Transcriptomic analysis of *Mysis diluviana* vertical migration

Matthew J. Wargo, Ph.D.[2]

[2] Dept of Microbiology & Molecular Genetics, University of Vermont Larner College of Medicine, Burlington, VT 0540

October 2020

ABSTRACT:

Mysis diluviana is a key species in the Great Lakes and many other large, deep lakes. *Mysis* links pelagic and deep benthic habitats through diel vertical migration (DVM), serving as a conduit for energy and nutrient transfer as an omnivorous predator and prey for both benthic and pelagic fish species. While much work has been done on *Mysis* in terms of ecology and population dynamics, molecular insights into *Mysis* DVM and high-content genetic markers to study their population genetics are lacking. Our pilot project focused on the genetics of *Mysis* DVM that will also address identification of genetic markers for this species. We collected *Mysis* from Lake Champlain and Lake Ontario and isolated RNA from the anterior portion (head) and DNA from the posterior portion (tails) from the Lake Champlain samples only- we had initially planned to do all but the resultant problems (described below) led to an alternate path. The head was chosen for transcriptomics due to the wider diversity in organs and tissues as well as containing all of the sensory organs. Initial RNA transcriptome sequencing via high-depth, short-read technology led to two realizations: (i) the *Mysis* transcriptome in the head is much larger and more diverse than we anticipated, and (ii) we could not use distantly related genomes (Pacific White Shrimp and Krill) as templates to analyze these RNA data. Thus, we embarked on a two-fold strategy to enable use of our RNA transcriptomic data. First, we used long-read direct RNA sequencing of mRNA isolated from total RNA to generate longer mRNA templates upon which the short-read sequences could be templated. Second, we used long-read sequencing to generate as much draft genome sequence from Mysis DNA as we could afford. Our long-read RNA sequencing led to the identification of 1,614 unique mRNAs, however, only 10% of those encode proteins with any similarity to those in publicly accessible genomes and, of those, only 12 encode proteins of known or well-predicted functions. Thus, though mapping our higher-depth, short-read sequencing onto these transcripts provided evidence for differential gene expression, none of the differentially expressed genes match any in public databases. Our DNA sequencing yielded over 15,000 contigs up to 71,000 bases in length, estimated to cover roughly 0.2% of the *Mysis* genomes (which is 6x larger than the human genome). Only ~5% of our mRNAs mapped to genomic contigs and there were many genomic reads that did not assemble into contigs. De novo annotation pipelines, splicing prediction, etc were attempted (each with a supercomputer runtime of 1-2 weeks) but did little to increase our

understanding of the differentially expressed genes. Further analysis would exceed our user time capacity on the UVM supercomputer (a single PSI-BLAST analysis with the *Mysis* mRNAs would take 3-4 weeks based on our priority level in these Covid-19 times). We came to the realization that our data were too large and the *Mysis* genome both too complex and too divergent from known genomes to allow completion of the original product. While the original transcriptomic analysis will not be completed in the near term, the genomic and transcriptomic data we generated will be fodder for the *Mysis* field and should provide new molecule information useable for those interested in *Mysis* genetics and ecology in the Great Lakes.