were used in a reciprocal blast (blastx NCBI) to confirm orthologs presence in *H. azteca.*

For the estrogen-related receptor (ERR), a draft gene model was identified as a likely ERR candidate. This gene model, HaztTmpM001378-RA, was edited to remove a small exon (exon 4 in the draft model) which had no read support, and to extend the 3' UTR based on strong read support. Exon 3 was also adjusted to better match read support and the resulting open reading frame encoded a protein that aligned well with other crustacean ERRs. Amino acid alignment of arthropod estrogen-related receptors (ERRs) was prepared with the aid of Geneious Pro (www.geneious.com) using the MAFFT E-INS-i algorithm.[11]

For Ecdysone receptor (EcR), a gene model encoding a putative EcR was identified in the *H. azteca* draft genome assembly by similarity searches using other crustacean EcR nucleotide sequences as queries. The name of the original model in the draft annotation was HaztTmpM005197-RA. Additional 5' exons were added based on strong support from RNA-Seq coverage and junction reads, high-scoring blast hits from EcR sequences from other crustaceans, and the presence of a longer open reading frame in the revised model that bore high similarity to other crustacean EcRs. Exon 3 of the original model was adjusted based on read coverage and the presence of junction reads for the revised position of the exon.

**Results**

*Overview*

A total of 17 whole or partial NR genes were annotated in the *H. azteca* genome. Among these genes were representatives from each of the seven major NR groups. Due to gaps in the assembly, several NR genes had sequences that were only partially elucidated (E78, HR3, RXR, and SVP).

Table S4.12.1 outlines the NR genes present in *H. azteca* as compared to those in *D. pulex and D. melanogaster*. The number of total NR genes in *H. azteca* is lower than both the closely related *D. pulex* (25 total NRs) and more distantly related *D. melanogaster* (21 total NRs).[12] This may be a factor of the low anticipated total number of genes in the *H. azteca* genome. Another possible explanation is the presence of gaps in the assembly, which may have obscured portions of genes or whole genes entirely. Transcript variants, of which few were identified, could have also likely been masked by these assembly gaps.

The diversity of receptors in each group was lower in *H. azteca* when compared to other species. The 0A (KNR-R1/R2), 1H (EcRa/b), 1L (HR97a/b/g), and 2E (PNR/TLL/DSF) groups of receptors in *H. azteca* contained fewer receptor subtypes than the same groups in *D. pulex* and *D. melanogaster*. Furthermore, certain subgroups in *H. azteca* did not have any representatives (1M, 1N, and 5B), contrary to these same groups in *D. pulex*, which each contain one NR subtype. It is thought that *D. pulex* underwent a genome duplication during its evolution, resulting in the larger variety of NR subtypes.[12] Evidence from the *H. azteca* genome does not indicate a similar situation. Nonetheless, it is difficult to say whether these are absent from the genome or if their presence was concealed by other factors (e.g. gaps in the assembly).

Several genes which were recently demonstrated to exist in crustaceans for the first time in *D. pulex*[12] and subsequently in *Daphnia magna*[13] were also found in *H. azteca*. These include members of the NR1 group (E78, HR96), NR2 group (HR78, HNF4), and NR6 group (HR4). This information from *H. azteca* supports the likelihood that these genes are common in other crustaceans as well.

Furthermore, a gene encoding an NR from group that is novel to the crustaceans, the 1L group, was identified in *H. azteca.* The 1L group was recently discovered in another crustacean, *D. pulex,* whose genome was previously sequenced.[12] Since then, HR97 isoforms have also been identified in another crustacean *D. magna*[13, 14] and the copepod *Tigriopus japonicus.*[15] HR97 receptors are closely related, but distinct from the HR96 receptor. Although there is currently no known ligand for the HR97 receptors, they are thought to play a role in developmental stage changes.

Overall, the distribution of subtypes among the different NR groups was very similar to those reported in other crustaceans. Information from this annotation supports evidence for NR groups that have recently been identified in crustacean species.

*Estrogen-related receptor*

Estrogen-related receptors (ERRs) are nuclear receptors (NRs) of the NR3 subfamily (estrogen receptor-like), and are classified as NR3B. ERRs are the only members of the NR3 subfamily known to exist in arthropods and have been identified in numerous crustaceans including *Daphnia pulex,*[12] *Daphnia magna,*[13] the copepods *Calanus finmarchicus,*[16] and *Tigriopus japonicas*,[15] and the freshwater shrimp *Macrobrachium nipponense*.[17]

The modified *H. azteca* gene model for ERR lacks exons encoding the amino-terminal domain present in canonical NRs (i.e. A/B domain) which is involved transcriptional coactivator interaction (Figure S4.12.1). The DNA-binding domain (C domain), hinge region (D domain) and ligand-binding domain (E domain) are present, suggesting that ligand-dependent binding with genomic DNA response elements is likely to be possible, but the lack of an A/B domain may prevent recruitment of the transcriptional machinery by this receptor. A carboxyl-terminal 'F' domain present in some NRs is not present in arthropod ERRs. Whether the lack of upstream sequence coding for the A/B domain is due to incomplete assembly, incomplete annotation or evolutionary loss of the domain is unclear at this stage. While there are no assembly gaps in the region, manual addition of exons upstream of the existing gene model (based on read coverage) did not extend the predicted open reading frame. An alignment with the top 10 BLAST hits from NCBI nr along with *Daphnia* ERRs is provided (Figure S4.12.2).

The redundans assembly (Hazt_2.0) contained no additional information and the most similar scaffold was identical to the existing assembly in this region. Predicting genes in this region using web augustus (http://bioinf.uni-greifswald.de/webaugustus/prediction/create, using *Drosophila*-trained parameters, which may not be ideal) did reveal two possible additional upstream exons, but one of them had no read support and the other contained no start codon and hence did not extend the reading frame; and neither had junction read support. However, this analysis suggests that further refinement of the model may reveal additional coding sequence corresponding to the A/B domain.

*Ecdysone receptor*

Ecdysone is the major steroid hormone in arthropods and is the precursor of the moulting hormone 20-hydroxyecdysone. The Ecdysone receptor (EcR), also known as ecdysteroid receptor, was first identified in *Drosophila* by screening for members of the nuclear receptor (NR) family capable of binding ecdysone analogues with high affinity.[18] EcR alone does not bind strongly to target response elements, even in the presence of ligand, but rather dimerises with the retinoid X receptor (RXR) homologue known as ultraspiracle protein (USP) to form the functional ecdysone receptor.[19]

A gene model encoding a putative EcR was identified in the *H. azteca* draft genome assembly by similarity searches using other crustacean EcR nucleotide sequences as queries. Although the length of the encoded protein is similar to one of the earliest cloned crustacean EcR sequences, from the fiddler crab *Uca pugilator*,[20] the lack of an intact A/B domain (transactivation domain) suggests that the *H. azteca* EcR gene may be incomplete (like the published *U. pugilator* sequence). Other reported crustacean EcRs generally include the A/B domain. There is a gap in the *H. azteca* draft assembly upstream from the putative EcR gene and no evidence of any additional read alignments supporting the existence of exons coding for a possible A/B domain.

Among crustaceans, the putative *H. azteca* EcR shares a maximum of 50% overall amino acid identity (with the range for comparison limited to the length of the alignment) with *Uca pugilator* (fiddler crab) ecdysteroid receptor, isoform 4 (AIE88264). Maximum identity within the DNA binding domain (C domain, 89 residues) is 85.4% with both *Homarus americanus* (American lobster) and *Crangon crangon* (brown shrimp) EcR. Maximum identity within the ligand binding domain (E domain, 238 residues) is 54.6% with *Uca pugilator* EcR. A neighbor-joining phylogenetic analysis shows that the *H. azteca* is most closely related to EcR sequences from *Daphnia magna* and *Sogatella furcifera* but is relatively divergent within the clade (Figure S4.12.3).

| Group | Scaffold | *H.azteca* | *D. pulex* | *D. melanogaster* |
|---|---|---|---|---|
| 0A | | KNR | Dappu-KNR-R1<br>Dappu-KNR-R2 | KNI<br>KNRL<br>EGON |
| 0B | | | | |
| 1A | | | | |
| 1B | | | | |
| 1C | | | | |
| 1D | | E75 | Dappu-E75 | E75 |
| 1E | | E78 * | Dappu-E78 | E78 |
| 1F | | HR3 * | Dappu-HR3 | DHR3 |
| 1G | | | | |
| 1H | | EcR | Dappu-EcRa<br>Dappu-EcRb | EcR |
| 1J | | HR96 | Dappu-HR96 | DHR96 |
| 1K | | | | |
| 1L | | HR97 | Dappu-HR97a<br>Dappu-HR97b<br>Dappu-HR97g | |
| 1M | | | Dappu-HR10 | |
| 1N | | | Dappu-HR11 | |
| 2A | | HNF4 | Dappu-HNF4 | HNF4 |
| 2B | | RXR * | Dappu-RXR | USP |
| 2C | | | | |
| 2D | | HR78 | Dappu-HR78 | DHR78 |
| 2E | | TLL<br>PNR | Dappu-TLL<br>Dappu-PNR<br>Dappu-DSF | TLL<br>PNR<br>DSF<br>FAX-I |
| 2F | | SVP * | Dappu-SVP | SVP |
| 3A | | | | |
| 3B | | ERR | Dappu-ERR | ERR |
| 3C | | | | |
| 4A | | HR38 | Dappu-HR38 | DHR38 |
| 5A | | FTZ-F1 | Dappu-FTZ-F1 | FTZ-F1 |
| 5B | | | Dappu-HR-39 | DHR39 |
| 6B | | HR4 | Dappu-HR4 | DHR4 |
| **Total** | | **17** | **25** | **21** |

**Table S4.12.1** Nuclear receptors in *H. azteca* and two closely related species, *D. pulex and D. melanogaster.* *Only partial sequences were available for these genes.

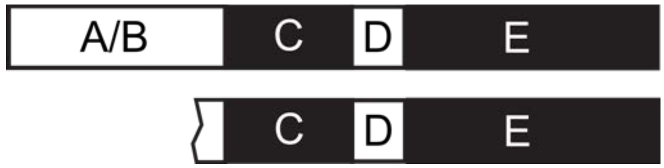
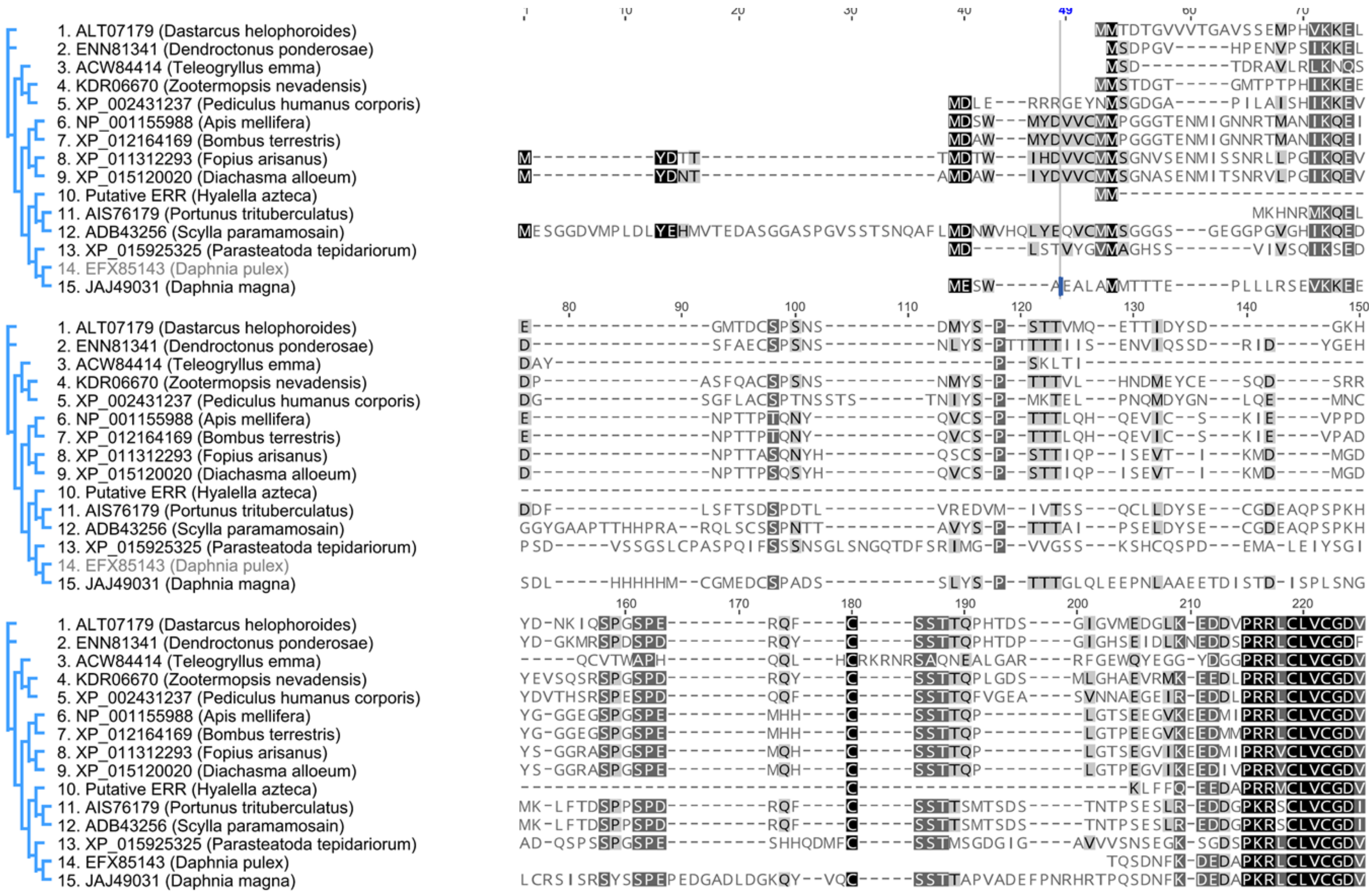
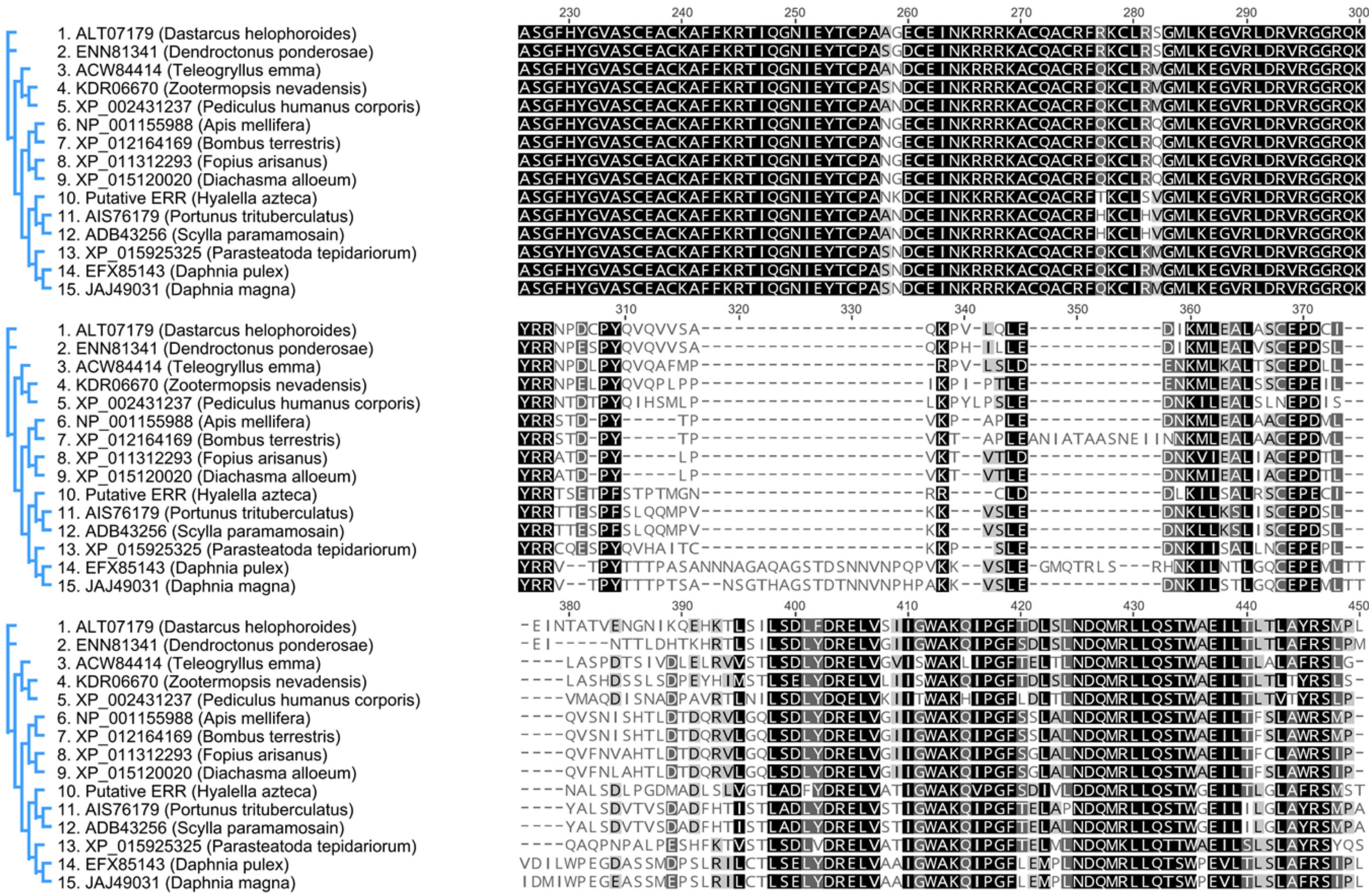



**Figure S4.12.1**. Domain structure of canonical nuclear receptors (top) compared with the domain structure of *Hyalella azteca* ERR derived from the current gene model (bottom).


1. ALT07179 (Dastarcus helophoroides)
2. ENN81341 (Dendroctonus ponderosae)
3. ACW84414 (Teleogryllus emma)
4. KDR06670 (Zootermopsis nevadensis)
5. XP_002431237 (Pediculus humanus corporis)
6. NP_001155988 (Apis mellifera)
7. XP_012164169 (Bombus terrestris)
8. XP_011312293 (Fopius arisanus)
9. XP_015120020 (Diachasma alloeum)
10. Putative ERR (Hyalella azteca)
11. AIS76179 (Portunus trituberculatus)
12. ADB43256 (Scylla paramamosain)
13. XP_015925325 (Parasteatoda tepidariorum)
14. EFX85143 (Daphnia pulex)
15. JAJ49031 (Daphnia magna)


1. ALT07179 (Dastarcus helophoroides)
2. ENN81341 (Dendroctonus ponderosae)
3. ACW84414 (Teleogryllus emma)
4. KDR06670 (Zootermopsis nevadensis)
5. XP_002431237 (Pediculus humanus corporis)
6. NP_001155988 (Apis mellifera)
7. XP_012164169 (Bombus terrestris)
8. XP_011312293 (Fopius arisanus)
9. XP_015120020 (Diachasma alloeum)
10. Putative ERR (Hyalella azteca)
11. AIS76179 (Portunus trituberculatus)
12. ADB43256 (Scylla paramamosain)
13. XP_015925325 (Parasteatoda tepidariorum)
14. EFX85143 (Daphnia pulex)
15. JAJ49031 (Daphnia magna)
230
240
250
260
270
280
290
300
310
320
330
340
350
360
370
380
390
400
410
420
430
440
450

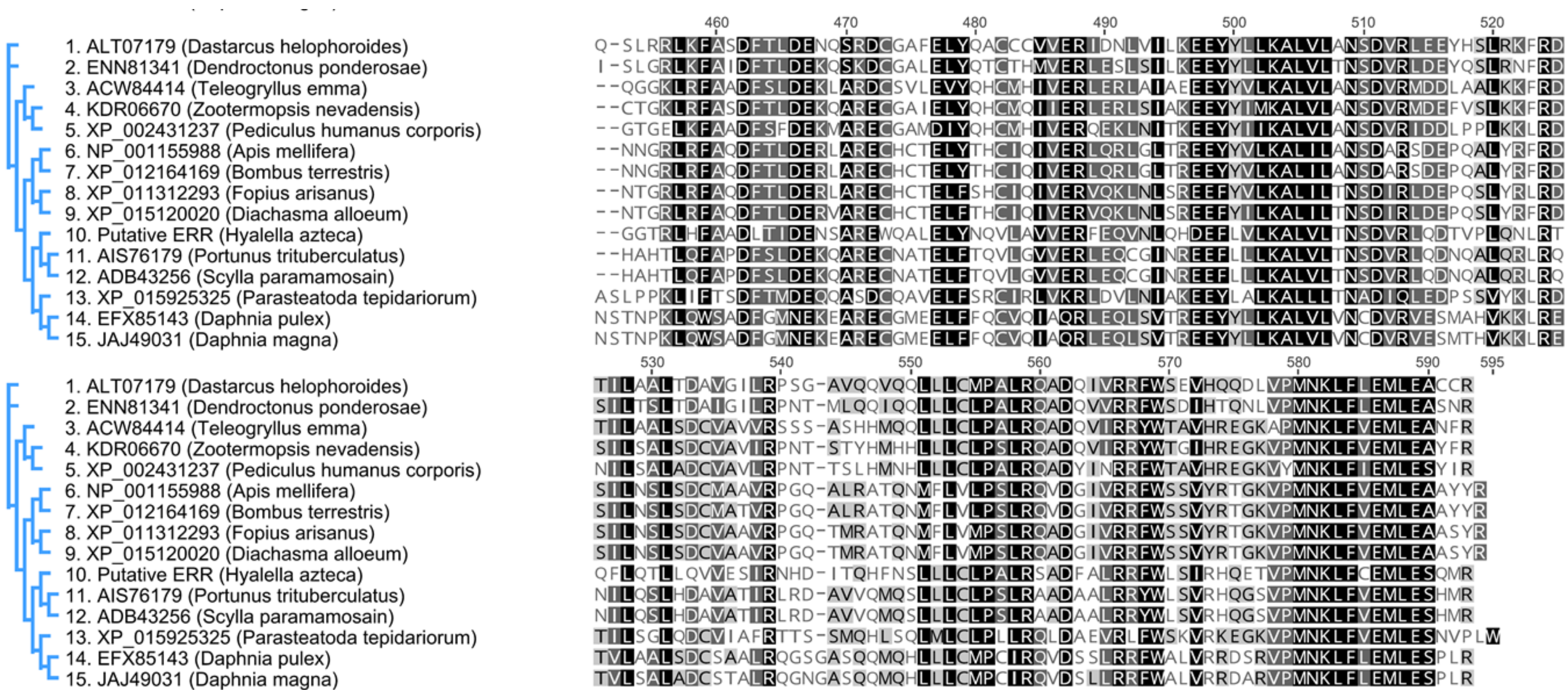


**Figure S4.12.2.** Amino acid alignment of arthropod estrogen-related receptors (ERRs) showing the amino-terminal truncation present in the *H. azteca* putative ERR deduced amino acid sequence. The alignment was prepared with the aid of Geneious Pro (www.geneious.com) using the MAFFT E-INS-i algorithm.[11]

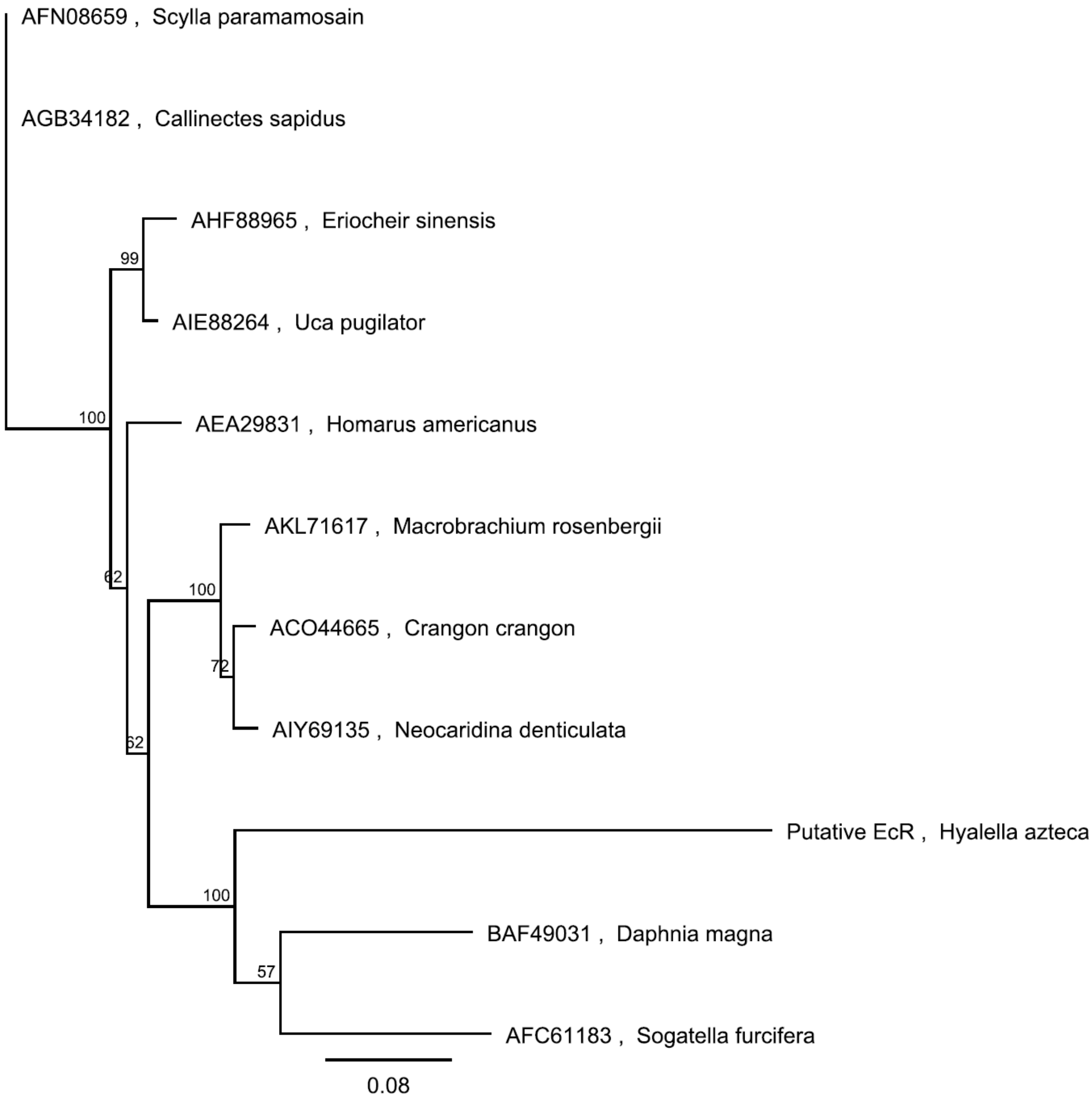



**Figure S4.12.3.** A consensus phylogenetic tree showing relationships among ecdysone receptors from selected arthropods. The tree was constructed with the aid of Geneious Pro (www.genious.com) using the neighbor-joining method based on Jukes-Cantor distances. Node labels represent bootstrap support values from 1000 replications (support values less than 50% not shown)

## S4.12. References


(1) Evans, R. M. The nuclear receptor superfamily: a rosetta stone for physiology. *Molecular Endocrinology* **2005,** *19*, (6), 1429-1438.

(2) Chawla, A.; Repa, J. J.; Evans, R. M.; Mangelsdorf, D. J. Nuclear receptors and lipid physiology: opening the X-files. *Science* **2001,** *294*, (5548), 1866-1870.

(3) Ogawa, S.; Eng, V.; Taylor, J.; Lubahn, D. B.; Korach, K. S.; Pfaff, D. W. Roles of Estrogen Receptor-α Gene Expression in Reproduction-Related Behaviors in Female Mice. *Endocrinology* **1998,** *139*, (12), 5070-5081.

(4) Kumar, R.; Thompson, E. B. The structure of the nuclear hormone receptors. *Steroids* **1999,** *64*, (5), 310-319.

(5) Nuclear Receptot Nomenclature Committee. A unified nomenclature system for the nuclear receptor superfamily. *Cell* **1999,** *97*, (2), 161-163.

(6) Colborn, T.; vom Saal, F. S.; Soto, A. M. Developmental effects of endocrine-disrupting chemicals in wildlife and humans. *Environmental Health Perspectives* **1993,** *101*, (5), 378.

(7) LeBlanc, G. A. Crustacean endocrine toxicology: a review. *Ecotoxicology* **2007,** *16*, (1), 61-81.

(8) Mazurová, E.; Hilscherová, K.; Triebskorn, R.; Köhler, H.-R.; Maršálek, B.; Bláha, L. Endocrine regulation of the reproduction in crustaceans: identification of potential targets for toxicants and environmental contaminants. *Biologia* **2008,** *63*, (2), 139-150.

(9) Colbourne, J. K.; Singan, V. R.; Gilbert, D. G. wFleaBase: the Daphnia genome database. *BMC Bioinformatics.* **2005,** *6*, 45.

(10) Attrill, H.; Falls, K.; Goodman, J. L.; Millburn, G. H.; Antonazzo, G.; Rey, A. J.; Marygold, S. J. FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. *Nucleic Acids Research.* **2016,** *44*, (D1), D786-792. .

(11) Katoh, K.; Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **2013,** *30*, (4), 772-780.

(12) Thomson, S.; Baldwin, W.; Wang, Y.; Kwon, G.; LeBlanc, G. Annotation, phylogenetics, and expression of the nuclear receptors in Daphnia pulex. *BMC Genomics* **2009,** *10*, (1), 500.

(13) Litoff, E. J.; Garriott, T. E.; Ginjupalli, G. K.; Butler, L.; Gay, C.; Scott, K.; Baldwin, W. S. Annotation of the *Daphnia magna* nuclear receptors: Comparison to *Daphnia pulex*. *Gene* **2014,** *552*, (1), 116-125.

(14) Li, Y.; Ginjupalli, G. K.; Baldwin, W. S. The HR97 (NR1L) group of nuclear receptors: a new group of nuclear receptors discovered in Daphnia species. *General and Comparative Endocrinology* **2014,** *206*, 30-42.

(15) Hwang, D.-S.; Lee, B.-Y.; Kim, H.-S.; Lee, M. C.; Kyung, D.-H.; Om, A.-S.; Rhee, J.-S.; Lee, J.-S. Genome-wide identification of nuclear receptor (NR) superfamily genes in the copepod Tigriopus japonicus. *BMC Genomics* **2014,** *15*, (1), 993.

(16) Tarrant, A. M.; Baumgartner, M. F.; Hansen, B. H.; Altin, D.; Nordtug, T.; Olsen, A. J. Transcriptional profiling of reproductive development, lipid storage and molting throughout the last juvenile stage of the marine copepod Calanus finmarchicus. *Frontiers in Zoology* **2014,** *11*, (1), 1-15.

(17) Jiang, H.; Li, X.; Sun, Y.; Hou, F.; Zhang, Y.; Li, F.; Gu, Z.; Liu, X. Insights into Sexual Precocity of Female Oriental River Prawn *Macrobrachium nipponense* through Transcriptome Analysis. *PLoS ONE* **2016,** *11*, (6), e0157173.

(18) Koelle, M. R.; Talbot, W. S.; Segraves, W. A.; Bender, M. T.; Cherbas, P.; Hogness, D. S. The drosophila EcR gene encodes an ecdysone receptor, a new member of the steroid receptor superfamily. *Cell* **1991,** *67*, (1), 59-77.

(19) Yao, T.-P.; Forman, B. M.; Jiang, Z.; Cherbas, L.; Chen, J. D.; McKeown, M.; Cherbas, P.; Evans, R. M. Functional ecdysone receptor is the product of EcR and Ultraspiracle genes. *Nature* **1993,** *366*, (6454), 476-479.
(20) Chung, A. C. K.; Durica, D. S.; Clifton, S. W.; Roe, B. A.; Hopkins, P. M. Cloning of crustacean ecdysteroid receptor and retinoid-X receptor gene homologs and elevation of retinoid-X receptor mRNA by retinoic acid. *Molecular and Cellular Endocrinology* **1998,** *139*, (1–2), 209-227.

## S4.13. Opsins

Markus Friedrich and Jeffery Jones
Wayne State University, Detroit MI, USA

Correspondence to friedrichwsu@gmail.com

### Introduction

Whether marine or terrestrial, arthropods deploy a diversity of light sensing mechanisms. Arguably the most predominant is light capture in photoreceptor cells through the expression of opsins: light-sensitive, seven-transmembrane G-protein coupled opsin receptor proteins. As is generally the case for arthropods, the visual behavior of crustaceans is driven by both retinal and non-retinal sensing of light. Retinal light perception takes place in two types of eyes: the single-chambered median eyes and lateral compound eyes. These main visual organs express retinal or visual opsins, which includes exclusively rhabdomeric opsins (r-opsins). In most cases, a single arthropod photoreceptor cell expresses one of several rhabdomeric opsin (r-opsin) genes, which differ in wavelength-sensitivity (WS). The nomenclature of these subfamilies reflects their diversified wavelength-specific maximal sensitivities. In panarthropods (Crustacean + Hexapoda), this includes the long wavelength-sensitive subfamily 2 (LWS2), the middle wavelength-sensitive subfamilies 1 and 2 (MWS1 and MWS2), and the short wavelength-sensitive subfamily 2 (SWS2).[1] Further duplications within these subfamilies potentially expanded this repertoire to a total of nine subfamilies in early pancrustaceans.[1]

Non-retinal light perception is mediated by the expression of specific subfamilies of opsin genes in other parts of the body. Arthropods are known to express separate opsin subfamilies in diverse non-photoreceptor cell types. This includes the Rh7,[2] arthropsin,[3 4 5] c-opsin,[6] and peropsin subfamilies.[7]

### Methods

The *H. azteca* genome draft Hazt_1.0 was searched by BLAST in Apollo environment with query sequences from all arthropod opsin subfamilies.[1] Best matches were tested by reciprocal BLAST to confirm opsin gene family homology. Candidate opsin gene family members were further investigated by gene tree reconstruction and analysis.

To generate global crustacean opsin gene trees, we deployed the multiple sequence alignment program MUSCLE.[8] Ambiguous sites removed with Gblock,[9] using the least stringent settings (Minimum Number Of Sequences For A Conserved Position: 64; Minimum Number Of Sequences For A Flanking Position: 64; Maximum Number Of Contiguous Nonconserved Positions: 8; Minimum Length Of A Block: 5; Allowed Gap Positions: With Half). This resulted in 255 sites of 126 sequences for gene tree reconstruction, which was carried out with the Neighbor Joining method in MEGA [10] using the Jones-Taylor-Thornton (JTT) model for pairwise distance estimated corrected for across site substitution variation applying a gamma distribution with four rate categories. Branch support was estimated by nonparametric bootstrapping from 100 sequence replicates.

To investigate the two *H. azteca* LWS opsin genes at higher resolution, we used web PRANK[11] for multiple sequence alignment and maximum likelihood for tree estimation. Ambiguous alignment sites were filtered with Gblocks as described above. Branch support was investigated by nonparametric bootstrap from 100 sequence sample replicates.

**Results and Discussion**

Considerable variation exists in the conservation and expression of opsin genes across species as well as between major arthropod clades, ranging from massive subfamily expansions to specific subfamily losses. The opsin repertoires of several malacostracan crustaceans have been characterized in detail, but comparatively little is known about the opsin repertoire of amphipods. The *H. azteca* genome thus adds a new important data point towards elucidating crustacean visual diversity. More specifically, *H. azteca* represents two firsts: 1) The first genome wide search for opsin genes for a family member of the Gammaridae; 2) The first extension of previous studies that explored opsin conservation in the freshwater *Gammarus minus*.[12] The latter work detected two closely related MWS opsins in *G. minus*, which likely predate the emergence of the species.

Our survey revealed the presence of three opsin genes in the *H. azteca* genome. A global analysis of crustacean opsin genes placed one of them into the MWS1 subfamily and two into the LWS subfamily (Fig. S4.13.1). The two LWS opsin paralogs are closely linked, separated by only about 6,000 bp (Table S4.13.1). Maximum likelihood analysis with a subset of closely related malacostracan LWS opsins moderately supported the two *H. azteca* LWS opsins as 1:1 orthologs of the two LWS opsins previously reported for *G. minus*.[12] In conclusion, the LWS opsin duplicate pair conserved in *H. azteca* and *G. minus* is likely ancient, predating at least the origin of amphipod Crustacea. The *H. azteca* MWS opsin, by contrast, represents the first reported amphipod MWS opsin and is not closely related to currently known malacostracan MWS opsin. Taken together, these findings suggest that amphipod crustaceans are equipped with a minimally diversified set of three opsin genes. Given that early crustaceans possessed a larger number of opsin subfamilies, future studies are likely to pinpoint opsin gene losses along towards the amphipod clade in the crustacean tree of life. Most strikingly perhaps, this seems to include all of the four non-retinal opsin subfamilies. It remains to be seen whether these candidate gene losses were more intimately associated with the adaptation of *H. azteca* to its crepuscular visual ecology or reflects a more ancient trend in amphipods.

While median eyes appear to be absent from amphipods,[13] the organization compound eyes varies greatly,[14] reflecting their highly divergent visual ecologies of deep sea, planktonic, symbiotic, benthic, freshwater, and terrestrial species. The eyes of oceanic hyperiid amphipods, for instance, feature an exceptional variation in shape and size, with deep sea species possessing lens diameter differences between dorsal and ventral areas while surface-living species have smaller, more homogenous eye structures.[15] The mesopelagic *Streetsia challengeri* sports 2,500 in a medially fused cyclopic compound eye.[16] At the cellular level, all amphipod compound eyes are characterized by regressive features.[14] Only two instead of four cone cells secrete the lens, which covers only five photoreceptor cells instead of the canonical eight in most other Crustaceans and insects, i.e. Pancrustacea.[14, 17-19]

Little is known yet specifically on the visual organization and behavior of *H. azteca*. Preliminary inspection reveals that each of the pair of strongly pigmented compound eyes consists of approximately 40 ommatidia, indicative of vision adapted to low light levels, fitting well with the known benthic ecology of *H. azteca*.[20] The best reference species is likely the benthic amphipod *Pontoporeia affinis*,[21] which inhabits low light environments of the Baltic sea and in northern European lakes. *P. affinis* eye size ranges 40-50 ommatidia. Previous work indicated the expression of screening pigments with absorption maxima in the 460-500 nm and 540-580 nm ranges, and the function of a single opsin gene product with an absorption maximum at 548 nm. The circadian activity rhythm of the supratidal amphipod *Talorchestia longicornis* has been reported to be responsive to light and temperature. The species expresses two visual pigments with absorption maxima near 420 and 520 nm, and it is the latter which mediates the entrainment by light [22]. These data lead to the prediction that the *H. azteca* MWS1 and LWS opsins might account for the green range sensitivities between 540-580 nm thus far reported for gammarid eyes, while visual pigments account for the blue range sensitivities of amphipods, as previously suggested.

**Acknowledgements:** Kaley Major

| ortholog | scaffold | replaced models | Coordinates |
|---|---|---|---|
| MWS Opsin | Scaffold58 | HaztTmpA003622-RA HaztTmpA003621-RA | 498381-501855 |
| LWS Opsin 1 | Scaffold28 | HaztTmpA002759-RA | 1115868-1117013 |
| LWS Opsin 2 | Scaffold28 | HaztTmpA002760-RA | 1123101..1124216 |

**Table S4.13.1: Location and scaffold coordinates for the three opsin genes identified in the *H. azteca* genome**. Note the close linkage (6078 bp) between the LWS opsins on Scaffold 28.

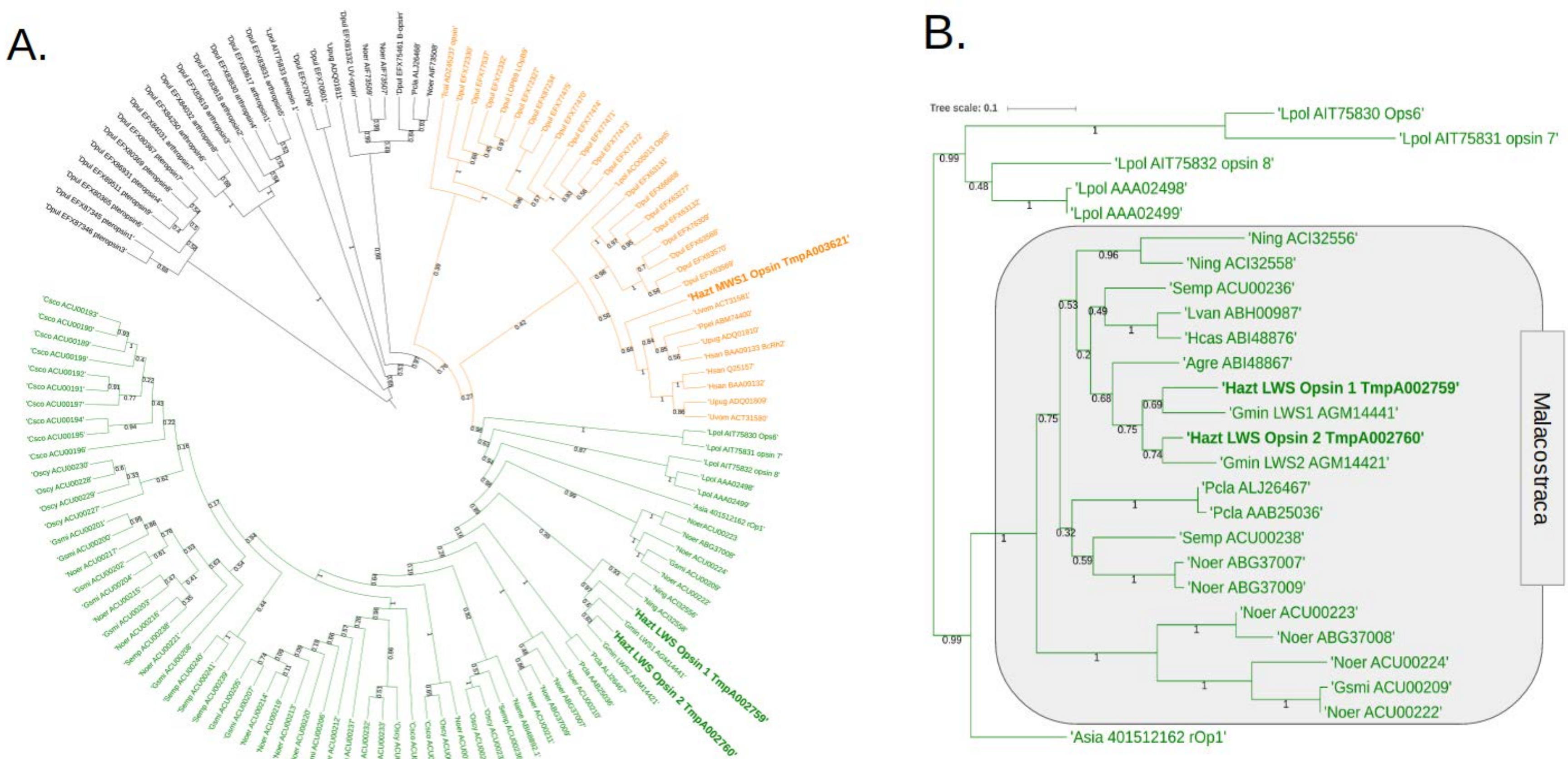



**Figure S4.12.1: Phylogenetic placement of the three H. azteca opsin genes. A.** Placement in a global phylogeny of crustacean opsin genes. Numbers at branches reflect nonparametric bootstrap support from 100 neighbor joining replicates. **B.** Placement of the two LWS opsin genes among publicly available malacostracan LWS opsins. Numbers at branches reflect nonparametric bootstrap support from 100 maximum likelihood replicates. Species abbreviations: Agre = *Archaeomysis grebnitzkii*, Asia = *Argulus siamensis*, Csco = *Coronis scolopendra*, Dpul = *Daphnia pulex*, Gmin = *Gammarus minus*, Gsmi = *Gonodactylus smithii*, Hazt = *Hyalella azteca*, Hcos = *Holmesimysis costata*, Hsan = *Hemigrapsus sanguineus*, Lpol = *Limulus polyphemus*, Lvan = *Litopenaeus vannamei*, Name = *Neomysis americana*, Ning = *Neognathophausia ingens*, Noer = *Neogonodactylus oerstedii*, Oscy = *Odontodactylus scyllarus*, Pcla = *Procambarus clarkii*, Ppel = *Portunus pelagicus*, Semp = *Squilla empusa*, Tcal = *Tigriopus californicus*, Upug = *Uca pugilator*, Uvom = *Uca vomis*

## S4.13. References


(1) Henze, M. J.; Oakley, T. H. The dynamic evolutionary history of pancrustacean eyes and opsins. *Integrative and Comparative Biology* **2015,** *55*, (5), 830-842.

(2) Brody, T.; Cravchik, A. *Drosophila melanogaster* G Protein–Coupled Receptors. *The Journal of Cell Biology* **2000,** *150*, (2), F83-F88.

(3) Colbourne, J. K.; Pfrender, M. E.; Gilbert, D.; Thomas, W. K.; Tucker, A.; Oakley, T. H.; Tokishita, S.; Aerts, A.; Arnold, G. J.; Basu, M. K. The ecoresponsive genome of *Daphnia pulex*. *science* **2011,** *331*, (6017), 555-561.

(4) Eriksson, B. J.; Fredman, D.; Steiner, G.; Schmid, A. Characterisation and localisation of the opsin protein repertoire in the brain and retinas of a spider and an onychophoran. *BMC Evolutionary Biology* **2013,** *13*, (1), 186.

(5) Hering, L.; Mayer, G. Analysis of the opsin repertoire in the tardigrade Hypsibius dujardini provides insights into the evolution of opsin genes in panarthropoda. *Genome Biology and Evolution* **2014,** *6*, (9), 2380-2391.

(6) Velarde, R. A.; Sauer, C. D.; Walden, K. K.; Fahrbach, S. E.; Robertson, H. M. Pteropsin: a vertebrate-like non-visual opsin expressed in the honey bee brain. *Insect Biochemistry and Molecular Biology* **2005,** *35*, (12), 1367-1377.

(7) Nagata, T.; Koyanagi, M.; Tsukamoto, H.; Terakita, A. Identification and characterization of a protostome homologue of peropsin from a jumping spider. *Journal of Comparative Physiology A* **2010,** *196*, (1), 51.

(8) Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **2004,** *32*, (5), 1792-1797.

(9) Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **2000,** *17*, (4), 540-552.

(10) Kumar, S.; Tamura, K.; Nei, M. MEGA: molecular evolutionary genetics analysis software for microcomputers. *Bioinformatics* **1994,** *10*, (2), 189-191.

(11) Löytynoja, A.; Goldman, N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* **2010,** *11*, (1), 579.

(12) Carlini, D. B.; Satish, S.; Fong, D. W. Parallel reduction in expression, but no loss of functional constraint, in two opsin paralogs within cave populations of *Gammarus minus* (Crustacea: Amphipoda). *BMC Evolutionary Biology* **2013,** *13*, (1), 89.

(13) Elofsson, R. The nauplius eye and frontal organs in malagostraca (crustacea). *Sarsia* **1965,** *19*, (1), 1-54.

(14) Hallberg, E.; Nilsson, H. L.; Elofsson, R. Classification of amphipod compound eyes—the fine structure of the ommatidial units (Crustacea, Amphipoda). *Zoomorphology* **1980,** *94*, (3), 279-306.

(15) Land, M. The eyes of hyperiid amphipods: relations of optical structure to depth. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* **1989,** *164*, (6), 751-762.

(16) Meyer-Rochow, V. B. The eyes of mesopelagic crustaceans. II. Streetsia challengeri (amphipoda). *Cell Tissue Res* **1978,** *186*, (2), 337-349.

(17) Buschbeck, E. K.; Friedrich, M. Evolution of insect eyes: tales of ancient heritage, deconstruction, reconstruction, remodeling, and recycling. *Evolution: Education and Outreach* **2008,** *1*, (4), 448-462.

(18) Mezzetti, M. C.; Scapini, F. Aspects of spectral sensitivity in Talitrus saltator (Montagu) (Crustacea-Amphipoda). *Marine and Freshwater Behaviour and Physiology* **1995,** *26*, (1), 35-45.

(19) Ugolini, A.; G. Borgioli; G. Galanti; L. Mercatelli; Hariyama, T. Photoresponses of the Compound Eye of the Sandhopper *Talitrus saltator* (Crustacea, Amphipoda) in the Ultraviolet-Blue Range. *The Biological Bulletin* **2010,** *219*, (1), 72-79.

(20) Wang, F.; Goulet, R. R.; Chapman, P. M. Testing sediment biological effects with the freshwater amphipod *Hyalella azteca*: the gap between laboratory and nature. *Chemosphere* **2004,** *57*, 1713-1724.
(21) Donner, K. O.; Langer, H.; Lindström, M.; Schlecht, P. Visual pigment, dark adaptation and rhodopsin renewal in the eye of Pontoporeia affinis (Crustacea, Amphipoda). *Journal of Comparative Physiology A* **1994,** *174*, (4), 451-459.
(22) Forward Jr, R. B.; Bourla, M. H.; Darnell, M. Z.; Cohen, J. H. Entrainment of the circadian rhythm of the supratidal amphipod Talorchestia longicornis by light and temperature: mechanisms of detection and hierarchical organization. *Marine and Freshwater Behaviour and Physiology* **2009,** *42*, (4), 233-247.

# S5. Gene Expression Data Sets

Simone Hasenbein
Aquatic Systems Biology Unit, Technical University of Munich, Freising, Germany

Correspondence to: simone.hasenbein@tum.de

**Table S5.1 Differentially expressed transcripts following exposure to cadmium.**

| Hazt_0.5.3 gene ID | Sequence Description | Log2ratio |
|---|---|---|
| GO:0023052: signaling | | |
| HAZT000761-RA-CDS | ankyrin-3-like isoform X3 | 1.407 |
| HAZT005466-RA-CDS | immune deficiency | 1.433 |
| HAZT007918-RA-CDS | activin receptor type-1 | 1.438 |
| GO:0044237: cellular metabolic process | | |
| HAZT002123-RA-CDS | transcription factor AP-2 | -0.539 |
| HAZT000808-RA-CDS | cAMP-responsive element modulator isoform X2 | 0.918 |
| HAZT001984-RA-CDS | Cyclic AMP-responsive element-binding 3 4 | 1.054 |
| HAZT002426-RA-CDS | transducin-like enhancer 4 isoform X2 | 1.662 |
| HAZT009794-RA-CDS | lysine--tRNA ligase isoform X2 | 0.773 |
| HAZT009654-RA-CDS | phosphatidylinositol phosphatase PTPRQ-like | 0.812 |
| HAZT003471-RA-CDS | 40S ribosomal S29 | 1.722 |
| HAZT006116-RA-CDS | DNA-binding RFX7-like | 1.011 |
| HAZT007307-RA-CDS | serine threonine- kinase ULK2 | 0.806 |
| HAZT007345-RA-CDS | cysteine--tRNA cytoplasmic-like | 1.274 |
| HAZT007854-RA-CDS | DNA-binding D-ETS-6-like | 1.692 |
| HAZT008119-RA-CDS | integrase core domain | 0.840 |
| HAZT009794-RA-CDS | lysine--tRNA ligase isoform X2 | 0.773 |
| HAZT010772-RA-CDS | transcription factor kayak isoform X3 | 0.834 |
| HAZT010857-RA-CDS | histone acetyltransferase p300-like | 1.736 |
| HAZT011247-RA-CDS | mothers against dpp | 0.685 |
| HAZT012199-RA-CDS | deoxynucleoside kinase-like | 2.198 |
| HAZT003221-RA-CDS | NFX1-type zinc finger-containing 1 | 0.994 |
| GO:0016043: cellular component organization | | |
| HAZT007870-RA-CDS | poly [ADP-ribose] polymerase 11-like | 2.268 |
| GO:0006950: response to stress, GO:0042221: response to chemical | | |
| HAZT011515-RA-CDS | integumentary mucin-like | 1.608 |
| HAZT009327-RA-CDS | ubiquitin ISG15 | 1.900 |
| HAZT002640-RA-CDS | thioredoxin peroxidase | 1.621 |
| HAZT005366-RA-CDS | multidrug resistance 1 | 0.849 |
| HAZT002784-RA-CDS | antilipopolysaccharide factor isoform 2 | 1.508 |
| HAZT011637-RA-CDS | glutathione S-transferase 2-like | 1.388 |
| HAZT011815-RA-CDS | glutathione S-transferase Mu 1-like isoform X2 | 1.148 |

**Table S5.1 continued**

| Hazt_0.5.3 gene ID | Sequence Description | Log2ratio |
|---|---|---|
| GO:0048856: anatomical strucutre development | | |
| HAZT004707-RA-CDS | Homeobox extradenticle | 0.874 |
| HAZT006868-RA-CDS | aminopeptidase | 1.999 |
| HAZT002071-RA-CDS | type I cytoskeletal 9-like | 1.897 |
| HAZT002420-RA-CDS | GSK-3-binding -like | 1.297 |
| GO:0071704: organic substance metabolic | | |
| HAZT011272-RA-CDS | cuticular analogous to peritrophins 3-A1 precursor | -0.360 |
| HAZT008523-RA-CDS | pancreatic lipase-related 2-like | -1.125 |
| HAZT000391-RA-CDS | zinc finger DPF3-like | 0.834 |
| HAZT009494-RA-CDS | cysteine sulfinic acid decarboxylase | 1.203 |
| HAZT001357-RA-CDS | GDP-fucose O-fucosyltransferase 2-like | 0.636 |
| HAZT004722-RA-CDS | homeodomain-interacting kinase 2 isoform X2 | 0.574 |
| HAZT005286-RA-CDS | ornithine decarboxylase-like | 0.637 |
| HAZT005410-RA-CDS | E3 ubiquitin- ligase RNF14-like | 1.086 |
| HAZT006199-RA-CDS | glucose-6-phosphate 1-dehydrogenase isoform X2 | 0.957 |
| HAZT007025-RA-CDS | mesencephalic astrocyte-derived neurotrophic factor homolog | 2.119 |
| HAZT008239-RA-CDS | tubulin polyglutamylase TTLL4 | 2.268 |
| HAZT009494-RA-CDS | cysteine sulfinic acid decarboxylase | 1.547 |
| HAZT010225-RA-CDS | myosin light chain smooth muscle-like | 0.885 |
| HAZT010374-RA-CDS | tyrosine decarboxylase | 1.440 |
| HAZT001425-RA-CDS | cyclin-dependent serine threonine- kinase | 1.771 |
| HAZT001132-RA-CDS | multidrug resistance-associated 5 | 0.949 |
| HAZT011471-RA-CDS | serine threonine- kinase pim-3-like | 1.180 |
| GO:0050789: regulation of biological process | | |
| HAZT006344-RA-CDS | LIM domain only 7 isoform X2 | -0.717 |
| HAZT006146-RA-CDS | ETS-related transcription factor Elf-5-like | -0.729 |
| HAZT001054-RA-CDS | Fibrillin-1 | -0.886 |
| HAZT004899-RA-CDS | tryptophan--tRNA cytoplasmic | 1.915 |
| HAZT007577-RA-CDS | nuclear pore complex DDB_G0274915-like | 1.418 |
| HAZT005438-RA-CDS | crustin-like peptide type | 1.433 |
| HAZT008887-RA-CDS | fizzy-related homolog | 1.990 |
| HAZT009188-RA-CDS | transcription factor SOX-14 | 0.943 |
| HAZT002980-RA-CDS | ubiquitin-conjugating enzyme E2 J2-like | 0.915 |
| GO:0065008: regulation of biological quality | | |
| HAZT000682-RA-CDS | ferritin peptide | 1.454 |
| GO:0007155: cell adhesion | | |
| HAZT005104-RA-CDS | serine-rich adhesin for platelets-like | 0.663 |
| HAZT007568-RA-CDS | CD63 antigen | 0.828 |
| HAZT011581-RA-CDS | A-agglutinin anchorage subunit | 1.312 |

**Table S5.1 continued**

| Hazt_0.5.3 gene ID | Sequence Description | Log2ratio |
|---|---|---|
| GO:0008219: cell death | | |
| HAZT007846-RA-CDS | dynamin-like 120 kDa mitochondrial isoform X5 | 1.335 |
| HAZT011233-RA-CDS | cathepsin L | 1.150 |
| GO:0006955: immune response | | |
| HAZT009324-RA-CDS | 2'-5'-oligoadenylate synthase 1 | 2.062 |
| GO:0007017: microtubule-based process | | |
| HAZT007649-RA-CDS | tubulin alpha-3 chain-like | 0.777 |
| GO:0051716: cellular response to stimulus | | |
| HAZT008240-RA-CDS | AN1-type zinc finger 2A | 2.556 |
| HAZT000821-RA-CDS | tyrosine- kinase BAZ1B-like | 0.627 |
| HAZT000282-RA-CDS | cold shock domain-containing CG9705 | 0.757 |
| HAZT000516-RA-CDS | transcription factor AP-1 | 1.200 |
| HAZT006418-RA-CDS | suppressor of cytokine signaling 2 | 1.671 |
| HAZT009155-RA-CDS | DNA polymerase beta | 1.053 |
| GO:0009605: response to external stimulus | | |
| HAZT004858-RA-CDS | polycystic kidney disease 1-like 2 | 1.530 |
| GO:0006807: nitrogen compound metabolic process | | |
| HAZT000034-RA-CDS | DBH-like monooxygenase 1 | -0.759 |
| HAZT001966-RA-CDS | single whey acidic domain-containing isoform 2 | 2.552 |
| HAZT010411-RA-CDS | Retrovirus-related Pol from type-2 retrotransposable element | 0.772 |
| HAZT000515-RA-CDS | tRNA-dihydrouridine(20a 20b) synthase [NAD(P)+]-like | 0.919 |
| HAZT000291-RA-CDS | serine ase 1 | 0.882 |
| HAZT000608-RA-CDS | mediator of RNA polymerase II transcription subunit 26 | 0.695 |
| HAZT006693-RA-CDS | nuclease HARBI1 | 0.871 |
| HAZT001814-RA-CDS | high-affinity choline transporter 1- | 1.228 |
| HAZT011834-RA-CDS | serine threonine- kinase SBK1 | 1.361 |
| HAZT010620-RA-CDS | serine protease easter-like | 2.824 |
| GO:0051234: establishment of localization, GO:1902578: single-organism localization | | |
| HAZT007219-RA-CDS | innexin 3 | 1.151 |
| HAZT007220-RA-CDS | innexin inx3 | 1.112 |
| HAZT000992-RA-CDS | zinc finger ZPR1 | 0.848 |
| HAZT000114-RA-CDS | mitochondrial thiamine pyrophosphate carrier | 1.131 |
| HAZT003399-RA-CDS | mitochondrial ornithine transporter 1 | 1.068 |
| HAZT001109-RA-CDS | Sodium-dependent nutrient amino acid transporter | 0.925 |
| HAZT001110-RA-CDS | Sodium-dependent nutrient amino acid transporter | 2.835 |
| HAZT001925-RA-CDS | sodium-dependent neutral amino acid transporter B(0)AT3-like | 1.302 |
| HAZT002520-RA-CDS | glutamate receptor delta-2-like | 1.746 |
| HAZT004580-RA-CDS | G -activated inward rectifier potassium channel 3-like | 0.793 |
| HAZT005033-RA-CDS | chloride channel 2-like | 0.772 |
| HAZT010399-RA-CDS | Y+L amino acid transporter | 1.584 |
| HAZT011007-RA-CDS | PREDICTED: uncharacterized protein LOC108675335 | 3.543 |

| Hazt_0.5.3 gene ID | Sequence Description | Log2ratio |
|---|---|---|
| GO:0006457: protein folding | | |
| HAZT005785-RA-CDS | heat shock 90 | 2.831 |
| HAZT006327-RA-CDS | dnaJ homolog 1 | 0.878 |
| HAZT007650-RA-CDS | dnaJ homolog subfamily A member 1 | 0.893 |
| HAZT008078-RA-CDS | heat shock 60 | 1.451 |
| HAZT011866-RA-CDS | heat shock 70 kDa cognate 4-like | 2.296 |
| HAZT004026-RA-CDS | heat shock 70 kDa 1-like | 0.685 |
| Other (no GO Term) | | |
| HAZT008455-RA-CDS | LIX1 isoform X2 | -0.733 |
| HAZT004838-RA-CDS | probable domain-containing histone demethylation 2C | -0.654 |
| HAZT002072-RA-CDS | PREDICTED: uncharacterized protein LOC108673944 | 3.443 |
| HAZT002229-RA-CDS | PREDICTED: uncharacterized protein LOC108668185 | 1.429 |
| HAZT002862-RA-CDS | PREDICTED: uncharacterized protein LOC108672909 | 0.546 |
| HAZT003731-RA-CDS | ---NA--- | 1.883 |
| HAZT003746-RA-CDS | ---NA--- | 2.091 |
| HAZT006015-RA-CDS | PREDICTED: uncharacterized protein LOC108683374 | 3.835 |
| HAZT006241-RA-CDS | PREDICTED: uncharacterized protein LOC108665911 | 0.819 |
| HAZT007069-RA-CDS | PREDICTED: uncharacterized protein LOC108667399 | 0.913 |
| HAZT007151-RA-CDS | PREDICTED: uncharacterized protein LOC106074525, partial | 1.066 |
| HAZT008358-RA-CDS | PREDICTED: uncharacterized protein LOC108678020 | 2.410 |
| HAZT008924-RA-CDS | PREDICTED: uncharacterized protein LOC108676787 | 2.134 |
| HAZT008927-RA-CDS | PREDICTED: uncharacterized protein LOC108665752 | 4.658 |
| HAZT009215-RA-CDS | PREDICTED: uncharacterized protein LOC108674485 | 1.828 |
| HAZT009372-RA-CDS | PREDICTED: putative uncharacterized protein DDB_G0286901 | 0.813 |
| HAZT009397-RA-CDS | Transcription elongation factor | 1.125 |
| HAZT009734-RA-CDS | PREDICTED: uncharacterized protein LOC108670605 | 1.343 |
| HAZT010287-RA-CDS | dentin sialophospho -like | 1.594 |
| HAZT010758-RA-CDS | PREDICTED: uncharacterized protein LOC108667785 | 1.545 |
| HAZT011738-RA-CDS | PREDICTED: uncharacterized protein LOC108667128 | 1.195 |
| HAZT008241-RA-CDS | PREDICTED: uncharacterized protein LOC108664476 | 1.125 |

**Table S5.1 Differentially expressed transcripts following exposure to cadmium.** Differentially expressed contigs were aligned to the *H. azteca* genome using blastn to identify full length transcripts and corresponding gene ID from the official gene set (OGS; Hazt_0.5.3). A total of 395 unique transcripts were identified. Differentially expressed transcripts were functionally annotated manually and sequences were then mapped to GO terms using Blast2Go. Differentially expressed transcripts were then grouped according to similar GO terms and are shown here with the average log2 transformed expression ratio of Cd vs. solvent control.

| Hazt_0.5.3 gene ID | Sequence Description | ZnS_low vs ctrl | ZnS_high vs ctrl | ZnO low vs ctrl | ZnO high vs ctrl |
|---|---|---|---|---|---|
| GO:0006950: response to stress | | | | | |
| HAZT000015-RA-CDS | Chorion peroxidase | 0.107 | 1.453 | 0.705 | 1.090 |
| GO:0042221: response to chemical | | | | | |
| HAZT001749-RA-CDS | Glutamate receptor 1 | -0.028 | -0.148 | -0.055 | -0.511 |
| GO:0051234: establishment of localization, GO: 0042221: response to chemical | | | | | |
| HAZT003132-RA-CDS | MFS-type transporter SLC18B1-like | -0.273 | -0.860 | -0.187 | -0.586 |
| HAZT005362-RA-CDS | multidrug resistance homolog 49 isoform X1 | -0.687 | -0.612 | -0.362 | -0.552 |
| HAZT003399 | mitochondrial ornithine transporter 1 | 0.155 | 0.679 | 0.038 | 0.336 |
| HAZT000118-RA-CDS | glycine-rich cell wall structural -like | 0.597 | 1.864 | 1.226 | 1.814 |
| HAZT007519-RA-CDS | uncharacterized PE-PGRS family PE_PGRS54-like | -0.110 | -0.262 | -0.037 | -0.636 |
| GO:0071704: organic substance metabolic process, GO:0006807: nitrogen compound metabolic process | | | | | |
| HAZT011272-RA-CDS | cuticular analogous to peritrophins 3-A1 precursor | 0.216 | 2.102 | 0.709 | 1.436 |
| HAZT008429-RA-CDS | aspartyl asparaginyl beta-hydroxylase-like | -0.100 | -0.300 | -0.251 | 0.180 |
| HAZT009726-RA-CDS | twitchin-like isoform X16 | -0.139 | -0.130 | -0.206 | -0.283 |
| HAZT004549-RA-CDS | TRAF and TNF receptor-associated | 0.285 | 0.065 | 0.017 | 0.214 |
| GO:0055114: oxidation-reduction process | | | | | |
| HAZT006913-RA-CDS | 3-hydroxyacyl- dehydrogenase type-2 | 0.406 | 0.345 | 0.006 | 0.300 |
| GO:0048869: cellular developmental process | | | | | |
| HAZT008835-RA-CDS | N-alpha-acetyltransferase auxiliary subunit | -0.176 | -0.254 | 0.016 | -0.261 |
| HAZT005749-RA-CDS | myb-related B-like | -0.115 | -0.013 | -0.235 | -0.496 |
| GO:0048856: anatomical structure development, GO:0032501: multicellular organismal process | | | | | |
| HAZT008291-RA-CDS | 2-hydroxyacylsphingosine 1-beta-transferase | -0.331 | -0.696 | -0.122 | -0.762 |
| Other (no GO Term) | | | | | |
| HAZT000356-RA-CDS | probable H ACA ribonucleo complex subunit | 1.068 | 1.465 | 1.295 | 2.063 |
| HAZT003524-RA-CDS | jmjC domain-containing 8-like | 0.166 | 0.012 | -0.073 | -0.206 |
| HAZT006852-RA-CDS | uncharacterized LOC106124634 precursor | 0.674 | 1.383 | 1.391 | 1.871 |
| HAZT006934-RA-CDS | neuralized 4 | 0.077 | 0.065 | -0.199 | -0.248 |
| HAZT008927-RA-CDS | PREDICTED: uncharacterized protein LOC108665752 | -0.162 | 1.606 | 0.701 | 1.752 |

**Table S5.2 Differentially expressed transcripts following exposure to zinc or ZnO NPs.** Differentially expressed contigs were aligned to the *H. azteca* genome using blastn to identify full length transcripts and corresponding gene ID from the official gene set (OGS; Hazt_0.5.3). A total of 60 unique transcripts were identified. Differentially expressed transcripts were functionally annotated manually and sequences were then mapped to GO terms using Blast2Go. Differentially expressed transcripts were then grouped according to similar GO terms and are shown here with the average log2 transformed expression ratio of treatment vs. control.

**Table S5.3 Differentially expressed transcripts following exposure to cyfluthrin.**

| Hazt_0.5.3 gene ID | Sequence Description | Log2ratio |
|---|---|---|
| GO:0065009: regulation of molecular function | | |
| HAZT005627-RA-CDS | small G signaling modulator 1- | 0.760 |
| GO:0051234: establishment of localization | | |
| HAZT004443-RA-CDS | facilitated trehalose transporter Tret1-like | -1.807 |
| HAZT005476-RA-CDS | complexin isoform X1 | 1.000 |
| HAZT004047-RA-CDS | ROP isoform X1 | 1.325 |
| GO:0033036: macromolecule localization | | |
| HAZT008393-RA-CDS | Shroom isoform X4 | 0.825 |
| GO:0065008: regulation of biological quality | | |
| HAZT007293-RA-CDS | alpha-tocopherol transfer -like | 0.707 |
| HAZT006047-RA-CDS | plasma membrane calcium-transporting ATPase 3 | 0.754 |
| GO:0023052: signaling | | |
| HAZT011551-RA-CDS | ras GTP exchange son of sevenless | -0.802 |
| GO:0006950: response to stress | | |
| HAZT006090-RA-CDS | alpha,alpha-trehalose-phosphate synthase [UDP-forming] | -0.592 |
| HAZT008461-RA-CDS | integumentary mucin -like | 0.608 |
| HAZT006430-RA-CDS | leucine--tRNA cytoplasmic | 1.026 |
| GO:0071704: organic substance metabolic | | |
| HAZT004762-RA-CDS | argininosuccinate lyase | -0.628 |
| HAZT003539-RA-CDS | D-3-phosphoglycerate dehydrogenase | -0.912 |
| HAZT004791-RA-CDS | 26S protease regulatory subunit 6B | -0.756 |
| HAZT000131-RA-CDS | RNA polymerase I-specific transcription initiation factor | -0.591 |
| HAZT006090-RA-CDS | alpha,alpha-trehalose-phosphate synthase [UDP-forming] | -0.592 |
| HAZT008553-RA-CDS | Gypsy retrotransposon integrase | 0.530 |
| GO:0032259: methylation | | |
| HAZT001523-RA-CDS | methyltransferase C9orf114 isoform X1 | -0.235 |
| GO:0044085: cellular component biogenesis | | |
| HAZT011764-RA-CDS | nucleolar GTP-binding 2 | 0.588 |
| GO:0044237: cellular metabolic process | | |
| HAZT003186-RA-CDS | COP9 signalosome complex subunit 2 | -0.121 |
| HAZT003685-RA-CDS | autophagy-related 101 | -0.754 |
| HAZT011265-RA-CDS | Ornithine decarboxylase antizyme 1 | -0.492 |
| HAZT006038-RA-CDS | enoyl- delta isomerase mitochondrial-like | -0.482 |
| HAZT003807-RA-CDS | serine arginine repetitive matrix 2-like | 0.600 |
| HAZT001184-RA-CDS | RNA polymerase II polypeptide A small phosphatase 1 | 0.315 |
| HAZT009481-RA-CDS | glucosylceramidase 3 | 1.106 |
| HAZT009461-RA-CDS | U7 snRNA-associated Sm LSm10 | 0.771 |
| HAZT009494-RA-CDS | cysteine sulfinic acid decarboxylase | 0.468 |

| Hazt_0.5.3 gene ID | Sequence Description | Log2ratio |
|---|---|---|
| GO:0048869: cellular developmental process | | |
| HAZT007949-RA-CDS | Cullin-associated NEDD8-dissociated 1 | -0.521 |
| Other (no GO Term) | | |
| HAZT008134-RA-CDS | PREDICTED: putative uncharacterized protein DDB_G0278921 | 0.707 |
| HAZT009358-RA-CDS | PREDICTED: uncharacterized protein LOC108675535 | 0.746 |
| HAZT000149-RA-CDS | PREDICTED: uncharacterized protein LOC108675536 | 0.676 |
| HAZT011393-RA-CDS | PREDICTED: uncharacterized protein LOC108675537 | 0.605 |

**Table S5.3 Differentially expressed transcripts following exposure to cyfluthrin.** Differentially expressed contigs were aligned to the *H. azteca* genome using blastn to identify full length transcripts and corresponding gene ID from the official gene set (OGS; Hazt_0.5.3). A total of 126 unique transcripts were identified. Differentially expressed transcripts were functionally annotated manually and sequences were then mapped to GO terms using Blast2Go. Differentially expressed transcripts were then grouped according to similar GO terms and are shown here with the average log2 transformed expression ratio of cyfluthrin exposed vs. control.

| Hazt_0.5.3 gene ID | Sequence Description | Log2ratio |
|---|---|---|
| GO:0006807: nitrogen compound metabolic | | |
| HAZT010921-RA-CDS | Vanin 1 | 1.019 |
| GO:0008150-ND,GO:0016021: integral component of membrane | | |
| HAZT002331-RA-CDS | Transmembrane 53 (transmembrane protein of unknown function) | 1.273 |
| HAZT010075-RA-CDS | Transmembrane 53 (transmembrane protein of unknown function) | 0.839 |
| Other (no GO Term) | | |
| HAZT007105-RA-CDS | Growth horomone secretagogue receptor type 1-like | -1.984 |
| HAZT003591-RA-CDS | thyroid hormone-inducible hepatic spot 14 | 0.609 |

**Table S5.4 Differentially expressed transcripts following exposure to PCB126.** Differentially expressed contigs were aligned to the *H. azteca* genome using blastn to identify full length transcripts and corresponding gene ID from the official gene set (OGS; Hazt_0.5.3). A total of 21 unique transcripts were identified. Differentially expressed transcripts were functionally annotated manually and sequences were then mapped to GO terms using Blast2Go. Differentially expressed transcripts were then grouped according to similar GO terms and are shown here with the average log2 transformed expression ratio of PCB126 vs. solvent control.

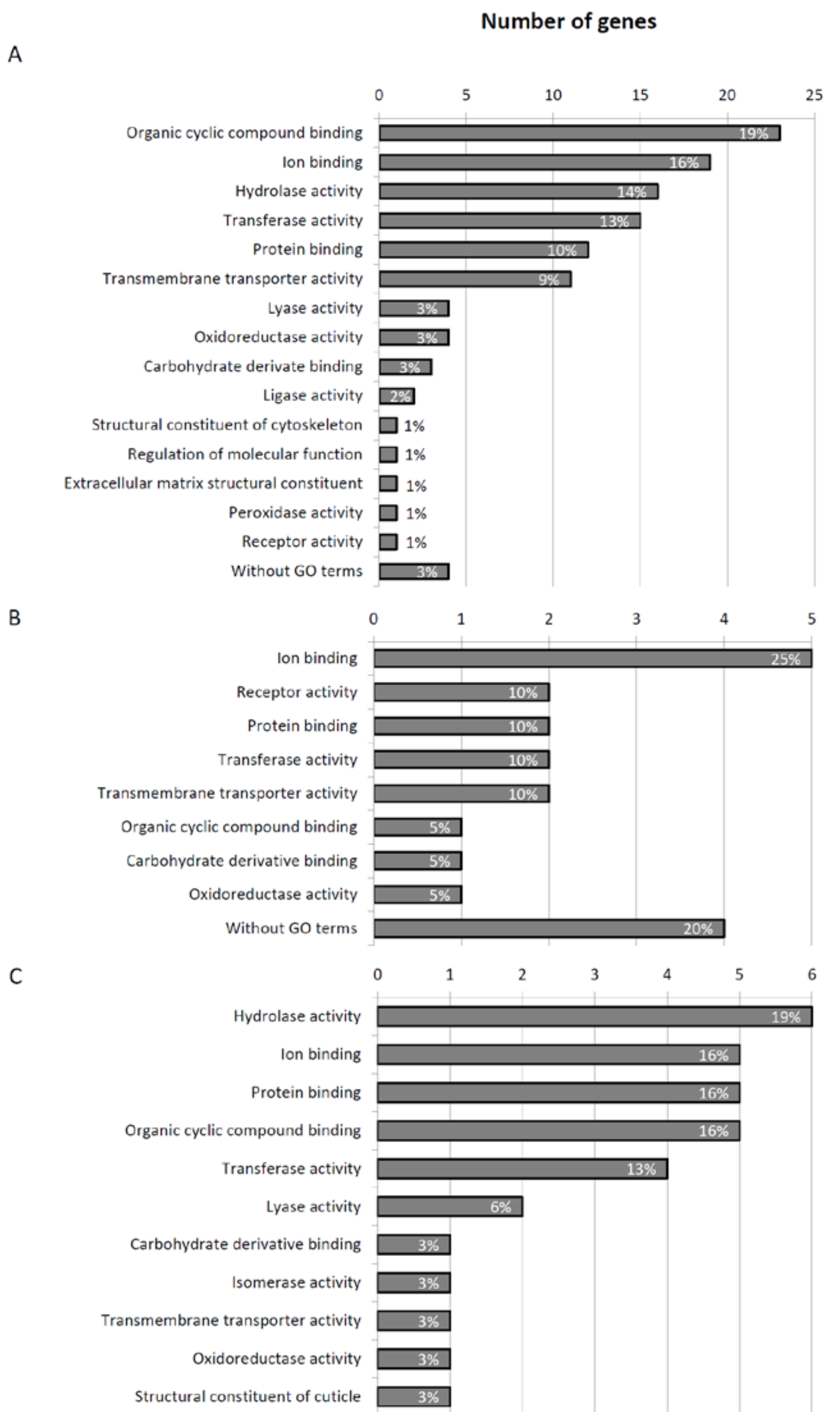



**Figure S5: Molecular processes gene ontology (GO) terms representing the differentially expressed transcripts from Cd (A), Zn and ZnO NPs (B), and cyfluthrin (C)**. The number of genes mapped to each of the GO terms is shown by the length of the bars, while the percentage of total transcripts is marked at the end of each bar. For PCB126, only one of the 12 annotated transcripts were mapped to biological processes GO terms. Similar graphs for biological processes GO terms can be found in the main text (Figure 3).

# S6 Additional Supplemental File Descriptions

**Supplemental File S6.1, excel:**

**Table S6.1: Summary of miRNAs identified from *Hyalella azteca*.** *H*. *azteca* miRNAs were predicted as described in S.2. All animal mature miRNAs (miRBase Release 21) were searched against *H*. *azteca* genome by BLAST (e-value < -4) for potential miRNA coding sites. A total of 1,261 candidate miRNA coding sites were identified by BLAST with 148 predicted after hairpin structure identification.

**Table S6.2: Cytochrome P450 Sequences and Official Names.** P450 genes were identified in the *Hyalella azteca* genome as described in S4.3. The P450 sequences were named by the cytochrome P450 nomenclature committee (Dr. D.R. Nelson, University of Tennessee). Names were determined using evolutionary relationships based on expansive phylogenetic trees consisting of the total known complement of P450s. While there are no strict percent identity requirements, generally P450s in the same family share at least 40% identity, while those in the same subfamily share 55% identity. Clans are a level designated by the nomenclature committee with generally lower percent identity among members, and are determined by clade groupings on the same phylogeny used for naming.

**Supplemental File S6.2, text:**

**Chemoreceptor Sequence File.** Protein sequences for the crustacean (*Hyalella azteca*, *Daphnia pulex*, and *Eurytemora affinis*) chemoreceptors described in S4.1. This includes 155 HaztGr proteins (p1-26), 3 updated DpulGr proteins (p 26), 67 EaffGr proteins (p26-37), 118 HaztIr proteins (p37-63), 154 DpulIr proteins (p63-94) and 22 EaffIr proteins (p94-100).

## S7. Author Contributions

***Writing team and project organization***
Helen C. Poynton, Simone Hasenbein, Joshua B. Benoit, Maria S. Sepulveda, Stephen Richards

HCP recruited and organized the manual annotation project, participated in the annotation, analyzed gene expression data, provided gDNA and RNA, and wrote the manuscript.

SH participated in the annotation, analyzed the gene expression data, and wrote the manuscript.

JBB participated in the annotation, analyzed genome quality and completeness, and wrote the manuscript.

MSS organized the epigenetics and miRNA sections, constructed the *H. azteca* distribution map, contributed to the Cd and PCB126 gene expression data, and wrote the manuscript.

SR conceived the project, managed genome sequencing, assembly and automated annotation of the genome, and wrote the corresponding methods for manuscript.

***Manual annotation teams***
Shuai Chen, Karl M. Glastad, Michael A. D. Goodisman, John H. Warren, Joseph H. Vineis, Jennifer L. Bowen, Markus Friedrich, Jeffery Jones, Hugh M. Robertson, René Feyereisen, Alexandra Mechler-Hickson, Nicholas Mathers, Carol Eunmi Lee, Andrew J. Rosendale, Andrew G. Cridge, Monica C. Munoz-Torres, Peter A. Bain, Austin R. Manny, Kaley M. Major, Faith Lambert, Christopher D. Vulpe

SC completed the miRNA analysis and wrote the corresponding results section.

KMG and MADG conducted the DNA methylation analysis and wrote the corresponding results section.

JHW identified bacterial contaminants in the genome, JHV and JLB assembled and annotated the bacterial genomes, and all wrote the corresponding results section.

MF and JJ annotated the opsin genes and wrote the corresponding results section.

HMR annotated the chemoreceptor genes and wrote the corresponding results section.

RF, AMH, and CEL annotated the P450 genes and wrote the corresponding results section.

NM and CEL annotated the ion transporters and wrote the corresponding results sections.

AJR, AGC, MMT, PB, ARM, KM, FL, and CDV participated in the manual annotation and wrote annotation reports.

***Genome assembly, automated annotation, and genome curation teams***
Monica F. Poelchau, Daniel S.T. Hughes, Swetha C. Murali, Monica C. Munoz-Torres, Yu-Yu Lin, John H. Werren, Mei-ju May Chen, Christopher P. Childers

MFP led the curation quality control (QC) steps, generated the Official Gene Set using Merge, and wrote the corresponding methods section for the manuscript.

DSTH conducted the genome assembly and automated annotation on Hazt_1.0.

SCM conducted the genome assembly and automated annotation on Hazt_1.0.

MCMT conducted training sessions for all annotators.

YYL removed bacterial contamination from the OGS.

JHW identified bacterial contamination from the OGS.

MJMC wrote the QC and Merge software.

CPC set up and maintained the JBrowse and Apollo instances.

***Genome sequencing team***
Jiaxin Qu, Shannon Dugan, Sandra L. Lee, Hsu Chao, Huyen Dinh, Yi Han, HarshaVardhan Doddapaneni, Kim C. Worley, Donna M. Muzny, Richard A. Gibbs

JQ, SD, SLL, HC, HD, YH, HD, KCW, DMM, RAG conducted the genome sequencing and assembly.

***Other contributors***
John K. Colbourne, Adam Biales, J. Spencer Johnston, Gary A. Wellborn, Padrig Tuck, Bonnie Blalock, Mark E. Smith, Hugo Ochoa-Acuña

JKC performed microarray analysis on Cd and PCB126 exposed *H. azteca.*

AB contributed a regulatory perspective and wrote the corresponding sections of the manuscript.

JSJ conducted flow cytometry to determine genome size.

GAW provided data for the *Hyalella* distribution map.

PT provided the gDNA for the genome sequencing.

BB created the in-bred *Hyalella azteca* lines.

MES performed exposures for the Cd and PCB126 gene expression analysis.

HOA contributed to the microarray construction.

**Email addresses of all contributors:**

Helen C. Poynton, helen.poynton@umb.edi
Simone Hasenbein, simone.hasenbein@tum.de
Joshua B. Benoit, benoitja@ucmail.uc.edu
Maria S. Sepulveda, mssepulv@purdue.edu
Monica F. Poelchau, monica.poelchau@ars.usda.gov
Daniel S.T. Hughes, dsthughes@gmail.com
Swetha C. Murali, shwethacm@gmail.com
Shuai Chen, shuai@purdue.edu
Karl Gladstad, karlglastad@gmail.com
Michael Goodisman, michael.goodisman@biology.gatech.edu
John H. Werren, jack.werren@rochester.edu
Joseph H. Vineis, jvineis@gmail.com
Jennifer L. Bowen, je.bowen@northeastern.edu
Markus Friedrich, friedrichwsu@gmail.com
Jeffery Jones, jonesjefferyw@gmail.com
Hugh Robertson, hughrobe@uiuc.edu
René Feyereisen, rene.feyereisen@plen.ku.dk
Alexandra Mechler-Hickson, alexandra.mechlerhickson@wisc.edu
Nicholas Mathers, nmathers@wisc.edu
Carol Eunmi Lee, carollee@wisc.edu
John K. Colbourne, j.k.colbourne@bham.ac.uk
Adam Biales, biales.adam@epa.gov
J. Spencer Johnston, spencerj@tamu.edu
Gary Wellborn, gwellborn@ou.edu
Andrew J. Rosendale, rosendaw@ucmail.uc.edu
Andrew Cridge, andrew.cridge@otago.ac.nz
Monica Munoz-Torres, moni@phoenixbioinformatics.org
Peter Bain, peter.bain.0@gmail.com
Austin R. Manny, austinmanny@g.harvard.edu
Kaley Major, kaley.major@gmail.com
Faith Lambert, fnlambert@ufl.edu
Chris Vulpe, cvulpe@ufl.edu
Padrig Tuck, padrig@gmail.com
Bonnie Blalock, bonnie.blalock001@umb.edu
Yu-Yu Lin, ifishlin324@gmail.com
Mark E. Smith, NA
Hugo Ochoa-Acuña, hugo.ochoa.acuna@gmail.com
Mei-ju May Chen, arbula@gmail.com
Christopher P. Childers, christopher.childers@ars.usda.gov
Jiaxin Qu, jiaxinqu@gmail.com
Shannon Dugan, sdugan@bcm.edu
Sandra L. Lee, sllee@bcm.edu
Hsu Chao, hchao@bcm.edu
Huyen Dinh, hdinh@bcm.edu
Yi Han, yhan@bcm.edu
HarshaVardhan Doddapaneni, doddapan@bcm.edu
Kim C. Worley, kworley@bcm.edu
Donna M. Muzny, donnam@bcm.edu
Richard A. Gibbs, agibbs@bcm.edu
Stephen Richards, stephenr@bcm.edu