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

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

Abstract:

A unique compensatory growth behavior, in response to deafferentation, is observed in the prothoracic ganglion of the cricket, *Gryllus Bimaculatus*. However, the molecular pathways underlying this growth response are not well understood. Previous researchers have assembled a transcriptome of the prothoracic ganglion in order to determine the molecules involved in the compensatory response. A new transcriptome was assembled using multiple individual assemblies constructed at varying k-mer lengths to create a more representative data set. Two programs, EdgeR and DESeq, were used to perform the differential expression analysis on filtered and approved samples. These steps have improved the reliability of the data set and generated a listing of differentially regulated genes that can be further analyzed to assess a broader picture of the molecular basis of the observed neuronal plasticity.

Project Objectives:

Neural plasticity broadly is the ability for the nervous system to adapt to changes. This ability is extremely valuable when it comes to responding to injury because injury or damage to the central nervous system can have significant and potentially permanent effects. Most adult organisms, especially mammals, are limited in their capacity to adapt. The cricket is used as a model organism for the study of neuroplasticity because of its unique level of plasticity, even into adulthood (Horch et al., 2011). There is still little known about the molecular pathways that are involved in the neural plasticity of the cricket interneurons.

As an initial analysis of the molecular basis of the compensatory behavior, previous researchers assembled and analyzed a transcriptome. This transcriptome was mined for development guidance molecules and several were determined to be present in the transcriptome including slit, robo, ephrin, eph (Fisher et al., 2018). While the transcriptome can be searched for certain molecules, the next step in the analysis would be to determine molecules that change in expression levels as a result of deafferentation. However, in order to proceed with any differential expression analysis there were still many steps that could be taken in order to produce the most reliable set of differentially regulated genes. This research aimed to look into these extra analysis steps, such as filtering outliers and sequencing depth analysis, and then proceed with the differential expression calculations and analysis. The goal of this project was to further knowledge on the molecular basis of the compensatory growth behavior by understanding the differentially regulated genes after deafferentation and the broader functions of these genes in a molecular network underlying neuroplasticity.

Methodology Used:

Trinity (v.2.2.0) software was used to assemble 5 individual transcriptomes. K-mer lengths varied for each assembly (k = 21, 25, 27, 30, 32). The following parameters were used: library normalization with maximum read coverage 50, and RF strand specific read orientation, maximum memory, 250GB, and 32 CPUs. Evidential Gene was used to combine the five assemblies into one, non-redundant assembly. Raw reads were mapped back to the *de novo* transcriptome using Bowtie2. Sequencing depth was calculated with samtools. The mean and standard deviation of each contig was calculated and plotted to reveal differences. Differential expression was performed using two programs: EdgeR and DESeq. Counts were filtered to exclude contigs with less than one cpm in less than 2 samples.

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 not a part 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 distribution visually appeared similar across all the individual assemblies.

Mapping the raw reads from Illumina sequencing back against each assembly resulted in an overall alignment rate ranging from 98.58-98.74% for all assemblies. The percentage of reads that multi-mapped, mapped to more than one contig, was quite high for all assemblies, ranging from 90.04-93.33%. Both of these alignment statistics are relatively consistent across all assemblies (Table 1).

Combined Multiple K-Mer Assembly

After combining all five individual assemblies, a total of 2,099,002 potentially redundant contigs were present in the transcriptome. The main, okay set output from Evidential Gene contained 55,895 contigs and the alternate, okalt set output from the program contained 143,462. Both sets were combined for a final multiple k-mer assembly consisting of 199,357 contigs.

Differential Expression Analysis

The pairwise comparisons from the EdgeR program gave results on what genes were upregulated or downregulated at each of the three time points when comparing the deafferented group to the control group. It showed that there are 2,234 genes upregulated at one day post-deafferentation, 1,860 genes upregulated at three days post-deafferentation, and 290 genes upregulated at seven days post-deafferentation. Additionally, 261 genes are downregulated at one day post-deafferentation, 1,675 genes are downregulated three days post-deafferentation, and 580 genes are downregulated seven days post-deafferentation. A similar pairwise comparison was performed using the DESeq2 software and revealed that 3,589 genes are upregulated at one day post-deafferentation, 1,424 genes are upregulated at three days, and 535 genes are upregulated at seven days. There are 985 genes downregulated at one day, 3,049 genes downregulated at three days, and 448 genes downregulated at seven days.

After the differentially regulated genes were analyzed within each individual software, comparisons were made between the two programs. This was done to generate a primary set of 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 that sample could be higher, even though this is only a result of the sequencing depth. The sequencing depth across the remaining samples appeared similar and normalization for remaining differences was run as part of the differential expression analysis.

Combined Multiple K-Mer Assembly

While a single k-mer based assembly can be a representative transcriptome, many times a transcriptome assembly created from combining multiple k-mer assemblies is more representative because they capture more variability in the assembly process due to different assembly parameters (Mamrot et. al, 2017). By choosing only one k-mer assembly there is a risk of losing a set of more accurate contigs generated from another assembly. Therefore, in order to generate a more comprehensive transcriptome we chose to combine all five assemblies and subsequently use a program to reduce redundancies and fragments.

Differential Expression Analysis

Using the generated read counts matrix from the combined multiple k-mer assembly, the differential expression analysis was run using two programs in order to create a set of high confidence upregulated and downregulated genes. Both programs employ slightly different algorithms when deciding what genes are upregulated and downregulated (Evans et. al, 2016). The decision was made to use EdgeR and DESeq2 because these programs are commonly used in differential expression analysis of transcriptomic data for smaller sample sizes fitting the negative binomial distribution. The names of these proteins have been briefly examined and some predicted proteins that are developmental guidance molecules or from regenerative pathways, were noticed. However, further functional analysis of these proteins needs to be conducted before generalized statements can be made about the types of molecules upregulated and downregulated after deafferentation.

Figures/Charts:

Table 1. Individual K-Mer Assembly Statistics

	k=21	k=25	k=27	k=30	k=32
Total # bases assembled	293,992,611	404,116,670	408,831,054	406,965,539	403,174,726
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
Average contig length (bp)	724.77	921.39	946	978.52	990.22
Median contig length (bp)	376	397	397	399	400
Maximum contig length (bp)	37,575	44,287	44,328	44,352	44,331
N50 (bp)	1219	2000	2141	2272	2341
GC count for complete assembly (%)	40.94	40.39	40.32	40.19	40.14
Overall alignment	98.61	98.58	98.66	98.71	98.74
Reads mapped 1 time (%)	4.4	3.69	2.71	2.08	2.55
Reads mapped >1 time (%)	90.04	90.95	92.21	93.33	92.9

Acknowledgements and References:

I would like to thank Professor Horch for her continued support and guidance throughout this project. Additional thanks to Maeve, Harris, and Professor O'Brien for their previous and continuing work on this research.

Funded by Maine Space Grant Consortium

Evan C., Hardin J., Stoebel D.M. (2016). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefing in Bioinformatics*

Fisher H.P., Pascual M.G., Jimenez S.I., Michaelson D.A., Joncas C.R., Quenzer E.D., Christie A.E., Horch H.W. (2018). *De novo* assembly of a transcriptome for the cricket *Gryllus bimaculatus* prothoracic ganglion: An invertebrate model for investigating adult central nervous system compensatory plasticity. *PLoS ONE,* 13(7)

Horch, H. W., Sheldon, E., Cutting, C. C., Williams, C. R., Riker, D. M., Peckler, H. R., Sangal, R. B. (2011). Bilateral consequences of chronic unilateral deafferentation in the auditory system of the cricket gryllus bimaculatus. *Developmental Neuroscience, 33*(1), 21-37

Mamrot J., Legaie R., Ellery S.J., Wilson T., Seemann T., Powell D.R., Gardner D.K., Walker D.W., Temple-Smith P., Papenfuss A.T., Dickinson H. (2017). De novo transcriptome assembly for the spiny mouse (Acomys cahirinus). Scientific Reports 7:8996