Transcriptome profile of the fathead minnow (*Pimephales promelas*)

Wentworth S. A.[1], Thede K.[2], Aravindabose V.[2], Monroe I.[1], Thompson A. W.[1], Garvin J.[2], Packer R.[1]

1. Department of Biological Sciences, Columbian College of Arts and Sciences, The George Washington University, Washington, DC
2. Department of Physiology and Biophysics, School of Medicine, Case Western Reserve University, Cleveland, Ohio

Introduction

Modern RNASeq technology provides a practical and accurate method for exploring the transcriptome and expressed proteins of non-model organisms. One such organism, the fathead minnow (*Pimephales promelas*), is a widely used toxicology model, however the lack of knowledge about its genetics has hindered further study into the interactions between this organism and its environment. In this study, we present the fully assembled and annotated gill transcriptome of the rosy red strain of *P. promelas* to establish it as a foundation for further genetic and physiological research.

Materials and Methods

RNA Extraction and Sequencing:

1. Fathead minnows were acclimatized to 25°C before tissue extraction.
2. Gill samples were perfused to clear blood prior to RNA extraction.
3. Stranded cDNA libraries were constructed from each RNA sample and sequenced by an Illumina HiSeq 2500.

Transcriptome Assembly:

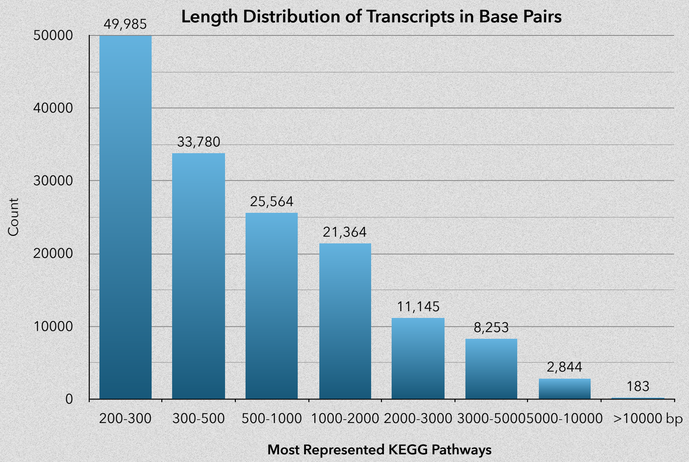
1. Quality control was performed with the fasts toolkit to trim bases that have been shown to be biased in Illumina sequencing[1].
2. Cleaned Reads were assembled *de novo* into transcripts using the Broad Institute's Trinity software package[2,3] using the default parameters.

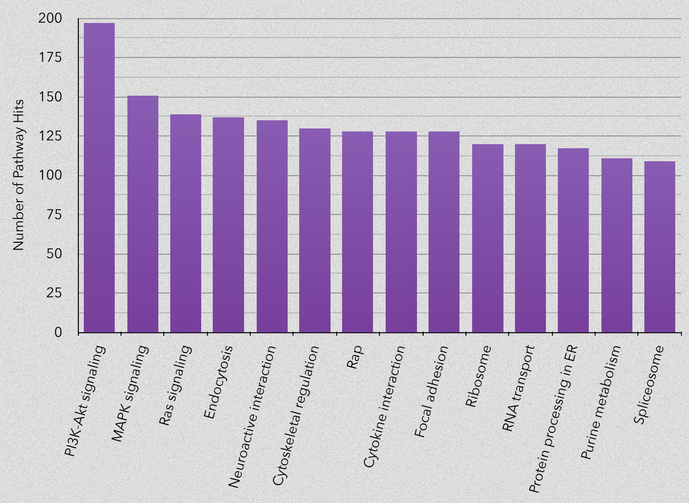
Transcriptome Annotation:

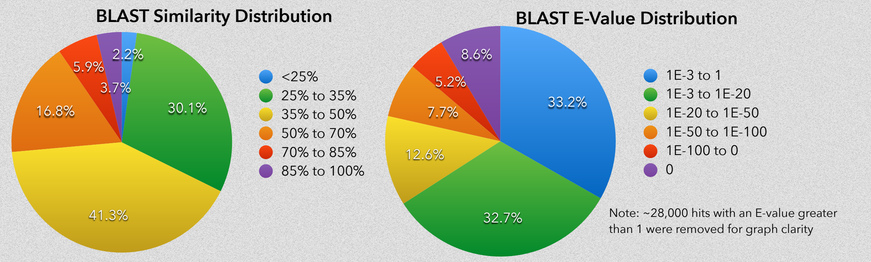
1. The assembled transcripts were annotated using the Broad Institute's Trinotate workflow[2,3].
2. The transcripts were BLASTed[4] against the Uniprot/SwissProt[5] database.
3. Further annotation was done using software such as Transdecoder[2,3], SignalP[6], tmhmm[7], and RNAMMER[9], which annotate other gene and transcript features.
4. Additional annotation using a WEGO level 2 gene ontology classification[9] and a Kyoto Encyclopedia of Genes and Genomes pathway classification[10] determined the categories of gene function and protein pathways represented in the transcriptome.

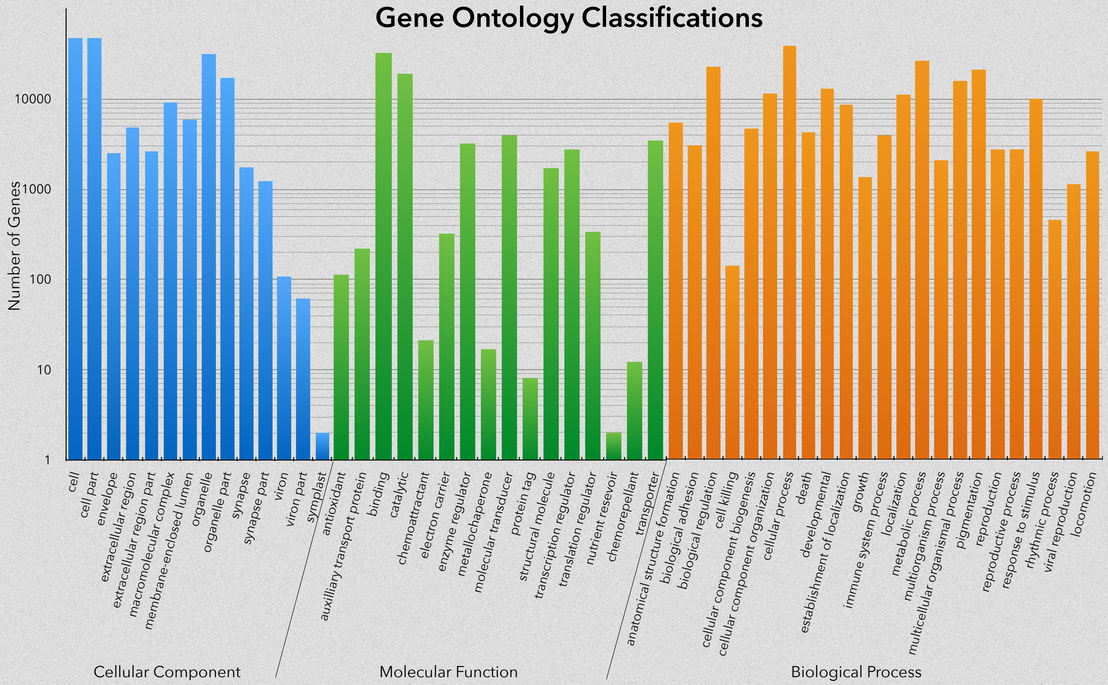
Summary of Transcriptome Data	
Total raw reads	470,813,940
Total cleaned reads	470,812,874
Length of reads(bp)	101
Total length of reads	47.5 Gbp
Total length of assembly	152 Mbp
Number of Transcripts	153,118
Transcript N50	2,271
Mean length of transcript	996
Median length of transcript	435
Transcripts annotated	150,773
Unique genes represented	72,334

Results









Conclusion and Discussion

Overall, the statistics of the assembled transcriptome denote that it is a high quality assembly. The coverage of the reads back to the transcriptome is ~270x, a coverage substantially higher than what is generally accepted (~30x)[11]. Furthermore, the N50 statistic–the contig length at which all longer contigs will comprise at least 50% of the total assembled bases of the transcriptome–at 2,271 bp is also higher than is normally found with other transcriptome profiles (more around ~1,500 bp).

It may be noted that the BLAST similarity distribution shows a majority of sequences having less than 50% identity with known gene sequences, however this is not surprising and can be explained. First, the transcriptome was only BLASTed against the UniProt/SwissProt database which, while very extensive and complete, is manually reviewed and annotated unlike the TrEMBL or nr database. This database is less likely to include sequences which are not extensively reviewed and may not provide hits for more recently sequenced and closely related organisms. Additionally, because *P. promelas* was BLASTed against a database containing a majority of mammalian genes, and the closest relative of the fathead minnow which has been extensively annotated is *Danio rerio* and, even though it is in the same family of *cyprinidae*, many regions may simply be evolutionarily homologous, yet not sufficiently identical to provide high similarity scores. Finally, while these similarity scores are not of the best quality, the E-values, even those removed from the above graph, show remarkable significance and quality.

Future Work

In continuation of our findings we would like to explore changes in the transcriptome for fathead minnows acclimatized to the extremes of seasonal water conditions (~5°C and ~25°C) to determine what changes in the expression of various genes with relation to changing environmental conditions. Additional explorations into a variety of environmental factors such as salinity or acidity could add to the larger picture, and could even lead to explanations of how the fathead minnow maintains homeostasis as environmental conditions change.

Acknowledgements

The authors would like to thank Adam Wong and Brian Haas for their technical support, Colonial One for providing the high-performance computing power required form this work, the Case Western Reserve University Genomics Core for the RNA sequencing, Dr. Tara Scully, and the Harlan Undergraduate Research Program for funding this project.

*References available upon request