The Toxicogenome of *Hyalella azteca*: A Model for Sediment Ecotoxicology and Evolutionary Toxicology


† School for the Environment, University of Massachusetts Boston, Boston, Massachusetts 02125 United States
‡ Aquatic Systems Biology Unit, Technical University of Munich, D-85354 Freising, Germany
§ Department of Biological Sciences, University of Cincinnati, Cincinnati, Ohio 45221 United States
∥ Forestry and Natural Resources, Purdue University, West Lafayette, Indiana 47907 United States
¶ Agricultural Research Service, National Agricultural Library, U.S. Department of Agriculture, Beltsville, Maryland 20705 United States
‖ Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030 United States
© Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104 United States
◆ School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332 United States
○ Biology Department, University of Rochester, Rochester, New York 14627 United States
∞ Department of Marine and Environmental Sciences, Marine Science Center, Northeastern University, Nahant, Massachusetts 01908 United States
☆ Department of Biological Sciences, Wayne State University, Detroit Michigan 48202 United States
> Department of Entomology, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801 United States
Ω Department of Plant and Environmental Sciences, University of Copenhagen, DK-1871 Frederiksberg, Denmark
∇ Center of Rapid Evolution (CORE) and Department of Integrative Biology, University of Wisconsin, Madison, Wisconsin 53706 United States
○ Department of Biosciences, University of Birmingham, BirminghamB15 2TT U.K.
/ National Exposure Research Laboratory, United States Environmental Protection Agency, Cincinnati, Ohio 45268 United States
\ Department of Entomology, Texas A&M University, College Station, Texas 77843 United States
π Department of Biology, University of Oklahoma, Norman, Oklahoma 73019 United States
ρ Laboratory for Evolution and Development, Department of Biochemistry, University of Otago, Dunedin, 9054 New Zealand
● Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720 United States
% Commonwealth Scientific and Industrial Research Organisation (CSIRO), Urrbrae SA 5064 Australia
◎ Department of Microbiology & Cell Science, University of Florida, Gainesville, Florida 32611 United States

Received: February 12, 2018
Revised: April 6, 2018
Accepted: April 10, 2018
Published: April 10, 2018
ABSTRACT: *Hyalella azteca* is a cryptic species complex of epibenthic amphipods of interest to ecotoxicology and evolutionary biology. It is the primary crustacean used in North America for sediment toxicity testing and an emerging model for molecular ecotoxicology. To provide molecular resources for sediment quality assessments and evolutionary studies, we sequenced, assembled, and annotated the genome of the *H. azteca* U.S. Lab Strain. The genome quality and completeness is comparable with other ecotoxicological model species. Through targeted investigation and use of gene expression data sets of *H. azteca* exposed to pesticides, metals, and other emerging contaminants, we annotated and characterized the major gene families involved in sequestration, detoxification, oxidative stress, and toxicant response. Our results revealed gene loss related to light sensing, but a large expansion in chemoreceptors, likely underlying sensory shifts necessary in their low light habitats. Gene family expansions were also noted for cytochrome P450 genes, cuticle proteins, ion transporters, and include recent gene duplications in the metal sequestration protein, metallothionein. Mapping of differentially expressed transcripts to the genome significantly increased the ability to functionally annotate toxicant responsive genes. The *H. azteca* genome will greatly facilitate development of genomic tools for environmental assessments and promote an understanding of how evolution shapes toxicological pathways with implications for environmental and human health.

INTRODUCTION

Sediment quality assessments serve as a metric for overall habitat integrity of freshwater ecosystems. Sediments provide a foundation for aquatic food webs, providing habitat for invertebrate species including crustaceans and insect larvae. However, they also concentrate pollution over time, especially hydrophobic contaminants that sorb to sediments, leading to bioaccumulation in the food web. While concentrations of a few legacy contaminants are declining in the United States due to regulatory efforts; chemicals designed as their replacements are becoming emerging contaminants that sorb to sediments, leading to bioaccumulation in the environment by chemicals. Here we describe the genome of *H. azteca* U.S. Lab Strain, the most widely used species within the complex for ecotoxicology, a relatively large amount of toxicity data has been amassed for the U.S. Lab Strain of *H. azteca*. Detailed characterization of the toxicogenome creates opportunities to reinterpret and exploit existing toxicity data to better predict risk posed to the environment by chemicals. Here we describe the genome of *H. azteca* with particular emphasis on genes related to key toxicological targets and pathways including detoxification, stress response, developmental and sensory processes, and ion transport. In addition, we begin the process of creating a “gene ontology” for environmental toxicology using *H. azteca* by...
annotating the function of genes responsive to model sediment contaminants, thereby shinning light on the toxicogenome underlying adverse outcome pathways.

MATERIALS AND METHODS

Hyalella azteca Strain, Inbreeding, and Genomic DNA (gDNA) Extraction. Hyalella azteca (U.S. Laboratory Strain) cultures were reared according to standard test conditions. These organisms have been maintained by the U.S. Environmental Protection Agency since their original collection by A. Nebeker (ca. 1982) from a stream near Corvallis, Oregon. They share highest genetic similarity with populations collected in Florida and Oklahoma (Figure 1, Clade 8) and recently in California. Several lines of full sibling matings were maintained for four generations, after which all lines were unable to produce offspring, likely due to inbreeding depression. Twenty animals including both males and females from a single inbred line were collected and gDNA extracted from individual H. azteca using the QIAGEN DNeasy Blood & Tissue Kit (QIAGEN, Germantown, MD) with slight modifications. Because of the gDNA quantities required, multiple individuals were pooled for library construction.

gDNA Sequencing, Assembly, and Annotation. H. azteca is one of 30 arthropod species sequenced as part of a pilot project for the i5K Arthropod Genomes Project at the Baylor College of Medicine Human Genome Sequencing Center, Houston, TX. Sequencing was performed on Illumina HiSeq 2000s (CASVA v. 1.8.3_V3) generating 100 bp paired end reads. The amount of sequences generated from each of four libraries (nominal insert sizes 180 bp, 500 bp, 3 kb, and 8 kb) is noted in Supporting Information (SI) Table S1 with NCBI SRA accessions. See SI S1 for more details on library preparation. Reads were assembled using ALLPATHS-LG (v35218) and further scaffolded and gap-filled using in-house tools Atlas-Link (v.1.0) and Atlas gap-fill (v.2.2) (https://www.hgsc.bcm.edu/software/). This yielded an initial assembly (HAZT_1.0; SI Table S1; NCBI Accession GCA_000764305.1) of 1.18 Gb (596.68 Mb without gaps within scaffolds), compared with genome size of 1.05 Gb determined by flow cytometry (see SI for methods). To improve assembly contiguity, we used the Redundans assembly tool. With Redundans using standard parameters, HAZT_1.0 scaffolds and all Illumina input reads given to ALLPATHS-LG when producing HAZT_1.0 as data inputs, generated a new assembly (HAZT_2.0, SI Table S1; NCBI accession GCA_000764305.2) of 550.9 Mb (548.3 Mb without gaps within scaffolds).

The HAZT_1.0 genome assembly was subjected to automatic gene annotation using a Maker 2.0 annotation pipeline tuned specifically for arthropods. The core of the pipeline was a Maker 2 instance, modified slightly to enable efficient running on our computational resources. See SI for additional details. The automated gene sets are available from the BCM-HGSC Web site (https://hgsc.bcm.edu/arthropods/hyalella-azteca-genome-project), Ag Data Commons and the National Agricultural Library (https://i5k.nal.usda.gov/Hyalella_azteca) where a web-browser of the genome, annotations, and supporting information is accessible. The HAZT_2.0 assembly was annotated by the automated NCBI Eukaryotic Genome Annotation Pipeline and is available from NCBI (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Hyalella_azteca/100/). Manual Annotation and Official Gene Set Generation. Automated gene prediction greatly facilitates the generation of useful genomic annotations for downstream research; however,
producing accurate, high-quality genome annotations remains a challenge.\textsuperscript{33} Manual correction of gene models generated through automated analyses, also known as manual annotation, can provide improved resources for downstream projects.\textsuperscript{38} The \textit{Hyalella} genome consortium recruited 25 annotators to improve gene models predicted from the genome assembly scaffolds,\textsuperscript{39} adhering to a set of rules and guidelines during the manual annotation process (https://isk.nal.usda.gov/content/rules-web-apollo-annotation-isk-pilot-project). Manual annotation occurred via the Apollo software,\textsuperscript{38} which allows users to annotate collaboratively on the web via the JBrowse genome browser, version 1.0.4\textsuperscript{40} (https://apollo.nal.usda.gov/hyazz/jbrowse/). Two transcriptomes (see below) were provided as external evidence for manual annotation and as an additional resource to search for missing genes. After manual annotation, models were exported from Apollo and screened for general formatting and curation errors (see https://github.com/NAL-i5K/ISKNAL_OGS/wiki). Models that overlapped with potential bacterial contaminants were removed. Bacterial contamination included regions identified via the procedure outlined in the SI S1, S2 as well as potential contamination identified by NCBI (Terence Murphy, personal communication). The remaining corrected models were then merged with MAKER gene predictions HAZTv_0.5.3 and miRNA predictions (see SI S3) into a nonredundant gene set, OGSv1.0;\textsuperscript{41} for details on the merge procedure see https://github.com/NAL-i5K/ISKNAL_OGS/wiki/Merge-phase. The manual annotation process generated 911 corrected gene models, including 875 mRNAs and 46 pseudogenes. All annotations are available for download at the i5k Workspace@NAL (https://isk.nal.usda.gov/Iskatalog/). Additional details pertaining to the annotation of specific gene families can be found with the annotation reports of the SI S4.

**RNA Sequencing and Transcriptome Libraries.** Two sets of transcriptomic data were generated from nonexposed \textit{H. azteca} to assist in gene prediction, and were recently published as part of a de novo transcriptome assembly project to identify peptide hormones\textsuperscript{33} (see S1 for details). RNaseq reads from this transcriptome project were aligned to the \textit{H. azteca} genome scaffolds (HAZTv_1.0) using TopHat 2.0.14 with bowtie 2\textsuperscript{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.

**Gene Expression Data Sets.** To assist in the manual annotation of genes related to toxicant stress, additional gene expression data sets were utilized, consisting of differentially expressed genes identified by microarray analysis following exposure to model pollutants (see Table 1). Two of these data sets were previously published\textsuperscript{30,44} while a third data set is new. Details related to these microarray experiments can be found in the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo; Accession number: GPL17458) and SI S1. Contigs corresponding to the differentially expressed microarray probes were aligned to the genome using Blastn within the Apollo genome browser. Genes were manually annotated in the areas of the genome where the contigs aligned using available MAKER and AUGUSTUS gene models and RNaseq reads to correct exon and intron boundaries if needed.

**RESULTS AND DISCUSSION**

**Description of the \textit{Hyalella azteca} Genome.** Analysis of the \textit{H. azteca} U.S. Lab strain genome size using flow cytometry gave an average genome size of 1.05 Gb (females 1C = 1045 ± 8.7 Mb; males 1C = 1061 ± 10.2 Mb). Our genome size is significantly smaller than recent estimates from \textit{H. azteca} representing different species groups collected throughout North America, which have been shown to vary by a factor of 2.\textsuperscript{45} Following the strategy of the i5k pilot project, we generated an assembly of 1.18 Gb (Hatz_1.0; SI Table S1). However, because of the high gap fraction and an assembly size larger than the experimentally determined genome size, a second assembly of 550.9 Mb was later generated from the same sequencing data using Redundans (Hatz_2.0).\textsuperscript{33} This assembly greatly improved contig N50 and produced a more complete gene set (Figure 2, SI Table S1). Because of the timing of the availability of Hatz_2.0, most of the manual annotations were performed using Hatz_1.0; therefore, both assemblies are presented here. RNaseq reads\textsuperscript{43} mapped equally well to both genome assemblies (Figure 2E), but significantly more reads mapped to the improved gene models of the Hatz_2.0 assembly (Figure 2F). BUSCO analysis\textsuperscript{46} was performed on both \textit{H. azteca} genome assemblies (Figure 2C) and predicted gene sets (Figure 2D) to assess the completeness of the genome. Hatz_2.0 contains a higher percentage (Genome = 91.0%, Gene set = 94.2%) of complete BUSCOs in contrast to the Hatz_1.0 assembly (Genome 85.3%, Gene set = 67.6%). When compared to other genomes of ecotoxicological relevance,\textsuperscript{47–53} the Hatz_2.0 showed comparable quality, ranking fourth in completeness out of the eight genomes assessed.

**Associated Bacteria and Lateral Gene Transfers.** Using two complementary approaches to screen the \textit{H. azteca} genome for bacterial contaminants,\textsuperscript{54,55} we recovered two draft bacterial genomes, or metagenome assembled genomes (MAGs), and evidence of lateral gene transfers (LGTs) (SI S2). MAG1 is a flavobacterium with distant affinities to currently sequenced bacteria (92% 16S rRNA identity to the most closely identified genera Chishiuella and Empedobacter). MAG2 is related to bacteria in the genus Ideonella (98% 16S rRNA identity to \textit{I. paludis}), of which some are able to degrade plastics,\textsuperscript{56} in particular an isolate from the wax moth \textit{Galleria mellonella}.\textsuperscript{57} Whether these bacteria are close associates of \textit{H. azteca} or components of their diet is not currently known, but their interesting gene repertoires could be relevant to \textit{H. azteca} ecotoxicology. An analysis of broad functional categories indicates that both genomes contain multiple genes related to metal detoxification and resistance to toxins (i.e., antibiotics) and possibly other organic pollutants (SI Figure S2.2). In addition, strong LGT candidates from Rickettsia-like bacteria were found on five genome scaffolds (SI Table S2.2).

**Genome Methylation and MicroRNAs.** We performed two genome-wide analyses to characterize DNA methylation patterns in the genome and characterize the full complement of \textit{H. azteca} microRNAs. DNA methylation is an epigenetic mechanism by which a methyl group (CH\textsubscript{3}) binds to DNA that may alter gene

---

**Table 1. Model Toxics and Exposure Concentrations for Microarray Gene Expression Analysis**

<table>
<thead>
<tr>
<th>pollutant class</th>
<th>chemical</th>
<th>end points</th>
<th>concentrations</th>
<th>references</th>
</tr>
</thead>
<tbody>
<tr>
<td>heavy metals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cadmium</td>
<td></td>
<td>5.5 µg/L</td>
<td>this study</td>
</tr>
<tr>
<td></td>
<td>zinc</td>
<td>1/10</td>
<td>25 µg/L</td>
<td>Poynton et al. (2013)\textsuperscript{44}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LC\textsubscript{10}</td>
<td>104 µg/L</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18 µg/L</td>
<td>Poynton et al. (2013)\textsuperscript{44}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LC\textsubscript{25}</td>
<td>70 µg/L</td>
<td></td>
</tr>
<tr>
<td>insecticide</td>
<td>cythuthin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NOEC</td>
<td>18 µg/L</td>
<td></td>
<td>Westen et al. (2013)\textsuperscript{30}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LC\textsubscript{90}</td>
<td>104 µg/L</td>
<td></td>
</tr>
<tr>
<td>nanomaterial</td>
<td>ZnO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1/10</td>
<td>18 µg/L</td>
<td>Poynton et al. (2013)\textsuperscript{44}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LC\textsubscript{25}</td>
<td>65 µg/L</td>
<td></td>
</tr>
<tr>
<td>organic pollutant</td>
<td>PCB126</td>
<td></td>
<td>7.0 µg/L</td>
<td>this study</td>
</tr>
</tbody>
</table>
Because methylated cytosines tend to mutate to thymines over evolutionary time, we used signatures of CpG dinucleotide depletion in the *H. azteca* genome to uncover putative patterns of DNA methylation. Our analyses showed a significant depletion of CpG dinucleotides in the *H. azteca* genome compared to other species, as illustrated in Figure 2. (A) Comparison of the contig, scaffold, and assembly size between the two assemblies. (B) Comparison of predicted gene sets from the two assemblies. The original MAKER gene set for Hazt_1.0 is Hazt_0.5.3 while the gene set developed by NCBI for Hazt_2.0 is referred to as Hazt_2.0. (C) BUSCO analysis compared to other genomes of ecotoxicological relevance including the amphipod *Parhyale hawaiensis*, the copepod *Eurytemora affinis*, *Daphnia pulex*, the aquatic midge *Chironomus tentans*, the terrestrial springtail *Folsomia candida*, the fathead minnow *Pimephales promelas*, and the killifish *Fundulus heteroclitus*. Dotted line corresponds to the total number of BUSCOs (single copy, duplicated, or fragmented) for the Hazt_2.0 assembly. (D) BUSCO comparison for the predicted gene sets. Dotted line corresponds to the total number of BUSCOs (single copy, duplicated, or fragmented) for the Hazt_2.0 gene set. (E) Percentage of RNA-seq reads that matched to the genome (left) and the predicted protein coding genes (right). Illumina data sets were acquitted from the NCBI Bioproject (PRJNA312414). Reads were mapped according to methods described in Rosendale et al. and Schoville et al.
that \textit{H. azteca} possesses strong indications of genomic DNA methylation, including CpG depletion of a subset of genes (SI Figure S3.1AB) and presence of the key DNA methyltransferase enzymes, DNMT1 and DNMT3 (SI Table S3.1). Furthermore, genes with lower levels of CpG observed/expected (CpG o/e; i.e., putatively methylated) displayed a strong positional bias in CpG depletion (SI Figure S3.1C, black line), with S regions of these genes being considerably depleted of CpGs. In contrast, genes with higher CpG o/e displayed no such positional bias (SI Figure S3.1C, gray line). Several insects where DNA methylation has been empirically profiled at the single-base level possess such patterns\textsuperscript{68} suggesting \textit{H. azteca} has similar patterns of DNA methylation as most insects (but see Glastad et al.\textsuperscript{66}).

MicroRNAs (miRNAs) are a family of short noncoding RNAs (~22 nt in length) that play critical roles in post-translational gene regulation.\textsuperscript{61} Recent research revealed that miRNAs are involved in aquatic crustaceans' response to environmental stressors (e.g., hypoxia and cadmium exposure),\textsuperscript{62,63} which makes miRNAs promising biomarkers for future aquatic toxicological research. We predicted \textit{H. azteca} miRNAs based on sequence homology and hairpin structure identification. A total of 1261 candidate miRNA coding sites were identified by BLAST. After hairpin structure identification, we predicted 148 \textit{H. azteca} miRNAs, which include several highly conserved miRNAs (e.g., miR-9 and let-7 family) (SI Table S3.2, sequences available in SI S6.1). Several Cd-responsive miRNAs in \textit{D. pulex} (miR-210, miR-71 and miR-252)\textsuperscript{62} were also predicted in \textit{H. azteca}, suggesting a conserved role of these miRNAs. This number of predicted miRNAs is comparable to what has been reported for other arthropods (SI Figure S3.2).

**Functional Annotation.** The manual annotation of the \textit{H. azteca} U.S. Lab strain genome resulted in the characterization of 13 different gene families (see detailed annotation reports in SI S4.1-S4.13). Given its importance in ecological and ecotoxicological studies, a particular focus was given to genes involved in environmental sensing (chemoreceptors and opsins), detoxification and response to stress (cytochrome P450s, cuticle proteins, glutathione peroxidases, glutathione S-transferases, heat shock proteins, and metallothionein proteins), as well as genes involved in important toxicological pathways (ion transporters, early development genes, insecticide target genes, and nuclear receptors). Here we highlight significant findings of gene family expansions or contractions as well as the characterization of genes of particular toxicological importance.

**Environmental Sensing.** To better understand how \textit{H. azteca} interacts with its environment, we annotated genes involved in light and chemical sensing. Arthropods deploy a diversity of light sensing mechanisms including light capture in photoreceptor cells through the expression of opsins: light-sensitive, seven-transmembrane G-protein coupled opsin receptor proteins. Our survey of the \textit{H. azteca} genome revealed only three opsin genes, a middle wavelength-sensitive subfamily 1 (MWS1) opsin and two belonging to the long wavelength-sensitive subfamily (LWS) opsins (SI Figure S4.13.1, Table S4.13.1). Maximum likelihood analysis with a subset of closely related malacostracan LWS opsins moderately supports the two \textit{H. azteca} LWS opsins as 1:1 orthologs of the two LWS opsins previously reported for \textit{Gammarus minus}.\textsuperscript{64} Thus, the LWS opsin duplicate pair conserved in \textit{H. azteca} and \textit{G. minus} is likely ancient, predating at least the origin of amphipod Crustacea. The \textit{H. azteca} MWS opsin by contrast, represents the first reported amphipod MWS opsin and is distinct from currently known malacostracan MWS opsins. These findings suggest that amphipod crustaceans are equipped with a minimally diversified set of three opsin genes and implies gene family losses for all of the nonretrin opsin subfamilies. This is in contrast to the 46 opsin genes characterized in \textit{Daphnia pulex}.\textsuperscript{47} It remains to be seen whether these candidate gene losses are associated with the adaptation of \textit{H. azteca} to its crepuscular visual ecology or reflect a more ancient trend in amphipods.

In contrast, the \textit{H. azteca} genome reveals gene expansions of chemoreceptors, which may be essential for \textit{H. azteca} given its epibenthic ecology and close association with sediments.\textsuperscript{6} Non-insect arthropods have two major families of chemoreceptors: gustatory receptor (GR) family, an ancient lineage extending back to early animals,\textsuperscript{49,65-67} and the ionotropic receptors (IRs) that are a variant lineage of the ancient ionotropic glutamate receptor superfamily known only from protostomes.\textsuperscript{68,69} These two gene families were manually annotated in the \textit{H. azteca} genome and improved models were generated for two other crustaceans, \textit{D. pulex} and \textit{Eurytemora affinis}\textsuperscript{69,71} for comparison (SI S4.1, sequences available in SI S6.2). With 155 GR genes, \textit{H. azteca} has over twice the number of GRs compared with \textit{D. pulex} (59)\textsuperscript{69} and \textit{E. affinis} (67), although many of the most recent gene duplicates are pseudogenes. Two candidate GR sugar receptors were identified in \textit{H. azteca} and \textit{D. pulex} (independently duplicated in both lineages), but not in \textit{E. affinis}. Otherwise these crustacean GRs form large species-specific expanded clades with no convincing orthology with each other or other conserved insect GRs such as the Gr43a fructose receptor (SI Figure S4.1.1). \textit{H. azteca} has 118 IR genes (two pseudogenic) compared with updated totals of 154 in \textit{D. pulex} (26 pseudogenic) and 22 intact genes for \textit{E. affinis}. All three species contain single copy orthologs of the highly conserved IR genes implicated in perception of salt, amines, amino acids, humidity, and temperature in insects, including Ir25a, Ir8a (missing in \textit{D. pulex}), Ir76b,\textsuperscript{72} and Ir93a.\textsuperscript{73} The remaining divergent IRs form largely species-specific expanded lineages (SI Figure S4.1.2). The many divergent IRs and GRs in these three crustaceans presumably mediate most of their chemical sense capability, but their great divergence from the proteins of \textit{Drosophila} for which functions are known preclude speculation as to specific roles.

**Cytochrome P450s.** 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.\textsuperscript{74} P450 genes function in metabolizing a wide range of endogenous and exogenous compounds, including toxins, drugs, plant metabolites, and signaling molecules.\textsuperscript{75-78} In the \textit{H. azteca} genome, we found 70 genes or gene fragments that contained a typical P450 signature (FxxGxxxC), where C is the heme thiolate ligand (SI S4.3, sequences available in SI S6.2). However, only 27 were complete genes. The 70 P450 genes were classifiable into one of four recognized P450 clans, with the CYP2 clan (SI Figure S4.3.1) being the largest with 48 genes. The most notable difference between the P450 complement of \textit{H. azteca} relative to hexapods (insects) was the expansion of the CYP2 clan P450s. Typical of expanded clades, we found several clusters of genes (and gene fragments) of the CYP2 clan. The CYP3 and CYP4 clans in \textit{H. azteca} were represented by eight and seven genes, respectively. The fourth P450 clan is the mitochondrial P450, with at least nine genes in \textit{H. azteca}. The number of P450s found in \textit{H. azteca} was greater to those found in other crustaceans, including the copepods \textit{Tigriopus japonicus} (52)\textsuperscript{79} and \textit{Paracyclopina nana} (46),\textsuperscript{80} but somewhat fewer than those
in hexapod taxa (including 106 in the mosquito Anopheles gambiae and 81 in the silkworm Bombyx mori).81

**Heat Shock Proteins.** The heat shock protein (HSP) molecular chaperones are a highly conserved family of proteins that facilitate the refolding of denatured proteins following stress, including thermal stress, but also in response to metals and other toxicants, oxidative stress, and dehydration.82 HSPs are 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/J-protein is a cofactor to HSP70 and delivers nonnative proteins to HSP70.83 HSPs were identified and annotated for each of these families (SI S4.8). The number of hsp70 (8 genes), hsp90 (3), hsp40 (3), and hsp60 (1) was well within the expected number found throughout Arthropoda.84 Of the eight hsp70 genes, five were found as a gene cluster on scaffold 277, which is similar to gene clusters identified in Drosophila melanogaster85 and Aedes aegypti.86 In agreement with Baringou et al.,87 the HSP70 proteins described here cannot be easily divided into inducible and cognate forms based on sequence characteristics. We instead decided to compare our eight sequences to sequence motifs described by Baringou et al. and classify the H. azteca HSP70s according to their framework (SI 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.87 finding that all amphipod HSP70s characterized to date are Group A proteins. One HSP70 contained slightly different motif characteristics and was grouped with A4 proteins, while the remaining sequences were grouped together in A5.

**Metallothionein Genes.** 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. Although their diversity makes it difficult to assign a specific function by class of MTs, their ability to bind metal ions has provided MTs with a role in detoxification, binding, and sequestration of toxic metals.88 Four mt genes were identified in the H. azteca genome by mapping Cd responsive contigs with homology to Callinectes sapidus CdtM-1 (AAF08964) to the HAZT_1.0 assembly (SI S4.11). These four genes were arranged as repeats on scaffold 460 and each contained three exons, the typical gene structure of mts (SI Figure S4.11.1). Mt-b and mt-d produce identical proteins of 61 amino acids, whereas mt-c is missing the downstream splice site on exon 1 and produces a truncated protein of 53 amino acids. Mt-a lacks a viable start codon, making it a likely pseudogene. Due to the similarity in the sequences of the remaining three genes, it is not possible to determine if they are all transcribed or regulated differently based solely on the RNaseq mapped reads. However, given their high degree of similarity and arrangement on scaffold 460, these genes are likely the result of recent gene duplications, which may provide an evolutionary advantage against high metal exposure. 1–4 mt genes have been identified in at least 35 other Malacostracan species. However, in most cases, the multiple MTs are not identical in amino acid sequence. For example, in the blue crab C. sapidus there are three mt genes. Two encode for Cd inducible forms with 76% sequence identity, while a third, codes for a longer copper-inducible form.89 Given that our strategy for identifying the mt genes in H. azteca relied on using the Cd inducible gene expression set, it is possible that a fourth, copper inducible form also exists in its genome.

**Ion Transport Proteins.** In arthropods, a subset of ion transporters are integral in maintaining cellular homeostasis and regulating epithelial transport of common ions such as H+, Na+, K+, and Cl−90−93 and are likely involved in the toxicity and uptake of metal ions. The proton pump V-type H+ ATPase (VHA, ATP6) is an evolutionarily conserved molecular machine having a wide range of functions. VHA actively translocates H+ across the membranes of cells and organelles allowing it to generate electrochemical H+ gradients that drive H+-coupled substrate transport of common bioavailable cations (Na+, K+, Li+).95 VHA is a large, two domain protein complex (V1 and V0) comprised of 13 subunits, which are ubiquitous in eukaryotes and thought to be expressed in virtually every eukaryotic cell.96 These 13 VHA subunits were identified in the H. azteca genome, but two accessory subunits were not (SI S4.10). A 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).97 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 sodium/hydrogen antiporters (NHA, SLC9B2, CPA2) are a subfamily of transmembrane ion transporters, which was only recently discovered in animal genomes and characterized in mosquito larvae.98−100 In both arthropods and mammals, evidence indicates that NHA is coupled to VHA as a secondary electroneutral transporter for ion uptake against concentration gradients.99,101−104 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 (SI S4.10).98

The minimal set of VHA subunits and expansion of NHA genes may have toxicological significance, particularly with respect to metal toxicity and transport. Metal speciation, and therefore toxicity, is highly pH-dependent.105 As a regulator of pH at the epithelial membrane, and thus electrochemical transmembrane gradients, VHA may play an important role in metal uptake, ion speciation, and solubility.106−108 The specific substrates for NHA have not yet been fully characterized for most species, but sodium and chloride are likely candidates.109 The presence of NHA gene duplicates in the H. azteca genome suggest a likely adaptation for ion uptake against transmembrane concentration gradients, but may also influence metal uptake and bioavailability (see SI S4.10 for further discussion). As many toxic metal ions are transported across the cell membrane via ion channels and/or ion transporters (e.g., ATPases),110 additional transporters should be explored for their direct roles in metal uptake and indirect influence on metal toxicity.

**The Toxico-Responsive Genome.** Contaminants such as heavy metals, organic compounds, and nanoparticles can adversely affect ecologically relevant organisms. In H. azteca, heavy metal contaminants (Zn, Cd) have been shown to have negative effects on the development of population growth rate, longevity, and reproduction112−114 while metal-based nanomaterials (i.e., ZnO NMs) cause increased toxicity which may be related to enhanced bioavailability.94,115 Organic compounds such as pyrethroid insecticides and polychlorinated biphenyls (PCBs) can cause detrimental effects on behavior, reproduction, and development.116−118 A primary goal of the H. azteca genome project was to expand the functional gene annotations for transcripts that respond to toxicant stress, referred to as the toxicogenome. We utilized two published gene expression studies (Zn, ZnO, NPs, and cyfluthrin)119 to
as well as two unpublished gene expression sets for Cd and PCB exposure (Table 1). During our original investigation, we were only able to annotate a small fraction of the differentially expressed transcripts (Cd: 13%; Zn: 29%; PCB126: 0%; Cylfluthrin: 20%) (Figure 3A). The ability to align these transcripts to the *H. azteca* genome allowed us to identify full length transcripts and more completely assemble transcripts aligning to the same genic region of the genome in a way that was not possible with a de novo transcriptome assembler (i.e., Newbler, Roche). This increased our ability to predict gene function increasing the fraction of annotated transcripts by 10−32% (Figure 3A). However, we also note that for each of these chemical challenges, over half of the

Figure 3. Annotation of toxicant responsive genes. (A) The annotation of differentially expressed transcripts was significantly improved using the *H. azteca* genome. The total number of unique differentially expressed transcripts is listed below each treatment. Pie graphs represent differentially expressed contigs and illustrate the percentage of original annotations (blue) and the additional annotations that were added when contigs were aligned to the genome (orange). For PCB126, none of the contigs were annotated prior to alignment to the genome. In many cases more than one contig aligned to the same transcript; therefore, the total number of contigs is greater than the number of transcripts. (B−D) Biological processes gene ontology (GO) terms representing the differentially expressed transcripts from Cd (B), Zn, and ZnO NPs (C), and cyfluthrin (D). 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, none of the 12 annotated transcripts were mapped to biological processes GO terms. Similar graphs for molecular function GO terms can be found in the SI Figure S5.
genes are still without annotations (Figure 3A), implying that these genes may be lineage specific. This is similar to the finding within the *D. pulex* genome that lineage specific genes were more likely to be differentially expressed following environmental challenges.

**Cadmium.** To explore the gene expression response and further annotate genes involved in heavy metal exposure, we conducted a gene expression study at ecologically relevant concentrations of Cd. Compared to controls, 116 genes were up-regulated in expression and 9 were down-regulated by Cd. These genes are related to several cell processes including digestion, oxygen transport, cuticular metabolism, immune function, acid-base balance, visual-sensory perception and signal transduction (SI Table S5.1). Categorizing the genes by biological processes illustrated that the metabolism of *H. azteca* was very broadly affected, with the cellular metabolic processes representing the largest GO term (Figure 3B). Heat shock proteins (general stress response) were significantly upregulated in response to Cd (see SI Table S4.8.1) consistent with other amphipod studies and showing a similar response to other stressors including heat stress, oxidative stress and changes in pH.

In addition, expression of the newly described MT genes were also significantly induced over 15-fold by Cd (SI Figure S4.11.3). Cd exposure also induced expression of genes involved in oxidative stress including glutathione-S-transferase, a commonly used biomarker in toxicity tests of pollutant exposure and oxidative stress, and thioredoxin peroxidase, a gene involved in protection against reactive oxygen species (ROS). Finally, genes involved in regulation of the cuticle were also upregulated. Differential expression of Chitinase and other cuticular proteins has been demonstrated previously in crustaceans in response to stress and has been correlated with impacts to growth and reproduction.

**Zinc.** We utilized a data set originally published in Poynton et al. that compared the toxicity of ZnO NPs to zinc sulfate (*ZnSO₄*) to increase the number of annotated genes that were responsive to metal exposure. Of the 60 differentially expressed genes, we annotated 25, including 15 genes that had not been annotated in the original publication (SI Table S5.2, Figure 3A). For example, choriorn peroxidase (*contig18799* in Poynton et al.) was induced in both the *ZnSO₄* and ZnO NP exposures and acts as an indicator of oxidative stress, as the gene is involved in ROS damage repair. *Contig000192* in Poynton et al. is another previously uncharacterized gene that was annotated as asparaginyl beta-hydroxylase-like protein, a regulator of muscle contraction and relaxation; its dysregulation suggests negative impacts to swimming behavior and movement. With the additional annotation results we were able to perform gene ontology analysis (SI Table S5.2) and observed that most genes were mapped to GO:0042221, response to chemical (Figure 3C).

**Cyfluthrin.** The pyrethroid insecticide cyfluthrin is one of the most widely applied insecticides worldwide and has been shown to be highly toxic to *H. azteca*. (EC₅₀ < 1 ng/L). We previously showed that cyfluthrin exposure at 1 ng/L caused differential expression of 127 sequences. Through the reanalysis of this data set, we were able to annotate 33 genes and successfully mapped them to GO terms (SI Table S5.3). Many affected genes were consistent with the known mechanism of pyrethroid toxicity, showing involvement in neurological system processes, synapse organization, and transmission of nerve impulses, but also stress response such as oxidation processes, damage repair, maintaining of homeostasis, and immune response (Figure 3D).

**PCB126.** For *H. azteca*, PCBs represent a major exposure and accumulation threat due to their habitat and feeding behavior in benthic areas. In fishes, dioxin-like PCBs (e.g., PCB126) are highly toxic as they bind to the aryl hydrocarbon receptor and induce the expression of CYP1 genes. Much less is known about this mechanism in crustaceans, but in general they appear more tolerant of PCBs. Following exposure of *H. azteca* to PCB126, the most potent and ubiquitous of the PCB congeners, we identified 21 differentially expressed sequences, representing seven genes, of which five were annotated (SI Table S5.4). Three of the five characterized genes are transmembrane proteins, while two are involved in endocrine processes (growth hormone, thyroid hormone). Neuroendocrine disruption of PCBs was described in crustaceans previously (see review in); however, most investigations on neuroendocrine disruption to date focus on effects in vertebrates. Our study demonstrates that potential impacts of neuroendocrine disruption on invertebrates deserves further attention.

In summary, with a total of 19 936 genes including 911 manually curated genes, the genome of the *Hyalella azteca* U.S. Lab strain provides a foundational tool for understanding the molecular ecology of benthic invertebrates as well as the mechanisms of toxicity of sediment associated pollutants. The critical gene families annotated here will serve as basis for studying toxicologically conserved pathways in invertebrates and developing adverse outcome pathways for sediment dwelling organisms. Overall, our results illustrate the advantage of applying a genome assembly to ecotoxicogenomic studies including the improved ability to annotate genes of interest and we strongly encourage the expansion of genomic, not just transcriptomic, resources for other species of ecotoxicological relevance.

The ever-growing list of chemical contaminants entering the environment poses a significant challenge in terms of risk assessment. The low-throughput and high cost of traditional toxicity testing, suggests that the need for alternative means to assess risk. The characterization of omics responses has emerged as a potential alternative. Measures on the cellular level provide valuable information on the mode of action of uncharacterized chemicals, the health status of exposed organisms, and can act as a means to extrapolate beyond model organisms, and can be integrated into predictive risk models. The interpretation of these omics responses within the context of the well-defined *H. azteca* genome described herein will greatly expand the utility and applications of omics responses to sediment ecotoxicology and risk assessment.

**ASSOCIATED CONTENT**

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.8b00837.

Additional method details (S1), detailed annotation reports (S2–S4), gene expression data tables and figures (S5), supplemental sequence files (S6) and detailed author contributions (S7) (PDF) (XLSX) (TXT)

**AUTHOR INFORMATION**

Corresponding Author

E-mail: helen.poynton@umb.edu.
Environmental Science & Technology

ORCID
Helen C. Poynton: 0000-0003-2663-9841
Maria S. Sepulveda: 0000-0003-0473-5760

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the staff at the Baylor College of Medicine Human Genome Sequencing Center for their contributions, Terence Murphy at NCBI for assisting in screening for bacterial contamination in the assembly and with removing contaminants from the genome, and Dawn Oliver at Purdue University for assistance with the distribution map (Figure 1). Additional thanks to Mitch Kostich of the U.S. EPA for his thoughtful comments on the manuscript. Funding for genome sequencing, assembly and automated annotation was provided by NHGRI grant U54 HG003273 to RAG. Funding for transcriptome sequencing was provided by NSF grant CBET-1437409 to HCP. Additional funding was provided by the State and Federal Contractors Water Agency (contract no. 15-14) to Richard Connon, UC Davis (supported S.H.), University of Cincinnati Faculty Development Research grant and partial support from NSF grant DEB165441 to JBB, and NSF grant DEB1257053 to J.H.W. M.M.T.’s work with Apollo was supported by NIH grants SR01GM080203 from NIGMS, and SR01HG004483 from NHGRI, and by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

REFERENCES

ecological hazard assessments. *Environ. Health Perspect.* 2013, 121 (9), 1002.
(35) 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.


