Spring 2011

The hidden transcriptome: discovery of novel, stress-responsive transcription in Daphnia pulex

Stephen Butcher
University of Iowa

Copyright 2011 Stephen Butcher

This thesis is available at Iowa Research Online: https://ir.uiowa.edu/etd/929

Recommended Citation
https://doi.org/10.17077/etd.g564j2de

Follow this and additional works at: https://ir.uiowa.edu/etd

Part of the Biology Commons
THE HIDDEN TRANSCRIPTOME: DISCOVERY OF NOVEL, STRESS-RESPONSIVE TRANSCRIPTION IN *DAPHNIA PULEX*

by

Stephen Butcher

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biology in the Graduate College of The University of Iowa

May 2011

Thesis Supervisor: Assistant Professor John Manak
This is to certify that the Master’s thesis of

Stephen Butcher

has been approved by the Examining Committee for the thesis requirement for the Master of Science degree in Biology at the May 2011 graduation.

Thesis Committee: ________________________________

John Manak, Thesis Supervisor

_________________________________________

Daniel Eberl

_________________________________________

Josep Comeron
# TABLE OF CONTENTS

## LIST OF TABLES  
iv  
## LIST OF FIGURES  
v  
## CHAPTER  

### I. ECOTOXICOLOGY AND ENVIRONMENTAL STRESSORS  
1  
- Environmental stressors 1  
- Techniques for identification of stress response genes 2  
- *Daphnia pulex*: An ecotoxicological model organism 2  
- Known stress responsive gene groups 3  
- Known stress responsive genes in *Daphnia pulex* 5  
- *Daphnia pulex* genome assembly and gene annotation 6  
- Transcriptional analysis of the stress response 7  

### II. RESULTS OF TILED MICROARRAY AND RNA-SEQUENCING EXPERIMENTS  
8  
- Microarray analysis of the *D. pulex* transcriptome 8  
- Validating and updating gene models 12  
- SOLiD whole transcriptome RNA sequencing 13  
- Analysis of *Daphnia* homologous stress responsive genes 13  
- Gene ontology (GO) analysis 16  
- Identification of uncharacterized DE genes 25  
- Gene expression analysis with RNA sequencing 26  
- DE analysis of genes induced by individual metals 30  
- Analysis of uncharacterized stress induced coding genes 30  
- RNA sequencing analysis of predicted non coding genes 35  

### III. DISCUSSION OF THE DATA ANALYSIS  
40  
- *Daphnia pulex* genome annotation 40  
- Identified stress responsive transcription 41  
- Conclusions 42  

### IV. MATERIALS AND METHODS  
44  
- *Daphnia pulex* predator exposure and RNA isolation 44  
- *Daphnia pulex* metal exposure and RNA isolation 44  
- *Daphnia pulex* whole genome tiling array 45  
- Processing of RNA samples for array hybridization 45  
- Microarray data analysis 46  
- RT-PCR validation of microarray data 46  
- RACE validation of microarray data 47  
- PCR product analysis 48
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO analysis of differentially expressed genes</td>
<td>48</td>
</tr>
<tr>
<td>cDNA library preparation for SOLiD RNA sequencing</td>
<td>48</td>
</tr>
<tr>
<td>Barcoding of individual cDNA libraries</td>
<td>50</td>
</tr>
<tr>
<td>Emulsion PCR and sequencing of cDNA libraries</td>
<td>50</td>
</tr>
<tr>
<td>Analysis of SOLiD RNA sequencing reads</td>
<td>51</td>
</tr>
<tr>
<td>Viewing of mapped RNA sequencing reads</td>
<td>51</td>
</tr>
</tbody>
</table>

REFERENCES                                                   | 53   |
LIST OF TABLES

Table 1. RT-PCR and RACE validation of novel transcription identified in the Chaoborus exposed microarray data 11

Table 2. Comparison of JGI gene annotation to the most current 2010 Daphnia pulex gene annotation 12

Table 3. Microarray evidence for stress responsive genes 22

Table 4. GO analysis of heavy metal stress responsive genes 24

Table 5. GO analysis of Chaoborus predation stress responsive genes 25

Table 6. Differentially expressed genes identified by microarray analysis and confirmed by RNA sequencing 29

Table 7. RNA sequencing transcriptional analysis of each individual metal identifies metal stressor specific gene induction 31
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Microarray data tracks loaded into SignalMap identify equally expressed regions of transcription corresponding to gene predictions.</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>RT-PCR and RACE validate novel transcription identified in the microarray data set</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Analysis of microarray data and RNA sequencing validates gene predictions</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>Predation stress induced cuticular protein identified in microarray data set and confirmed in RNA sequencing</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>Predation stress induced uncharacterized protein identified in microarray data set and confirmed in RNA sequencing</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>Metallothionein genes identified in the RNA sequencing data set as upregulated greater than two fold across every metal treatment</td>
<td>19</td>
</tr>
<tr>
<td>7</td>
<td>An oxidative stress induced growth inhibitor identified as metal stress induced in the microarray data and confirmed with RNA sequencing</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>A transcription initiation factor gene identified as metal stress induced in microarray data and confirmed with RNA sequencing</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>A serine-like protease gene repressed in under both stress conditions</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>Uncharacterized protein coding gene induced under metal stress and confirmed with RNA sequencing</td>
<td>27</td>
</tr>
<tr>
<td>11</td>
<td>Uncharacterized gene with low protein coding potential induced under metal stress and confirmed with RNA sequencing</td>
<td>28</td>
</tr>
<tr>
<td>12</td>
<td>Uncharacterized differentially expressed gene identified as being metal stress responsive</td>
<td>32</td>
</tr>
<tr>
<td>13</td>
<td>Manual gene annotation with paired end RNA sequencing allows for prediction of splice consensus sites</td>
<td>33</td>
</tr>
</tbody>
</table>
Figure 14.  Manually annotate nucleotide sequence assembled from paired end RNA sequencing confirms an uncharacterized gene with no known homology that is induced under metal stress 34

Figure 15.  Uncharacterized differentially expressed gene with low coding potential identified as being metal stress responsive 36

Figure 16.  Manually annotate nucleotide sequence assembled from paired end RNA sequencing confirms an uncharacterized gene with no known homology that only contains very small potential open reading frames. This stress metal stress induced gene may be a putative non-coding RNA. 37

Figure 17.  A region of differential expression manually assembled using RNA sequencing data confirms low protein coding potential. 38

Figure 18.  A differentially expressed region of low coding potential with a manually assembled nucleotide sequence indicates only small ORF potential. 39
CHAPTER I

ECOTOXICOLOGY AND ENVIRONMENTAL STRESSORS

Environmental stressors

Ecotoxicology studies have focused on the effect that environmental toxins have on an ecosystem and how organisms within the ecosystem respond to these toxins. Some organisms are extremely sensitive to changes in their environment and have developed mechanisms for adapting to different environmental stressors (1). One important area of ecotoxicology is focused on aquatic ecosystems. Toxins that leach into a water supply have far reaching effects from the smallest organisms, to larger animals and humans that are dependent on the water source for life. There are different types of aquatic stressors ranging from naturally occurring toxins, to toxins that are introduced to an environment by humans. Excessive bacterial growth, low availability of nutrients, hypoxia, and predation can all stress organisms in an aquatic environment (1). It is the introduction of toxins from human activities that can be the most devastating to an ecosystem. These toxins are introduced into the environment from different sources including pesticides, industrial factories, mines, and landfills. Heavy metal toxins are a common pollutant from these sources, and are a common pollutant found within aquatic environments. Even exposure to low levels of heavy metals toxins can be very detrimental to the health and viability of exposed organisms (2). Early identification of the environmental damage from pollutants like heavy metal toxins is the key to removal of the polluting toxins before serious damage is done to the affected organisms.
Techniques for identification of stress response genes

Determining the underlying mechanisms that allow an organism to adapt and survive environmental stress has been of recent interest, as it opens avenues for developing a tool for early indication of environmental pollution (1, 2, 3, 4). The availability of genomics tools, like whole genome microarrays and the development of new techniques like next generation sequencing, has advanced transcriptome studies (5, 6). These transcriptomics techniques have traditionally been used for the empirical analysis of transcription at specific developmental stages in an effort to validate or improve current gene models. This technology can also be utilized in toxicogenomics for the identification and quantification of stress responsive transcription, and can be used for the characterization of novel, stress responsive transcription (7). The incorporation of these tools into ecotoxicological studies can be applied to any organism of interest. This is an important first step to defining the genetic mechanism behind an organism’s stress response, and eventually developing this as a tool for detecting the presence of environmental stressors.

Daphnia pulex: An ecotoxicological model organism

*Daphnia pulex*, the common water flea, is a microcrustacean that is important in fresh water ecosystems as it makes up a large part of the herbivorous zooplankton biomass (8). Early aquatic toxicology studies focused on *Daphnia* because it is found in virtually all fresh water ecosystems, and has been shown to be sensitive to the presence of toxins in its environment (9). *Daphnia* are able to respond quickly to environmental stressors which make it an ideal organism for development as a tool for the early detection of aquatic pollutants (10). These organisms have a large brood sizes with short
generation time, and reproduce asexually, which allows for clonal populations to be easily cultured for experimental analysis (11). They are also easily collected and manipulated in field studies, making them useful for onsite water quality assessments (9, 12). *Daphnia* shows a definable defense response to predation stress brought on by the presence of the phantom midge larvae *Chaoborus*. This is well characterized by the development of elongated neck teeth as a defense against this predator (13). The regulation of this morphological response to predation by *Chaoborus* is of interest to researchers as it provided some of the first phenotypic evidence of an induced stress response in *Daphnia*. The *Chaoborus* stress response is a good example of how *Daphnia* is sensitive to environmental stress and has mechanisms in place to respond to them. Characterizing the regulation of the *Daphnia* stress response could be developed into an early indicator of fresh water pollution, as well as a model system for further characterization of the eukaryotic stress response (8, 9, 14). Further development of this tool requires a complete understanding of how *Daphnia* gene expression is regulated in response to specific toxins. Using tiled microarrays and next generation sequencing techniques, it is possible to identify transcriptional changes in response to environmental stress (6, 15).

**Known stress responsive gene groups**

Stress responsive transcription has been identified in other organisms. Protein coding genes like heatshock genes and cytochrome genes are part of a general stress responsive set of genes that are regulated in response to stressors in the environment (4, 16). More recently, it has been shown that some non-coding RNAs are also stress responsive. These stress responsive ncRNAs have been identified in many organisms
ranging from bacteria, to plants, to humans, and have been confirmed as having a regulatory role in the stress response (16, 17). *Escherichia coli* responds to low iron stress by up regulating the expression of non-coding RNAs that target iron binding protein genes for degradation in response to the low amounts of iron in the environment (18). Bacterial ncRNAs have been defined as playing a key role in bacterial response to other kinds of stressors including oxidative stress, osmotic stress, and nutrient limitation (17, 18, 19). Analysis of yeast transcriptomes under different environmental conditions indicated that greater than 90% of the yeast genome is transcribed, and identified a subset of these identified transcripts as non-coding (20). Moreover, it has been shown that ncRNAs are also involved in regulating the stress response in plants. microRNAs have been shown to regulate gene expression in response to several environmental stressors including oxidative stress, bacterial infection, phosphate starvation, drought, and heavy metal stress (21, 22, 23, 24, 25). Long non-coding RNAs have also been characterized as playing a role in regulating the stress response in *Arabidopsis thaliana* during drought conditions (26). ncRNAs have also been identified in humans and have been shown to play a role in development, epigenetic programming, and diseases (27, 28). The ncRNA pathway allows for a quick response to environmental stressors by targeting transcripts and/or transcriptional machinery to directly regulate gene expression, and could be a common mechanism for stress response regulation. It is possible that the *Daphnia pulex* stress response is also dependent on ncRNAs to quickly regulate stress-specific transcription in response to environmental stressors. Empirical analysis of the *Daphnia* transcriptome under stress and non-stressed conditions will not only define protein
coding genes that are stress responsive, but will also identify any ncRNA genes that are also stress responsive (15).

**Known stress responsive genes in *Daphnia pulex***

Researchers have been attempting to identify stress responsive genes in *Daphnia pulex* using a candidate gene approach. A study by Shaw et al. in 2006 was aimed at defining the *Daphnia pulex* response to cadmium stress using gene expression microarrays. These arrays were not representative of the entire *Daphnia pulex* transcriptome as the Daphnia genome sequencing project had not been completed. The gene probes for the microarray were developed using known *Daphnia* EST sequence data. The arrays were used to compare changes in gene expression from cadmium treated *Daphnia* to a non-cadmium treated control (4). After 48 hours of sub-lethal cadmium exposure, genes that were identified as being involved in storage of lipids and for the molting process showed changes in expression levels when compared to untreated *Daphnia*. This is in line with the decrease in size, lipid storage, and reproductive success that is seen in *Daphnia* populations that are exposed to sub-lethal amounts of cadmium for 21 days. This indicates that *Daphnia* respond quickly to environmental toxins, and this can be measured by changes in gene expression before any morphological changes are visible. Other genes identified as being responsive to cadmium exposures included metabolism, calcium binding genes, and uncharacterized genes. This approach also revealed previously uncharacterized metallothioneins in *Daphnia* that were increased in response to cadmium exposure. The authors concluded that they were able to identify *Daphnia* genes that are responding to cadmium toxicity after only twenty four hours of exposure to cadmium, indicating that this type of approach could be used as an early
detection tool of environmental stressors. It was also noted that a more complete analysis of the *Daphnia pulex* transcriptome would be necessary to completely characterize the transcriptional response of *Daphnia* to cadmium or to any other stressors.

**Daphnia pulex genome assembly and gene annotation**

The *Daphnia* Genomics Consortium (DGC) and the Joint Genome Institute (JGI) have worked in collaboration to sequence the *Daphnia pulex* genome and the genome assembly and minimum protein coding gene set have been published. The 200mb genome is currently assembled into 19,008 contigs in 5191 scaffolds. Current transcriptome analysis indicates that there is a minimum of 30,907 protein coding genes in this genome (8). Completion of the genome sequencing project made it possible to use techniques like tiled microarrays and RNA sequencing to empirically analyze *Daphnia pulex* transcription at specific developmental stages, as well as under specific stressed conditions. This provides an avenue to not only validate and improve current gene models in *Daphnia*, but also indicates transcribed regions in the genome that are responsive to specific stressors. Microarray analysis of stress responsive transcription can be used to characterize annotated genes with unknown function and to identify previously unannotated regions of the genome. The incorporation of RNA sequencing provides further validation of the microarray data and provides the exact sequence of transcribed regions that can be used for further analysis of uncharacterized genes. The combination of the two data sets defines for the first time a complete catalogue of stress responsive transcription in *Daphnia pulex*. 
Transcriptional analysis of the stress response

*Daphnia pulex* is a well defined ecotoxicology model organism that is known to be highly sensitive to changes in its environment (8, 9). It is becoming clear that environmental stressors cause transcriptional changes that can induce a defense response to the stressor (16, 17). In this study, tiled microarrays and RNA sequencing were used to empirically analyze regions of transcription in *Daphnia pulex* to both improve the current gene annotations and to define stress responsive transcription due to predation or heavy metal stress. The transcriptional profiles of the stressed animals reveals stress responsive transcription that corresponds to both annotated gene regions and previously unannotated regions of the genome. Some of the stress responsive genes identified have no known function or homology with other organisms and are likely uncharacterized stress response genes. These uncharacterized genes would likely never be revealed using traditional transcriptome methods where organisms are studied in optimal laboratory conditions. Transcriptome analysis of an organism grown under conditions that are representative of the stresses that are present in their natural environment, allows for the characterizing of a hidden transcriptome that is only revealed by exposure to environmental stressors.
CHAPTER II

RESULTS OF TILED MICROARRAY AND RNA-SEQUENCING EXPERIMENTS

**Microarray analysis of the D. pulex transcriptome**

Transcriptome analysis was performed using co-hybridizations of cDNA from stressed animals with labeled cDNA from non-stressed control hybridized to NimbleGen whole genome tiling microarrays designed for *Daphnia pulex*. This technique empirically interrogates the entire, non-repeat genome, which allows for regions of active transcription to be identified within populations of *Daphnia* collected at specific life stages, and grown in stressed or non-stressed conditions. This experiment was designed to both improve current gene annotations as well as to define regions of the genome that were transcribed or repressed in response to stress. Six different populations of *Daphnia pulex* were exposed to specific stressors. One population was exposed to stress from *Chaoborus* predation, and the five other populations were exposed to heavy metal stress from one of the following metals: Cadmium, Arsenic, Copper, Nickel, or Zinc. Animals were collected after exposure to *Chaoborus* predation. This chronic exposure was performed over the entire life of the animal until collection at the 1st instar juvenile stage. All of the collected animals had developed neckteeth, a visible morphological change that is a well characterized inducible defense to predation. Adult (17-24d) animals were collected after an acute exposure of 24 hours to heavy metals. This acute exposure was designed to capture some of the immediate transcriptional changes that occurred in response to heavy metal stress. Similarly aged non-stressed control populations were also collected for comparison of stressed versus non-stressed responsive transcription. Double stranded cDNA libraries were generated from animals collected from each stress
exposure and from non-stressed control populations. The individual metal exposed cDNA samples were pooled together to make up a mixed metals population that represented all of the transcription from the heavy metal exposures. This mixed metal population was co-hybridized with a non-metal exposed control cDNA sample. The *Chaoborus* exposed and non-*Chaoborus* exposed cDNA was also co-hybridized to a tiling microarray. Three biological replicate cDNA libraries were generated from stressed and non-stressed animals and each was co-hybridized separately to the tiled microarrays to replicate the observed transcriptional profiles. The tiled microarray data from the three biological replicate samples was quantile normalized, background corrected with a 1% false positive rate, and converted to log2 space, before analyzing the transcriptional profiles of each sample. This microarray data were viewed using SignalMap, which allowed each microarray data set to be viewed together with the exact same scale and genomic region. Regions of transcription were identified as not changing between the control and stress exposed microarray data sets. A 10.6 kb window of scaffold_21 showed equal expression in both of the exposed samples and their respective control samples as well as accurate coverage of gene models with precisely detectable exon/intron boundaries. (F1). This indicated that processing of the arrays and normalization of the data did not introduce a bias into the transcriptional data that would affect further differential gene expression analysis. Further validation of these data was done by performing reverse transcriptase (RT) PCR to show that regions of transcription seen on the array are truly represented in the cDNA samples. The RT-PCR was done using the *Chaoborus* exposed RNA sample as template. Fifty five regions were selected that represent the types of transcription identified on the microarrays including gene model validations, gene model
extensions, and regions of transcription that appear to not be connected to any gene model. 5’ rapid amplification of cDNA ends (RACE) was also employed to further validate novel transcriptional regions, and three of the 5’ extensions that were verified by RT-PCR were extended to the very 5’ end of the gene by RACE (T1). The PCR products were sequenced and aligned to the *Daphnia pulex* genome using the genome browser BLAST function (F2). This validated the microarray data, and makes it a useful tool for improving the current gene models.

![Figure 1](image)

**Figure 1.** Microarray data tracks loaded into SignalMap identify equally expressed regions of transcription corresponding to gene predictions. Tiled microarray data tracks from the Chaoborus and Metal exposures (orange) along with their corresponding controls (green) loaded into SignalMap. Each track is set to the same log2 scale to allow for direct comparisons of the data sets to identify any changes in transcription. A 10.6 kb (scaffold_21:56000-66600) window of the *Daphnia pulex* genome shows five annotated JGI gene regions that are well expressed in the microarray data and are not changing in the exposed versus the control samples. The star indicates a region of transcription that is well defined in the microarray data across all four conditions, but is not currently incorporated in the gene prediction.
Figure 2. RT-PCR and RACE validate novel transcription identified in the microarray data set. Novel regions of transcription were identified in the microarray data sets, and validation of novel expression was performed. A previously unidentified region of transcription 5’ to a JGI gene prediction showed co-regulation with the gene model across both the Chaoborus and Metal exposure microarray experiments. RT-PCR confirmed three novel 5’ exons that are linked to the current JGI gene prediction. RACE identified the very 5’ end of the gene model and added two more exons to the gene model that were not identified by RT-PCR. All together, this adds five novel 5’ exons to the JGI gene model.

Table 1. RT-PCR and RACE validation of novel transcription identified in the Chaoborus exposed microarray data

<table>
<thead>
<tr>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel 5’ JGI Gene Extensions</td>
<td>10</td>
</tr>
<tr>
<td>Novel 5’ JGI Gene Exons</td>
<td>13</td>
</tr>
<tr>
<td>Novel 5’ RACE JGI Gene Extensions</td>
<td>3</td>
</tr>
<tr>
<td>Novel Interenal JGI Gene Exons</td>
<td>3</td>
</tr>
<tr>
<td>Novel 3’ JGI Gene Extensions</td>
<td>5</td>
</tr>
<tr>
<td>Potential Stand-Alone Genes</td>
<td>5</td>
</tr>
<tr>
<td>Confirm One Gene Model Over Another</td>
<td>9</td>
</tr>
<tr>
<td>Failed-no PCR Product</td>
<td>10</td>
</tr>
<tr>
<td>Failed RACE</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total Regions</strong></td>
<td><strong>55</strong></td>
</tr>
</tbody>
</table>
Validating and updating gene models

The *Daphnia pulex* gene model JGI version 1.1 was generated in 2007 and was primarily based on EST sequences and protein coding genes in related species. This gene model prediction was able to incorporate 72% of the EST evidence available into 30,907 protein coding gene models. It also showed that 15,863 of these predicted genes had significant protein homology (T2). The newest prediction, Genes-2010, incorporated the whole genome tiled microarray data set as another predictor of gene expression. This data set incorporated 90% of the EST evidence available into 42,482 protein coding gene predictions. Of these predicted genes, 22,098 show significant protein homology. The data set predicted 6,405 genes that match exactly to the JGI gene model prediction, 24,502 genes that overlap part of a JGI gene model prediction, and 11,575 gene models that were previously unannotated (T2). An accurate gene model is important for further analysis of stress responsive transcription. The Genes 2010 gene model provides the most comprehensive gene model to date for *Daphnia pulex*.

Table 2. Comparison of JGI gene annotation to the most current 2010 *Daphnia pulex* gene annotation

<table>
<thead>
<tr>
<th></th>
<th>JGI-2007 Gene Annotation</th>
<th>2010 Gene Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EST Coverage Incorporated into Model</td>
<td>75%</td>
<td>90%</td>
</tr>
<tr>
<td>Gene Models with Coded Protein Homology (e-value&lt;= 1e-5)</td>
<td>15863</td>
<td>22098</td>
</tr>
<tr>
<td>Gene Models Matching In Both Predictions</td>
<td>6405</td>
<td>6405</td>
</tr>
<tr>
<td>Updated Gene Models</td>
<td>NA</td>
<td>24502</td>
</tr>
<tr>
<td>Gene Models not in JGI Prediction</td>
<td>NA</td>
<td>11575</td>
</tr>
<tr>
<td>Total Number of Predicted Genes</td>
<td>30907</td>
<td>42482</td>
</tr>
</tbody>
</table>
**SOLiD whole transcriptome RNA sequencing.**

Whole transcriptome RNA sequencing generated short read sequence information from actively transcribed regions within *Daphnia pulex*. The same stress exposures that were done for the microarray transcriptome analysis were also done for RNA sequencing. cDNA libraries were prepared specifically for the SOLiD sequencing platform, and each library was separately barcoded. The barcoding of the libraries allowed for each sample to be pooled together for sequencing, and then analyzed separately. Single end and paired end sequencing was done giving a 50bp from the 5’ end of each 150-250bp of cDNA fragment. The paired end sequencing also generated a 35bp read from the 3’ end of the same piece of cDNA which gave approximately 85bp of sequence per 150-250bp of cDNA fragment. A whole plate was sequenced for both the single end and paired end analysis which generated coverage of fifty million reads per sample. The mapped sequence reads from the RNA sequencing not only confirmed the microarray data, but also provided sequence information and strandedness of each region of transcription (F3). This information combined with the microarray data was useful for further validating genome annotation predictions and will be incorporated into future gene predictions and for further characterization of novel stress responsive regions of transcription.

**Analysis of *Daphnia* homologous stress responsive genes**

Bioinformatic analysis of the tiled microarray data identified genes that were increased or decreased in response to *Chaoborus* predation or heavy metal stress when compared to the control samples. These genes were analyzed across all three biological replicates to ensure that the genes were truly differentially expressed due to the stressed
Figure 3. Analysis of microarray data and RNA sequencing validates gene predictions. The microarray data tracks identify expression that matches part of the JGI gene model and also provides evidence for novel transcription that was missed by the gene prediction. The incorporation of the microarray data into the genes 2010 prediction algorithm extends the JGI gene model and shows evidence for an internal gene model. Solid RNA sequencing confirms the 2010 gene predictions. The raw Chaoborus control RNA sequencing track is loaded into the Integrative Genome Viewer (IGV) allowing for strandedness and connectivity of the reads to be identified. The sequencing of the Chaoborus control sample shows directionality of transcription, where red is top strand gene and blue is bottom strand transcription, matching the gene prediction. Also, the paired end data indicate connectivity of transcription and shows that the sequenced regions are connected as indicated by the 2010 gene model condition and not because of handling of the samples. Any genes that were up or down regulated two fold or more and had a significant P-value of less than or equal to 0.05 were characterized as being stress responsive. Known stress responsive genes were identified as differentially expressed in response to the Chaoborus predation or the heavy metal stress, further validating this data set. The array data identified a cuticular protein coding gene that was significantly upregulated in response to Chaoborus predation stress (F4). Eighteen cuticle genes were identified as significantly increased due to Chaoborus exposure, and were likely pivotal to the development or maintenance of neck teeth as a
defense against predation. Previously uncharacterized predation stress responsive genes were also shown to be upregulated. These genes have no known sequence homology, and upregulated in the Chaoborus predation stress response is the only known function of these genes (F5). These uncharacterized stress responsive protein coding genes tend to be expressed at low levels in the non-stressed control samples then upregulated in response to stress. The heavy metal stressed samples also showed expected changes in gene expression in response to the toxins. Metallothionein genes were known to be important for metal binding/transport, and two of these genes had been identified by gene expression arrays as upregulated in Daphnia exposed to cadmium stress (F6). These two genes were not tiled on the Daphnia microarray, but the RNA sequencing correlated very well to the gene prediction, and showed a very strong change in the number of total reads in every metal treated sample verses the metal control sample. The up regulation of the metallothionein genes is a strong indicator of heavy metal stress, and indicated that these animals were still in the process of responding to the stressors when they were collected. Other stress responsive genes were also identified in the microarray data set. An oxidative stress-induced growth inhibitor was upregulated in response to the metal stress, and likely correlates with the reduction in Daphnia pulex size that has been characterized during long exposures to sub-lethal amounts of cadmium (F7). A transcription factor initiation gene was shown to be significantly upregulated in response to the metal stress. This gene had not previously been characterized as stress responsive, but it may be playing a role in regulating the expression of other stress responsive genes (F8). Genes were also characterized as being down regulated in both stress responses. The genes that were down regulated in both stress responses are likely gene groups that are repressed as
part of a general stress response. A trypsin-like serine protease gene was identified from the microarray analysis as being significantly down regulated in both the metal response and the predation response (F9). General metabolism genes were expected to be reduced in the stress response as the stress response would take priority over other cellular functions. Microarray analysis of all the stress responsive genes indicates that 174 genes are induced two fold or greater, and 248 are repressed two fold or greater in the Chaoborus predation response, while 1165 genes are induced two fold or greater, and 1724 genes are repressed two fold or greater in the heavy metal response. Also, 86 of the same genes show two fold or greater repression in both stress responses, and 30 of these genes are described as having metabolic function. Four genes are showing significant increase in both stress exposures, indicating that specific genes are up regulated to respond to different types of stress (T3).

**Gene ontology (GO) analysis**

Gene ontology analysis was done using GO Slim terms that were generated as part of the *Daphnia pulex* genes 2010 annotation. The GO Slim terms are a scaled down version of GO ontologies that give a broad classification of gene products. This was particularly useful for this data set as many of the identified genes have not been functionally validated. Analysis of the GO Slim terms that are enriched generate a reliable overview of gene groups that were enriched in the different stress responses. GO term enrichment was analyzed via the agriGO classification tool, and the gene sets that were enriched significantly (P-value of at least 0.05) in response to *Chaoborus* predation or the heavy
Figure 4. Predation stress induced cuticular protein identified in microarray data set and confirmed in RNA sequencing. A cuticular protein coding gene is upregulated greater than two fold as indicated by the bioinformatic analysis of the microarray data for the Chaoborus exposed sample versus the control. The Chaoborus exposed (orange) and control (green) microarray tracks were loaded into SignalMap and set to the same log2 scale for a direct comparison of expression levels between the samples. The paired end RNA sequencing Chaoborus exposed and control RNA sequencing tracks are loaded into IGV. The total read counts are generated using the normalized single end read data set and indicates that there is a two-fold or greater difference in expression across the gene regions.
Figure 5. An uncharacterized protein coding gene with no known homology are upregulated greater than two fold as indicated by the bioinformatic analysis of the microarray data for the Chaoborus exposed sample versus the control. The Chaoborus exposed (orange) and control (green) microarray tracks were loaded into SignalMap and set to the same log2 scale for a direct comparison of expression levels between the samples. This same gene region is well covered by the RNA sequencing data. The paired end RNA sequencing Chaoborus exposed and control RNA sequencing tracks are loaded into IGV. The total read counts are generated using the normalized single end read data set and indicates that there is a two-fold or greater difference in expression across the gene regions.
Figure 6. Metallothionein genes identified in the RNA sequencing data set as upregulated greater than two fold across every metal treatment. Two predicted metallothionein genes that were known to be up regulated in Daphnia exposed to cadmium stress were identified in the RNA sequencing data. The two gene regions are loaded into the *Daphnia pulex* genome browser to view the mapped RNA sequencing reads. The scale to the right of each RNA sequencing track indicates the normalized single end read count and this shows a greater than two fold expression of each gene under every metal stressed condition as compared to the metal control sample. This indicates that the animals are actively responding to the metal stress across each metal exposed RNA sequencing sample at the time that they were collected for analysis.
Figure 7. An oxidative stress induced growth inhibitor identified as metal stress induced in the microarray data and confirmed with RNA sequencing. An oxidative stress induced growth inhibitor protein coding gene is upregulated greater than two fold as indicated by the bioinformatic analysis of the microarray expression in the metal exposed sample versus the control. The metal exposed (orange) and control (green) microarray tracks were loaded into SignalMap and set to the same log2 scale for a direct comparison of expression levels between the samples. This same gene region is well covered by the RNA sequencing data. The raw metal exposed and control paired end RNA sequencing tracks are loaded into IGV both confirm the gene model predictions. Read count analysis was done using normalized single end RNA sequencing gene counts to validate that the gene region is increased two fold in total read counts in the Arsenic exposure versus the Metal Control.
Figure 8. A transcription initiation factor gene identified as metal stress induced in microarray data and confirmed with RNA sequencing. A transcription initiation factor protein coding gene is upregulated greater than two fold as indicated by the bioinformatic analysis of the microarray expression in the metal exposed sample versus the control. The metal exposed (orange) and control (green) microarray tracks were loaded into SignalMap and set to the same log2 scale for a direct comparison of expression levels between the samples. This same gene region is well covered by the RNA sequencing data. The raw metal exposed and control paired end RNA sequencing tracks are loaded into IGV both confirm the gene model predictions. Read count analysis was done using normalized single end RNA sequencing gene counts to validate that the gene region is increased two fold in total read counts in the Arsenic exposure versus the Metal Control.
Figure 9. A serine-like protease gene repressed in under both stress conditions. Genes that are not immediately essential for life functions, like metabolism genes, are predicted to be down regulated as part of a general stress response pathway to shift transcriptional energy to the up regulation of specific stress responsive genes. The same serine-protease like gene is shown to be down regulated by more than two fold in both the metal and *Chaoborus* stress response as indicated by the microarray data. The Chaoborus and metal exposed (orange) and control (green) microarray tracks were loaded into SignalMap and set to the same log2 scale across each track to allow for a direct comparison of gene expression in all samples. This same gene region is well covered by the RNA sequencing data. The *Chaoborus* exposed and control RNA sequencing tracks are loaded into IGV, and a two-fold or greater change in expression is verified by the read counts.

<table>
<thead>
<tr>
<th>Gene Evidence</th>
<th>Chaoborus Induced</th>
<th>Chaoborus Repressed</th>
<th>Metals Induced</th>
<th>Metals Repressed</th>
<th>Shared Induced</th>
<th>Shared Repressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes with Homologs</td>
<td>78</td>
<td>184</td>
<td>373</td>
<td>1103</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>Genes with Paralogs</td>
<td>56</td>
<td>46</td>
<td>51</td>
<td>326</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>Uncharacterized Array Expr.Coding Evidence</td>
<td>33</td>
<td>17</td>
<td>197</td>
<td>248</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Uncharacterized Array Expr./small ORF</td>
<td>7</td>
<td>1</td>
<td>544</td>
<td>47</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>174</strong></td>
<td><strong>248</strong></td>
<td><strong>1165</strong></td>
<td><strong>1724</strong></td>
<td><strong>4</strong></td>
<td><strong>86</strong></td>
</tr>
</tbody>
</table>

stress responsive genes identified by the microarray data set are changing at least 2fold and have a P-value less than or equal to 0.05
metal stress are summarized in Tables 4 and 5, and GO analysis terms that were shared across both stress responses are summarized in Table 6. The four genes that are upregulated in both stressors have no known homology and do not have GO terms assigned to them. The analysis indicated that within the same functional group there were some genes that are induced and were likely playing an important stress response role, while others are repressed in the heavy metal or the predation stress response as a consequence of cellular resources being devoted to responding to the stressor. The repressed genes that were identified likely represent a set of genes that the organism can temporarily shut down as energy is diverted to the initiation of the stress response. Gene groups that were enriched in the heavy metal induced gene set are likely to be directly responding to the stress exposure. Interestingly, the GO analysis showed enrichment in transcription associated gene groups including transcription factor activity, transcription regulation activity. This indicated that the animals were still in the process of regulating transcriptional events in an effort to respond to the stress. Also, metal binding, transport activity, and cell-cell signaling GO groups were identified as significantly enriched in the metal induced data set, further indicating that these animals are actively responding to the metal stress. The GO analysis from the *Chaoborus* predation data set showed that metabolism genes were repressed while structural molecule activity genes were induced. This was expected, but had not been verified in the *Daphnia pulex* predation stress response. Unlike acute metal toxicity, predation does not seem to induce transcription associated genes (T5). It is possible that the acute toxicity treatment with the heavy metals could be uncovering the genes that were involved in the early stages of the metal stress response. As predicted, GO functions that were significantly enriched in both
repressed data sets primarily consist of metabolism functions. It is likely that these genes make up a group of genes that *Daphnia* can temporarily repress in an effort to focus cellular energy on the up regulation of genes directly involved in defending against the detrimental effects of the environmental toxins.

Table 4. GO analysis of heavy metal stress responsive genes

<table>
<thead>
<tr>
<th>Induced GO Function</th>
<th>P-Value</th>
<th>Repressed GO Function</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>transcription factor activity</td>
<td>1.10E-06</td>
<td>translation</td>
<td>8.80E-10</td>
</tr>
<tr>
<td>transcription regulator activity</td>
<td>1.10E-05</td>
<td>protein metabolic process</td>
<td>9.00E-07</td>
</tr>
<tr>
<td>cell-cell signaling</td>
<td>4.50E-05</td>
<td>generation of precursor metabolites and energy</td>
<td>6.70E-05</td>
</tr>
<tr>
<td>anatomical structure morphogenesis</td>
<td>0.00019</td>
<td>structural molecule activity</td>
<td>1.80E-15</td>
</tr>
<tr>
<td>anatomical structure development</td>
<td>0.00019</td>
<td>peptidase activity</td>
<td>2.00E-08</td>
</tr>
<tr>
<td>cell communication</td>
<td>0.0002</td>
<td>extracellular region</td>
<td>4.50E-25</td>
</tr>
<tr>
<td>metal ion binding</td>
<td>0.00039</td>
<td>vacuole</td>
<td>2.10E-06</td>
</tr>
<tr>
<td>transmembrane transporter activity</td>
<td>0.0015</td>
<td>extracellular matrix</td>
<td>7.40E-06</td>
</tr>
<tr>
<td>ion channel activity</td>
<td>0.002</td>
<td>proteinaceous extracellular matrix</td>
<td>7.40E-06</td>
</tr>
<tr>
<td>ion transmembrane transporter activity</td>
<td>0.002</td>
<td>lytic vacuole</td>
<td>8.00E-05</td>
</tr>
<tr>
<td>substrate-specific transmembrane transporter activity</td>
<td>0.002</td>
<td>lysosome</td>
<td>8.00E-05</td>
</tr>
<tr>
<td>passive transmembrane transporter activity</td>
<td>0.002</td>
<td>lipid particle</td>
<td>0.00063</td>
</tr>
<tr>
<td>substrate-specific transporter activity</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>receptor binding</td>
<td>0.0059</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Stress responsive genes identified by the microarray data set are changing at least 2-fold and have a P-value less than or equal to 0.05.
Table 5. GO analysis of *Chaoborus* predation stress responsive genes

<table>
<thead>
<tr>
<th>Induced GO Function</th>
<th>P-Value</th>
<th>Repressed GO Function</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>structural molecule activity</td>
<td>6.00E-06</td>
<td>extracellular region</td>
<td>5.50E-28</td>
</tr>
<tr>
<td>extracellular region</td>
<td>2.00E-05</td>
<td>peptidase activity</td>
<td>1.90E-18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hydrolase activity</td>
<td>5.00E-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>protein metabolic process</td>
<td>1.50E-07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>carbohydrate metabolic process</td>
<td>4.60E-06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>primary metabolic process</td>
<td>5.80E-06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>catalytic activity</td>
<td>1.60E-05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lytic vacuole</td>
<td>5.50E-05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lysosome</td>
<td>5.50E-05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>carbohydrate binding</td>
<td>7.40E-05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>receptor activity</td>
<td>0.00033</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lipid metabolic process</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

Stress responsive genes identified by the microarray data set are changing at least 2-fold and have a P-value less than or equal to 0.05

**Identification of uncharacterized DE genes**

A portion of the stress responsive genes identified by the microarray analysis are predicted protein coding genes with unknown function. This group of genes does not have GO functions ascribed to them and very little was known about them other than they were up regulated in response to stress. Predicted protein coding genes with open reading frames greater than 70 amino acids and no known protein homology were up regulated in both stress responses, but there are many more of these identified in the metal stress response (10). Genes with low coding potential, open reading frames of less than 70 amino acids, were also identified as upregulated in both data sets, but many more were identified as metal stress responsive (11). The low coding potential gene predictions are primarily based on the microarray data set and are a low confidence prediction, but the RNA sequencing data do provide transcriptional evidence for 134 of these predictions.
Little is known about these genes aside from their up regulation during stress exposure, but one possibility is that these regions of transcription have a primary function as part of the initial heavy metal stress response pathway. In the predation experiment, the animals had already generated their inducible predation defense by the time they were collected for the transcriptome analysis. Stress responsive genes were identified as differentially expressed in the predation experiments, but the initial stress responsive genes were likely no longer differentially expressed. This could explain why there were 741 uncharacterized or low coding genes that were up regulated and 295 down regulated in the heavy metals, while in the predation experiment only 40 of these genes were up regulated and 18 are down regulated (T3). It is possible that these uncharacterized genes were playing a specialized role in *Daphnia* that is only revealed under stressed conditions.

**Gene expression analysis with RNA sequencing**

The microarray transcriptome analysis revealed stress responsive regions of transcription, many of which corresponded to gene models with no known function in *Daphnia*. Counting of the normalized single end RNA sequencing reads per gene in a stress exposed sample versus its corresponding control sample provided a way to validate the differential gene expression identified by the microarray analysis (T6). The combination of these two data sets identified 28 genes that were induced and 190 genes that were repressed in the *Chaoborus* predation response, while 255 genes were induced and 606 genes were repressed in the metal stress response. The RNA sequencing data also validated some of the uncharacterized genes are up regulated two fold or greater in both stress responses. The differential analysis of the normalized single end RNA
Figure 10. Uncharacterized protein coding gene induced under metal stress and confirmed with RNA sequencing. Uncharacterized genes are identified as stress responsive in the microarray data. (A) Protein coding genes with no sequence homology and unknown function that are increased in response to metal stress may have a primary role in the stress response. These genes have been shown to be upregulated two fold or more in the metal stress microarray data set. The metal exposed (orange) and control (green) microarray tracks were loaded into SignalMap and set to the same log2 scale for a direct comparison of expression levels between the samples. The raw metal exposed and control RNA sequencing tracks are loaded into IGV, and the total read count across each gene region verifies that transcription is induced two fold or greater in the metal stress response.
Figure 11. Uncharacterized gene with low protein coding potential induced under metal stress and confirmed with RNA sequencing. Genes that are identified as having low protein coding potential are also identified as differentially expressed in response to metal stress in the microarray data. A subset of these low coding potential genes has been confirmed by RNA sequencing. Further analysis of these genes could identify them as stress responsive non-coding RNAs. These genes have been shown to be upregulated two fold or more in the metal stress microarray data set. The metal exposed (orange) and control (green) microarray tracks were loaded into SignalMap and set to the same log2 scale for a direct comparison of expression levels between the samples. The raw metal exposed and control RNA sequencing tracks are loaded into IGV, and the total read count across each gene region verifies that transcription is induced two fold or greater in the metal stress response.

sequencing data does not validate as many stress responsive genes that were identified via microarray. One possibility for the discordance in the microarray data set and RNA sequencing data set was that the expression of genes with paralogs in the genome could be generating false differential expression calls in the microarray data. To analyze this, only genes without any identified paralogs in the genome were analyzed for differential expression across both data sets. Even with the removal of genes with predicted paralogs
from the data analysis, 8 of 65 genes are correlating in the Chaoborus induced data sets, 27 of 52 genes are correlating in the Chaoborus repressed data sets, 124 of 923 genes are correlating in the metals induced data sets, and 49 of 705 genes are correlating in the metals repressed data sets. The incorporation of the paired end sequencing data set into the differential gene expression will help to further validate differential transcription identified in the microarray data set, and will provide the final list of stress responsive genes verified in both data sets.

Table 6. Differentially expressed genes identified by microarray analysis and confirmed by RNA sequencing

<table>
<thead>
<tr>
<th>Gene Prediction</th>
<th>Chaoborus Induced Genes</th>
<th>Chaoborus Repressed Genes</th>
<th>Metals Induced Genes</th>
<th>Metals Repressed Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characterized Protein Coding Genes</td>
<td>12</td>
<td>150</td>
<td>73</td>
<td>351</td>
</tr>
<tr>
<td>Uncharacterized Genes (ORF&gt;70)</td>
<td>16</td>
<td>40</td>
<td>71</td>
<td>236</td>
</tr>
<tr>
<td>Uncharacterized Genes (ORF&lt;70)</td>
<td>0</td>
<td>0</td>
<td>111</td>
<td>19</td>
</tr>
<tr>
<td>Correlating DE Genes without Predicted Paralogs</td>
<td>8 (65)</td>
<td>27 (52)</td>
<td>124 (923)</td>
<td>49 (705)</td>
</tr>
<tr>
<td>Total Correlating DE Genes</td>
<td>28</td>
<td>190</td>
<td>255</td>
<td>606</td>
</tr>
</tbody>
</table>

Stress responsive genes identified by the RNA sequencing data set are changing at least 2fold
DE analysis of genes induced by individual metals

Analysis of the SOLiD RNA sequencing allowed for the transcriptional responses from each individual metal exposure to be compared to the non-metal exposed control sample. This is something that could not be done with the microarray data since the heavy metal exposed RNA collections were mixed together before hybridization to the microarrays. To focus in on genes that were significantly induced in the individual metal exposures, the median fold induction of the two metallothionene (identified in figure 4) was calculated and to be 4.69 and 4.89 fold across each individual metal exposure. The metallothionene genes are known to be transcriptionally increased in *Daphnia* under cadmium stress, and genes that are increased at this same level or higher are likely to also have similarly important role in alleviation of heavy metal stress. The fold cutoff that was used to identify genes as significantly induced in the individual metal responses at least a 4.5 fold induction. Analysis of the metal exposed RNA sequencing data using this cutoff identified a total of 1223 genes that are induced at least 4.5 fold in response to each individual stressor (T7). The 880 genes that are identified as being induced greater than 4.5 fold in only one of stressors tested indicates that there are individual genes that are differentially expressed specifically in response to individual metal stressors. Further analysis of the genes that are responding to each specific metal will be pursued, but it is clear that each metal stressor induces specific sets of genes that likely function specifically to uniquely respond to the stress.

Analysis of uncharacterized stress induced coding genes

While further work is needed to determine the functional roles of these stress responsive uncharacterized regions of transcription, the RNA sequencing data can be
Table 7. RNA sequencing transcriptional analysis of each individual metal exposure identifies metal stressor specific gene induction

<table>
<thead>
<tr>
<th>Number of Genes Induced Across Metal Exposures</th>
<th>Genes Identified In Only One Individual Metal Exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1223</td>
<td>880</td>
</tr>
<tr>
<td></td>
<td>208</td>
</tr>
<tr>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>12</td>
</tr>
</tbody>
</table>

Genes were called significantly induced if they were expressed at least 4.5 fold greater than the corresponding non-metal exposed control sample.

used to confirm the accuracy of the low homology gene predictions. An uncharacterized gene was predicted in the Genes 2010 annotation as a top strand gene with three exons. It was located between two conserved bottom strand gene predictions. The direction of transcription identified by the RNA sequencing data set confirmed that there was a region of top strand transcription between the two regions of bottom strand transcription (12). The connectivity of transcribed regions is identifiable by the paired end sequence reads that map across exon intron boundaries, and confirms that there are five exons belonging to this gene model (13). Exon-intron splice consensus sites were identified accurately with RNA sequencing, and the data confirmed the predicted splice sites for the gene model (13). The RNA sequencing data actually removes some nucleotide sequence from both ends of the gene prediction, but this did not change the longest predicted open
reading frame (14). The RNA sequence analysis of this predicted stress responsive gene confirmed that it codes for a protein that has no known homology. The manual curation of this gene region with the RNA sequencing data confirmed the gene predication and validated that this stress responsive gene coded for a peptide that had no known homology. This result further validated that these uncharacterized stress responsive genes could represent a novel class of protein coding genes with a primary role in responding to environmental stress.

Figure 12. Uncharacterized differentially expressed gene identified as being metal stress responsive. Paired end RNA sequencing data provides a tool for manual assembly of gene models to further characterize stress responsive genes that have no predicted gene homology. A predicted protein coding gene with no known homology was identified as a top strand gene with a two-fold or greater induction under metal stress.
Figure 13. Manual gene annotation with paired end RNA sequencing allows for prediction of splice consensus sites. Paired end RNA sequencing data provides a tool for manual assembly of gene models to further characterize stress responsive genes that have no predicted gene homology. Mapped mate pairs provide evidence for connectivity across exon-intron boundaries which are identified in IGV by highlighting the mate pairs with matching colors. This also provides better coverage across the length of an expressed gene which can be used to identify the exact consensus (GT/AG) splice sites between these boundaries.
Figure 14. Manually annotate nucleotide sequence assembled from paired end RNA sequencing confirms an uncharacterized gene with no known homology that is induced under metal stress. Paired end RNA sequencing data provides a tool for manual assembly of gene models to further characterize stress responsive genes that have no predicted gene homology. This allows for the determination of the very start and end of a gene region, as well as the exact nucleotide sequence across each exon within the gene model. In this example the predicted exon-intron splice sites are validated by RNA-seq and nucleotide sequence was removed from the 5' end and 3' of the gene (red letters), but it does not change the longest open reading frame coded by this gene, and validates that this gene is a novel stress responsive gene identified in *Daphnia pulex*. 
RNA sequencing analysis of predicted non-coding genes

Regions of differentially expressed transcription were also identified that have low protein coding potential. These gene predictions contain small open reading frames, and while they could be coding for very small peptides, it is also possible that these genes represent non-coding RNAs that are stress responsive (15). Manual assembly of the transcribed region was done using the paired end RNA sequencing data, and a three frame translation was performed in the top strand direction. The longest ORF identified was 45 amino acids (green) and none of the three predicted ORFs showed any known homology (16). This same manual assembly of nucleotide sequence from predicted gene regions with low protein coding potential was performed on two other regions that were identified as stress responsive in both data sets (F17,18). Protein prediction from a three frame translation of the manually assembled data identifies small potential ORFs with no known homology. These regions of transcription are confirmed as having low protein coding potential, and are stress responsive in both data sets. It is possible that these are non-coding RNAs that has never before been characterized in *Daphnia pulex*, and is playing a role in the stress response.
Figure 15. Uncharacterized differentially expressed gene with low coding potential identified as being metal stress responsive. An RNA sequencing track is loaded into the Integrative Genome Viewer for further sequence analysis. The RNA sequencing data shows that the uncharacterized protein coding gene is a top strand (red color) gene instead of a bottom strand gene that the gene annotation predicted.
Figure 16. Manually annotate nucleotide sequence assembled from paired end RNA sequencing confirms an uncharacterized gene with no known homology that only contains very small potential open reading frames. This stress metal stress induced gene may be a putative non-coding RNA. The manual assembly of the sequence identified in this region does not identify any intron/exon splice sites, and top strand ORF analysis of the sequence predicts three small potential protein coding sequences that have no known homology. The ORF analysis confirms that this is a region of transcription with low protein coding potential that is induced under metal stress.
Figure 17. A region of differential expression manually assembled using RNA sequencing data confirms low protein coding potential. RNA sequencing tracks are loaded into the Integrative Genome Viewer for further sequence analysis of regions described as having low protein coding potential and are differentially expressed greater than two fold as identified by the normalized single end RNA sequencing read counts. The paired end data tracks are loaded to show coverage of the gene prediction, and were used for manual sequence assembly. The RNA sequencing data shows that the uncharacterized protein coding gene is a top strand (red color) gene instead of a bottom strand gene that the gene annotation predicted in both of these transcribed regions. The manual assembly of the sequence identified in this region does not identify any intron/exon splice sites, and top strand ORF analysis of the sequence predicts three small potential protein coding sequences that have no known homology. The ORF analysis confirms that this is a region of transcription with low protein coding potential that is induced under metal stress.
Figure 18. A differentially expressed region of low coding potential with a manually assembled nucleotide sequence indicates only small ORF potential. RNA sequencing tracks are loaded into the Integrative Genome Viewer for further sequence analysis of regions described as having low protein coding potential and are differentially expressed greater than two fold as identified by the normalized single end RNA sequencing read counts. The paired end data tracks are loaded to show coverage of the gene prediction, and were used for manual sequence assembly. The RNA sequencing data shows that the uncharacterized protein coding gene is a top strand (red color) gene instead of a bottom strand gene that the gene annotation predicted in both of these transcribed regions. The manual assembly of the sequence identified in this region does not identify any intron/exon splice sites, and top strand ORF analysis of the sequence predicts only two small potential protein coding sequences that have no known homology. The ORF analysis confirms that this is a region of transcription with low protein coding potential that is induced under metal stress.
CHAPTER III

DISCUSSION OF THE DATA ANALYSIS

*Daphnia pulex* genome annotation

The analysis of the tiled microarray data in conjunction with the paired end RNA sequencing provides a tool for development of accurate gene models for *Daphnia pulex*. This data can predict gene regions even if there is little other transcriptional evidence available. Analysis of *Daphnia* transcriptome under stress conditions is a unique approach to genome annotation. This approach has identified genes that are expressed under normal laboratory conditions and are not responding to stress, as well as identifying genes that are stress responsive. Interestingly there is a set of uncharacterized stress responsive genes that are expressed at low levels in the non-stressed controls and induced in the stress response. These genes lack any significant homology and may never have been identified if the transcriptome analysis had only been performed under non-stressed laboratory conditions. Microarray analysis of differentially expressed regions of transcription can be used to predict gene models, but the incorporation of RNA sequencing makes these predictions better. The RNA sequencing data provides sequence information, strandedness of transcription, exon intron splice sites, expression levels, and connectivity of transcribed regions. The incorporation of the paired end RNA sequencing into the gene predictions will generate a highly accurate gene annotation for *Daphnia pulex*. 
**Identified stress responsive transcription**

Differentially expressed genes are identified in the microarray data set and a subset of these genes was validated as differentially expressed based on the normalized single end read counts from the RNA sequencing data set. This data set of differentially expressed genes contains a mix of genes with known homologs and genes that have no known homology or function. Some conserved genes like the metallothionein genes were previously known to be involved in stress response. The unknown genes that showed a two-fold or greater change in expression may represent a new class of genes that are specifically functioning as part of the stress response. Deeper analysis of the RNA sequencing data can be used to identify sequence similarities or similar protein coding domains that are conserved specifically within these uncharacterized stress responsive genes. Another subset of differentially expressed genes are predicted as low protein coding potential and may be non-coding RNAs. Non-coding RNAs have certainly been identified as playing a role in the stress response of other organisms, and may be represented in this data set of *Daphnia* stress responsive genes. RNA sequencing analysis does validate some of these low coding gene predictions, but further analysis of these regions needs to be done to identify whether they are non-coding RNAs or genes that code for uncharacterized small peptides. There is discordance with differentially expressed genes identified between the RNA sequencing and microarray data set. One possibility was that the paralogous gene expression was the cause of this disparity, but it was confirmed that the discordance in the two gene sets exists between genes that are unique within the genome. It is likely that a subset of genes that are identified in the microarray data set are affected by spurious cross hybridizations of the probes on the
array which are giving a false indication of differential analysis. Also, the use of different clonal populations in between the metal stress exposures is also a likely cause of some discordance between differentially expressed genes identified in both data sets. The paired end RNA sequencing data set provides a data set that is likely more accurately mapped to gene predictions. The paired end RNA sequencing should provide better concordance between the two data sets and will be used to generate the final predicted stress responsive genes identified in these data sets.

Conclusions

Analysis of the *Daphnia pulex* transcriptome under stressed and non-stressed conditions reveals stress responsive regions of transcription that would likely not be identified under non-stressed laboratory conditions. Furthermore, the identification of previously uncharacterized genes as stress responsive provides a function to these genes that could not be identified under non-stressed conditions. Transcriptome analysis performed under non-stressed conditions is likely missing a subset of genes that are primarily expressed in response to environmental changes. These stress specific changes in gene expression can be used to develop *Daphnia* as an ecotoxicological tool for the early identification of environmental toxins. It is clear that *Daphnia* is responding transcriptionally to sub-lethal heavy metal stress after only 24 hours of exposure. It is possible that *Daphnia* would show some of these same transcriptional changes to even trace amounts of heavy metals. While more experiments need to be done to confirm this, the data shows the potential for development of *Daphnia* as a model system for early identification of heavy metal stress. The characterization of stress responsive transcription in *Daphnia pulex* provides evidence for a hidden transcriptome that is only
revealed when organisms are forced to adapt to changes in their environment. This suggests that previous transcriptome studies may have missed identifying interesting classes of genes that are only easily identified under stressed conditions. Future transcriptome studies should consider including samples collected under different environmental conditions in an effort to better understand these stressor responsive regions of transcription.
CHAPTER IV

MATERIALS AND METHODS

*Daphnia pulex* predator exposure and RNA isolation

*Daphnia pulex* were exposed to predation stress from the phantom midge *Chaoborus flavicans*. 60-70 *Chaoborus* larvae were added to a nylon net cage and put in the same tank as a clonal population of *Daphnia pulex* adult females so they are exposed to *Chaoborus* karymones without actual contact with the predators. Affected offspring were collected at the 1st juvenile instar stage as well as control 1st instar juveniles that were not exposed to *Chaoborus*. Induction of stress response is characterized by the development of neck teeth defenses, which are not present in control population. Total RNA was collected from both the *Chaoborus* exposed and non-exposed R9 clonal populations using the Ambion RNAqueous kit with a DNase treatment.

*Daphnia pulex* metal exposure and RNA isolation

Similarly staged adult *Daphnia pulex* were exposed to sublethal doses of a single heavy metal (20ug/L Cd, 1,348ug/L As, 200ug/L Zn, 200ug/L Ni, or 1ug/L Cu) for 24 hours. Two clonal populations were used for the individual metal exposures arsenic and copper exposures were performed using TCO clonal population and the cadmium, nickel, and zinc exposures were performed using PA33 clonal populations. After exposure, total RNA was extracted from both the metal exposed and non-exposed controls using the Ambion RNAqueous kit with a DNase treatment. The metal exposed RNA was mixed together in equal molar amounts to generate a mixed metal population for microarray analysis.
**Daphnia pulex whole genome tiling array**

A *Daphnia pulex* tiled genomic microarray set was designed by the *Daphnia* Genomics Consortium in concert with the bioinformatics team at Roche NimbleGen. This array set, which comprises two HD2 (2.1 million feature) microarrays, utilizes isothermal probe design (50-75pb in length) that are spaced an average of 30bp apart. The design names of the arrays are as follows: Daphnia_tiling1_HX1 and Daphnia_tiling2_HX1.

**Processing of RNA samples for microarray hybridization**

The collected total RNA from each individual metal treatment was pooled together to make a mixed metal total RNA collection. Total RNA collected from the *Chaoborus* treated, *Chaoborus* control, and mixed metal and metal control was amplified using MessageAmpTM II aRNA kit (Ambion). Double-stranded cDNA was generated starting with 10ug of cRNA using Double-Stranded cDNA Synthesis Kit and random hexamer primers (Invitrogen). The cDNA was amplified and labeled using either Cy3- or Cy5-coupled random nonomers (TriLink) and a competitive hybridization (using 15 ug of each labeled cDNA) was performed that always coupled both stressor treated cDNA and non stressed control cDNA samples in order to facilitate a direct comparison of transcription (see the Roche NimbleGen labeling and hybridization protocol http://www.nimblegen.com/products/lit/chip_userguide_v6p1.pdf for specific labeling and processing details). After hybridization for 16-20 hours, the arrays were washed, dried, and then scanned on an Axon GenePix 4200A microarray scanner from Molecular Devices.
Microarray data analysis

All microarray processing was performed by Don Gilbert and the DGC bioinformatics group. The signal to background noise was determined by measuring the fluorescence of random probes across the array and then calling experimental probes as expressed if they are higher than 99% of the random probe signal. This indicates that only 1% of the fluorescence of the experimental probes are false positives. The software packages LIMMA R and Bioconductor were used to further analyze the microarray data. All of the probes were quantile normalized across all three biological replicates and log2 expression scores were generated for each oligonucleotide probe. Expression values were generated using the normalized oligonucleotide probe expression values and the average of the expression is used as the final expression score for each transcribed region. The expression values for each region of transcription were analyzed across each replicate and probability values were generated using LIMMA. Differential expression was described as significant if there was at least a two-fold change in calculated expression, and this was replicated across three biological replicates with a significant P-value of less than or equal to 0.05.

RT-PCR validation of microarray data

Regions of novel transcription were identified by tiled microarray and were verified using reverse transcriptase(RT) PCR using Chaoborus exposed total RNA as template. First strand cDNA was generated by starting with 5μg of total RNA using the AffinityScript Multiple Temperature cDNA Synthesis Kit (Stratagene) with Oligo dT primers as directed by the manufacturer’s protocol. Transcript specific PCR was done using PFU Ultra II Phusion HS DNA Polymerase (Stratagene). PCR reaction parameters
for cDNA targets are as follows: Extension time of 30 seconds per 1kb, 50-500ng of FS-cDNA, 0.2uM of each transcript specific primer, 250 uM of each dNTP, 1X final concentration of reaction buffer, and 1 unit of DNA polymerase. Thermocycler settings were set as follows: 1 minute at 95 C followed by 30 cycles of 95 C for 20 seconds, Primer Tm-3 C for 20 seconds, and 72 C for 30s/kb of expected transcript length, and a final extension time of 72 C for 3 minutes.

RACE validation of microarray data

RACE was performed using ExactStart eukaryotic RACE kit (Epicentre). First strand cDNA was generated starting from total RNA extracted from Chaoborus treated Daphnia pulex. A 5’ adapter (P1) sequence was added to RNA transcripts by removing the 5’ cap to yield a transcript with a free 5’ monophosphate for P1 adapter linkage. cDNA synthesis was then performed using oligo dT with a 3’ adapter (P2) sequence attached to it. The 5’ and 3’ PCR primers, P1 and P2, are used with a transcript specific primer to define the 5’ or 3’ ends of a transcript. Transcript specific RACE PCR requires two rounds of amplification with two sets of primers. The first PCR requires the use of the original P1 or P2 PCR primers and matching transcript specific primers that are 45-48bp in length with a matching melting temp of the P1 and P2 primers. PCR reaction conditions were one cycle at 98 C (30s), 7 cycles of 98 C (10s), 72 C (6min), then 30 cycles of 98 C (10s) and 67 C (6 min). The second PCR amplification required the use of nested primers representative of the P1 or P2 sequence and a nested transcript specific primer. Both of these nested primers were 26-30bp with a melting temperature of 57-60 C. The parameters for the nested PCR are as follows: One cycle of 98 C for 30s, 5 cycles of 98 C (10s), 62 C (6 min), followed by 20 cycles of 98 C (10s) and 65 C (6 min).
**PCR product analysis**

Agarose gel electrophoresis was performed to analyze the PCR products as well as for size selection and purification of specific products. Gel purification was done using Qiagen gel purification columns as per the manufacturer’s protocol. Transcripts were cloned for sequencing using the Topo TA Cloning Kit for Sequencing (Invitrogen), with the pCR4-TOPO vector. Primers T3 and T7, provided with the kit, were used for sequencing off of this plasmid. Sequencing was performed by the Center for Comparative Genomics using the ABI 3730 DNA Analyzer.

**GO analysis of differentially expressed genes**

SlimGO terms were assigned to annotated *Daphnia pulex* genes that show significant gene homology (E-value <1e-5). The batch analysis of GO terms that were enriched in the stress repressed or induced gene sets were analyzed using the agriGO web based tool (29). This tool recognizes the GOslim terms and identifies significantly enriched GO terms in the data sets. The final output identifies the groups of genes with specific biological functions that are enriched due to stress response.

**cDNA library preparation for SOLiD RNA sequencing**

Total RNA from either stressor induced or no stress control Daphnia was collected as described above for the predation exposure experiment. The same R9 clonal population was used for the exposure collection as well as the non-exposed control. The metal exposures were performed using the same metals and metal concentrations but only the TCO clonal population was exposed and transcriptionally analyzed with RNA sequencing. Also, the metal exposures were not mixed together to make a mixed metal
population, but instead the samples were prepared and analyzed separately to identify the transcriptional response generated by each individual metal. Ribosomal RNA was depleted from these samples using two rounds of depletion with the Invitrogen RiboMinus Eukaryotic kit for RNA-Sequencing. cDNA libraries were constructed as described by the SOLiD Total RNA-Seq Kit protocol. 200-500ng of rRNA depleted sample was RNaseIII fragmented for 25 minutes at 37 C. The fragmented RNA was cleaned up using the RiboMinus Concentration Module (Invitrogen) following manufacturer’s protocol. The fragmentation was analyzed using BioRad Experion High Sensitivity chip following the manufacturer’s protocol. 100ng of fragmented rRNA-depleted samples was dried down using vacuum centrifugation at low heat, and suspended in 3 ul of water. The generation of an amplified whole transcriptome cDNA library was accomplished using the components from the SOLiD Total RNA-Seq Kit (Applied Biosystems). Hybridization and ligation of SOLiD Adapter Mix to the fragmented RNA and reverse transcription was performed following the manufacturer’s protocol. First strand cDNA was purified using MinElute PCR Purification Kit (Qiagen). First strand cDNA was mixed with equal volume of load dye and heated for 3 minutes at 95 C, then snap cooled on ice and run on a 5% pre-cast polyacrylamide denaturing urea gel (BioRad ) with a 50bp DNA ladder and stained with SYBR Gold nucleic acid gel stain (Invitrogen). DNA was visualized with UV illumination and gel between 100-200 nt was excised. This gel was cut into 4 vertical strips, and the 2 outside strips were stored at -20 C as a backup sample. Second strand cDNA synthesis and library amplification was performed directly on the gel strips using the 2 inner vertical gel strips in two separate reactions using components from SOLiD Total RNA-Seq Kit. Reaction conditions were
95 C (5 min), 15 cycles of 95 C (30s), 62 C (30s), and 72 C (30s), and 72 C (7 min). The products from both reactions were pooled and purified using PureLink PCR Micro Kit (Invitrogen). Size distribution of amplified cDNA was analyzed using BioRad Experion High Sensitivity chip to ensure that less than 20% of the amplified cDNA was in the 20-200pb range before proceeding with Solid System template bead preparation and sequencing.

**Barcoding of individual cDNA libraries**

During the construction of the amplified cDNA library, each metal treated sample and the no metal control sample was individually barcoded by substituting a barcode SOLiD 3’ primer (SOLiD RNA Barcoding Kit Module 17-32) for the 3’ primer provided by the SOLiD Total RNA-Seq Kit. Each SOLiD 3’ barcode primer contains sequence required for SOLiD emulsion PCR, a unique identifying sequence, and an internal adapter sequence required for sequencing of the barcode. The barcoded samples were pooled in color-balanced multiples of 4 at equimolar concentrations for sequencing on the SOLiD platform.

**Emulsion PCR and sequencing of cDNA libraries**

When less than 20% of the amplified DNA is in the 25–200 bp range, you can proceed with the SOLiD™ System templated bead preparation stage, in which each library template is clonally amplified on SOLiD™ P1 DNA Beads by emulsion PCR. Refer to the *Applied Biosystems SOLiD™ 4 System Templated Bead Preparation Guide* (PN 4448378). Template beads were generated for sequencing using standard manufacturers’ protocols. Beads from the libraries were deposited onto a full slide with 8 control or treated library preps. Two separate sequencing runs were performed using
single end (50bp) and paired end sequencing (50bp/35bp) by the DNA Core facility at the University of Iowa using LifeTechnologies SOLiD System V4.

**Analysis of SOLiD RNA sequencing reads**

The RNA sequencing reads were mapped using the SOLiD Bioscope Whole Transcriptome Analysis Pipeline V1.1. The software algorithms were written specifically for SOLiD whole transcriptome sequencing. The reads were aligned to the *Daphnia pulex* genome sequence by the DNA Core Facility at the University of Iowa as part of the RNA sequencing service. The reads are filtered to remove repetitive sequence reads that map to more than 100 genomic locations, and then each read is mapped uniquely within the genome. The mate pairs are mapped separately but can still be identified as belonging to the same short read sequence. The output alignment files are generated for each specifically barcoded and sequenced sample. Differential gene expression analysis was performed by Don Gilbert using the single end read data set. The single end reads were normalized using DEseq to compensate for the differences in total reads per library. The gene expression calls were generated by adding the total reads across each gene prediction in each sample. Significant changes in expression were defined by a two-fold or greater change in read counts in the stress exposed samples as compared to the control sample. The files contain the mapping and paired end information from the RNA sequencing and can be viewed using any available sequencing alignment viewer.

**Viewing of mapped RNA sequencing reads**

The Integrative Genomics Viewer was used for visualization of the mapped RNA sequencing reads (30). For viewing of the RNA sequencing data set, the mate pair RNA sequencing data sets were loaded into IGV which allowed for the identification of the
sequence of each mapped read as well as the location of mapped mate pairs. The identification of mate pairs confirms regions of transcription that are connected as part of the same gene. The sequence of the reads allowed for the identification of exon-intron boundaries based on splice consensus sites.
REFERENCES


