In Silico Differential Gene Expression Analysis of a De Novo Transcriptome of the Prothoracic Ganglion of the Cricket (Gryllus Bimaculatus)

Student Researcher: Felicia Wang
Advisor: Hadley Horch
Bowdoin College
Department of Neuroscience

Pairwise comparisons were made between time points and programs. Lists of upregulated and downregulated genes were run through the blastx program (e-value cutoff of 1e-3) to identify proteins with high similarity to the query.

Results Obtained:

Determining outliers

Initially a multi-dimensional scaling plot was used to assess the presence of potential outliers and showed that samples 7C2 and 1C1 were apart of the central cluster of samples and may be outliers. In addition to this analysis, a depth of sequencing coverage was calculated for the contigs in each sample and compared for any differences. The results of these depth plots were that sample 7C2 appeared very different when compared to all the other samples due to its decreased density of contigs and the difference in the shape of the plot. Additionally, sample 1C1 was also determined to be visually different from the other samples. These analyses resulted in the determination that samples 7C2 and 1C1 would be considered outliers and removed from the differential analysis so as not to bias the data.

Multiple K-Mer Analysis

Individual transcriptomes were assembled at five different k-mer lengths. As the k-mer length used in the assembly increased, the N50 value, maximum contig length of the assembly, mean, and median also increased and the total number of Trinity “genes” decreased. The total number of contigs decreased from k=25 to k=32, however the total number of contigs for the k=21 assembly was the lowest (Table 1.). The GC content for all five assemblies were all around 40%, ranging from 40.14 to 40.94 and inversely correlated to the k-mer length (Table 1.). The contig length di.3 ( )-l 0 93 0 Td 6--0.8 (i)-3.2 (.3 ( )

differentially regulated genes, which were determined to have a higher likelihood of being truly differentially regulated due to two lines of evidence pointing towards them. In comparing the two programs, 2,099 genes were found to be upregulated at day one, 1,043 genes were found to be upregulated at day three, and 272 genes were found to be upregulated at day seven. Additionally, 180 genes were found to be downregulated at day one, 1,604 genes were found to be downregulated at day three, and 367 genes were found to be downregulated at day seven.

At this point, only the genes in the intersection between the two programs have been BLASTed. Not all the transcripts input into the BLAST program resulted in hits.

Significance and Interpretation of Results:

Determining outliers

This outlier determination was very important to proceeding with the differential analysis because the differential analysis is reliant on relative rates of expression. Therefore, if one subject presents as an outlier this could skew the data in the direction of the outlier and present biased results. Additionally, it was important to understand how the depth of sequencing varied from sample to sample because if one sample was sequenced at a greater depth, then the number of reads mapping back for

Figures/Charts

Table 1. Individual-K-mer Assembly Statistics

	k=21	k=25	k=27	k=30	k=32
Total # bases assembled	293,992,61	404,116,67	408,831,05	406,965,53	403,174,72
Total # assembled contigs	405,638	438,593	431,712	415,901	407,158
Total # Trinity "genes"	351,829	302,633	297,584	288,135	283,278