Differential Gene Expression in *Holothuria glaberrima* using *Salmon* & *edgeR*

David John Ortiz Rivera
Computer Science Student
University of Puerto Rico, Rio Piedras Campus
david.ortiz11@upr.edu

Humberto Ortiz-Zuazaga
Computer Science Professor, Principal Investigator
University of Puerto Rico, Rio Piedras Campus
humberto.ortiz@upr.edu

July 28, 2017

Contents

List of Figures

1 Abstract

With a refined quantifying method we aim to find DGE using samples from the sea cucumber species; *Holothuria Glaberrima*. Samples for the species are taken from the uninjured organism and different stages of the regeneration process after evisceration. Among them we can find the eviscerated injured (*Day2, Day12, and Day20*), *Uninjured Normalized*, *non-normalized regenerating*, and *regenerating pooled normalized* groups.

MDS, *MA*, and *Volcano* plots were constructed to visualize and analyze DGE and the relation between groups and samples. The *MDS* plot clearly states the difference between the injured and the uninjured samples. During *MA* plot analysis, it was found that DGE among the injured groups possessed fairly uniform expression and that genes from *Day2* were slightly more expressed against *Uninjured Normalized* than the other 2 groups. This means that at the beginning of the regeneration process some genes tend to express themselves more in contrast to the later half of regeneration. *Volcano* plots showed that *Day2* and *Day12* had genes that possessed a higher level of significance than the genes from *Day20*, but overall. Additionally, the top 2 contigs from *Figure 6* showed a possible relationship with muscle development and regeneration.

DGE analysis presents the difference between injured and uninjured samples during the regeneration process of the *Holothuria Glaberrima* after evisceration. This difference in the expression of genes can be seen at the beginning of the regeneration process and even slightly at later stages, implying that there is a relation between the expressed genes and the regeneration process.

2 Introduction

Previous work in the field of Differential Gene Expression (DGE) analysis led us to believe that gene counting might not be best suited for our designed method [8]. Because gene counting goes through all of the genes in a sample to count them, we could reduce run-time whilst producing accurate results if we'd use a quantifying method. One which estimates the likely-hood of the genes in a sample, rendering gene counting unnecessary. The *Salmon* package, which features indexing, quasi-mapping, and inference algorithms such as *SCVB0* and *EM* [10], was selected to provide said method.

With our refined method we aim to find DGE using samples from the sea cucumber species; *Holothuria Glaberrima* (*Figure 1*). The organism is comprised of 12 sample files taken at different stages of the regeneration process after evisceration. Among them we can find the eviscerated injured (*Day2, Day12, and Day20*), *Uninjured Normalized*, *non-normalized regenerating*, and *regenerating pooled normalized* groups. More about the samples on [7].

(a) *Holothuria Glaberrima* [5]

Figure 1: *Holothuria Glaberrima*

3 Methods

The first 8 samples were obtained using *Illumina sequencing*, while the other 4 through *454 pyrosequencing*. It was decided that the latter would be left out of any type of visualization and analysis to avoid unexpected results produced by the different sequencing techniques. Since sample data for our previous project [8] wasn't trimmed nor normalized, we suspected that it was responsible for some of the unexpected results obtained. For that reason, trimming and normalizing were also prioritized for this project alongside quantifying. All files were trimmed, normalized, and assembled (single ended) using *Trimmomatic-0.36* [2], *khmer tools* [3], and *Trinity-v2.4.0* [6] respectively. *Salmon (0.8.2)* [10] was used for the indexing and the quantification process. For visualizing, the latest version of *edgeR (3.18.1)* [11]. A *Standard Nucleotide BLAST* was used to analyze some contigs [1]. All scripts used for this project are available on *GitHub* [9].

3.1 Quantifying & Visualizing

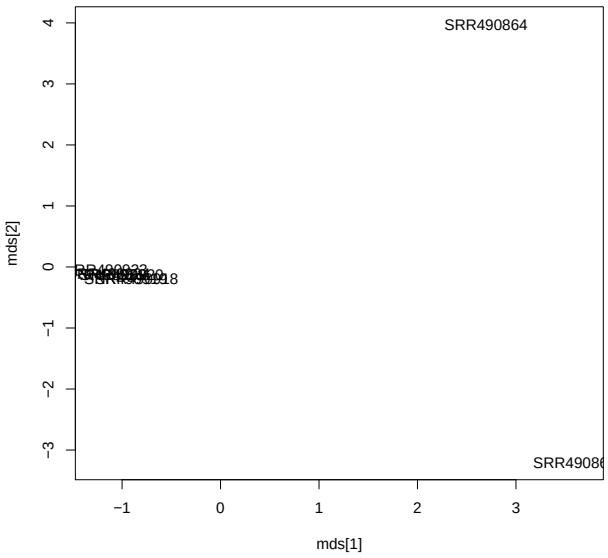
The generated transcript from *Trinity* was indexed; a hashing method that allows the mapping of reads to the transcript. Then, each individual sample was mapped to the indexed transcript to estimate the abundance of their genes using inference algorithms (*SCVB0 and EM*). This generated 12 sample count files. Only 8 will be used for visualizing and analyzing later on. The sample count files were divided into their respective groups. This resulted in only 4 groups: *Day2, Day12,* and *Day20* (injured) and *Uninjured Normalized*. We plot the different injured groups against the uninjured group in *edgeR* to produce *MDS* (1), *MA* (3), and *Volcano* plots (3). To identify some of the most expressed contigs, we created a top table using the same data used to model one of the volcano plots (*Figure 6*). Some of the contigs would later be analyzed using *BLAST*.

4 Results

We analyze the visualized data to find any relationship with the expressed genes and the regeneration process of the organism.

4.1 MDS plot

To view the similarities between the samples, a Multidimensional Scaling (MDS) plot was produced. *Figure 2* shows the uninjured group (samples SRR490864 and SRR490868) differing greatly from the injured group (samples SRR490919, SRR490918, SRR490921, SRR490920, SRR490923, and SRR490924) possibly due to the latter regenerating from evisceration.



(a) MDS plot from the samples in groups: Day2, Day12, Day20, and Uninjured Normalized. The injured group samples share similarities between themselves, but when compared to the uninjured group the differences are apparent.

Figure 2: MDS plot from Day2, Day12, Day20, and Uninjured Normalized

4.2 MA plots

Differential Gene Expression from *Day2 VS Uninjured*, *Day12 VS Uninjured*, and *Day20 VS Uninjured*, showed somewhat uniformly expressed genes. Looking at *Figures 3, 4, and 5* closely, we can see that: (a) At the beginning of the regeneration process some genes are expressed higher than at the later half of regeneration. (b) *Day2* showed the highest level of expressed genes out of all the groups.

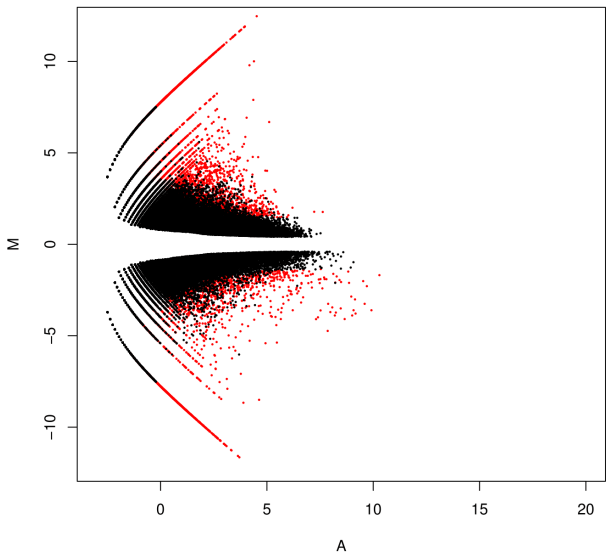


Figure 3: MA plot from Day2 VS Uninjured Normalized

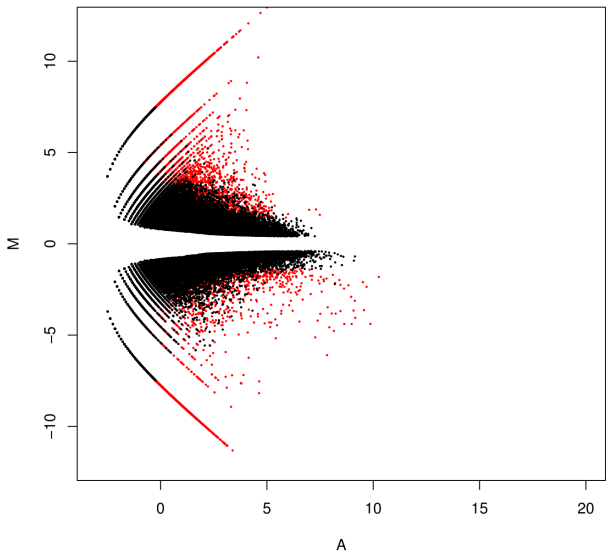


Figure 4: MA plot from Day12 VS Uninjured Normalized

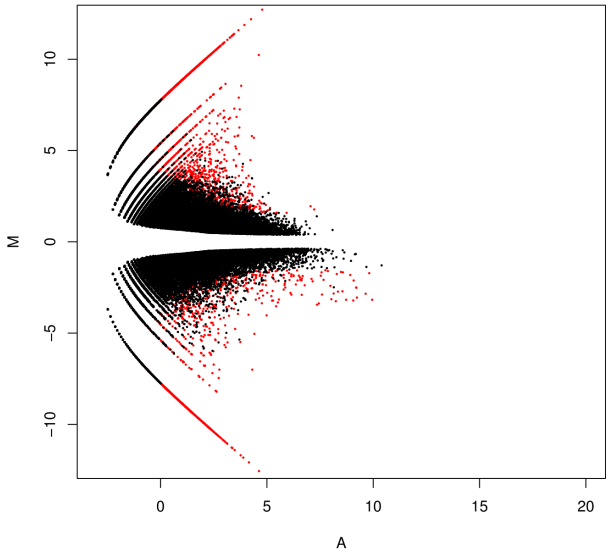


Figure 5: MA plot from Day20 VS Uninjured Normalized

4.3 Volcano plots

Day2 and *Day12* both depict a gene expressed above a *Log Odd* of 10, signaling the high level of significance of that gene. At first glance, *Day12* and *Day20* both give the impression that they posses a higher level of significance in their expressed genes when compared to *Day2*, but a closer look shows that both plots have a decrease in scale. Similarly to the MA plots, we see a decrease in the level of expression, in this case, a decrease in the level of significance at later stages in the regeneration process. Once again, signaling this relationship between the expressed genes and the regeneration process.

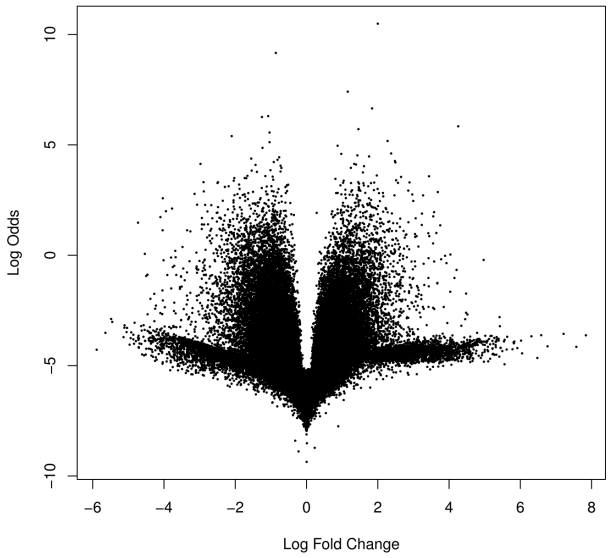


Figure 6: Volcano plot Day2 VS Uninjured Normalized

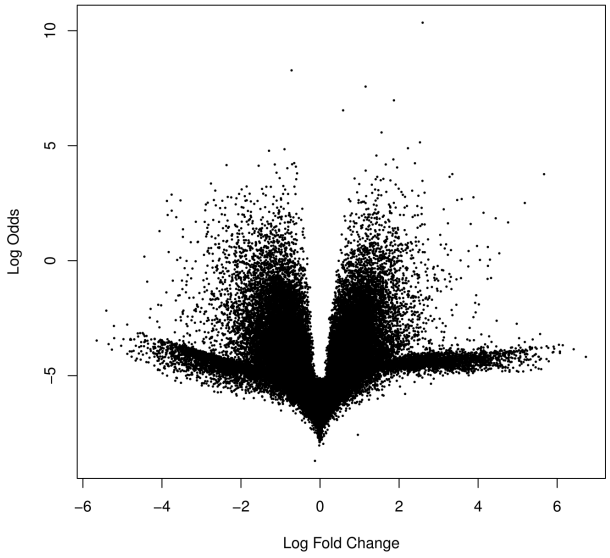


Figure 7: Volcano plot Day12 VS Uninjured Normalized

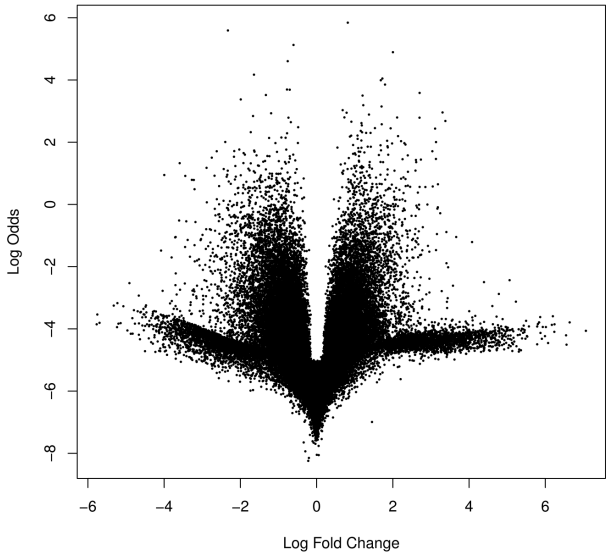


Figure 8: Volcano plot Day20 VS Uninjured Normalized

We take a closer look at some of the most expressed contigs from the *Day2 VS Uninjured Volcano* plot from *Figure 6* by creating a top table.

Day2 VS Uninjured						
Contig	logFC	AveExpr	t	P.Value	adj.P.Val	B
DN70006_c6_g1_i1	2.0012987	8.511831	87.82541	1.828483e-08	0.001709248	10.488470
DN77686_c3_g1_i1	-0.8626237	9.302421	-62.42930	8.509798e-08	0.003977437	9.162151
DN76771_c15_g4_i1	1.1594018	8.002912	40.61401	5.893981e-07	0.013774087	7.412668
DN72976_c4_g2_i2	1.8412865	8.082781	34.33991	1.253350e-06	0.020988837	6.649383
DN77579_c0_g1_i1	-1.0761521	8.373718	-32.65290	1.571710e-06	0.020988837	6.301126
DN77703_c1_g1_i2	-1.2513026	8.645890	-32.91845	1.515546e-06	0.020988837	6.258708
DN77672_c73_g1_i4	4.2589800	6.268616	41.89279	5.126781e-07	0.013774087	5.839361
DN62032_c1_g1_i6	1.4587754	7.383805	27.79526	3.239072e-06	0.032299067	5.710988
DN72988_c0_g2_i1	-1.0361479	7.547594	-26.98545	3.698467e-06	0.032299067	5.558072
DN77634_c1_g1_i5	-2.0997609	6.064940	-28.52180	2.884961e-06	0.032299067	5.396483

Table 1: Top table for *Day2 VS Uninjured*

4.4 BLAST

We used *NCBI's BLAST* to have a closer look at the some most expressed genes from *Table 1*. *DN70006_c6_g1_i1*'s hits were mostly related to myosin regulation. This might indicate that the sea cucumber needs myosin to form the muscles of its digestive system during its regeneration process. *DN77686_c3_g1_i1*'s hits were all related to titin-like, also thought to play a role in muscle development [4]. Note that it's possible for the hits of *DN70006_c6_g1_i1* to be related to something else since not all the hits are related to myosin regulation.

5 Conclusion

MDS, *MA*, and *Volcano* plots were constructed to visualize and analyze DGE and the relation between groups and samples. The *MDS* plot clearly states the difference between the injured and the uninjured samples. During *MA* plot analysis, it was found that DGE among the injured groups possessed fairly uniform expression and that genes from *Day2* were slightly more expressed against *Uninjured Normalized* than the other 2 groups. This means that at the beginning of the regeneration process some genes tend to express themselves more in contrast to the later half of regeneration. *Volcano* plots showed that *Day2* and *Day12* had genes that possessed a higher level of significance than the genes from *Day20*, but overall. Additionally, the top 2 contigs from *Figure 6* showed a possible relationship with muscle development and regeneration.

DGE analysis presents the difference between injured and uninjured samples dur-

ing the regeneration process of the *Holothuria Glaberrima* after evisceration. This difference in the expression of genes can be seen at the beginning of the regeneration process and even slightly at later stages, implying that there is a relation between the expressed genes and the regeneration process.

6 Future Work

Until now, only 8 of the 12 files have been used for visualization. These remaining 4 files are *SRR490772, SRR490752, SRR490669, and SRR490649*. *SRR490772, SRR490752, and SRR490669* are all from a non-normalized library, but they come to different days of the regeneration stage and *SRR490649* comes from a pooled normalized library. Because of this and the difference in sampling techniques, we are still not sure that we should visualize them. To assess what will be done with the remaining 4 files, further consultation is needed. It would be ideal to create a pipeline from this method. This would improve run time for this type of project somewhat, if set up correctly. Lastly, the use of this method on my cave-fish project [8] would improve run-time and generate more accurate results over its current one.

7 Acknowledgements

I would like to thank my PI, Humberto Ortiz-Zuazaga, Titus Brown and my lab partners Walter Baez, Kevin Legarreta, and Angel Sanquiche for helping with the development of the new method and the scripts. Special thanks to the developers of *Trinity, edgeR, Salmon, Trimmomatic, khmer tools. BLAST* for making this project possible and the University of Puerto Rico for the samples.

References

[1] G.-W. M. W. M. E. . L. D. Altschul, S.F. "basic local alignment search tool." j. mol. biol. 215:403-410., 1990.

[2] Bolger, A. M., Lohse, M., & Usadel, B. Trimmomatic: A flexible trimmer for illumina sequence data. bioinformatics, btu170, 2014.

[3] Crusoe MR, Alameldin HF, Awad S et al. The khmer software package: enabling efficient nucleotide sequence analysis [version 1; referees: 2 approved, 1 approved with reservations]. f1000research 2015, 4:900 (doi: 10.12688/f1000research.6924.1), 2015.

[4] EMBL-EBI. Titin-like domain (ipr023111), 2017.

[5] François Michonneau. Français: *Holothuria glaberrima*, 2012. [Online; accessed July 1st, 2017].

[6] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. nat protoc. 2013 aug;8(8):1494-512. open access in pmc doi: 10.1038/nprot.2013.084. epub 2013 jul 11. pubmed pmid: 23845962, 2013.

[7] NCBI. Experiment srp012442. national center for biotechnology information. u.s. national library of medicine, 2013. web. 29 july 2017.

[8] H. Ortiz, David; Ortiz-Zuazaga. Finding differential gene expression in cave and surface fish species using transcriptomic data obtained from trinity and edger. figshare. https://doi.org/10.6084/m9.figshare.4557115.v1, 2017.

[9] D. J. Ortiz-Rivera. (2017, july 29). kytrnd/bioinformatics: Differential gene expression. zenodo. http://doi.org/10.5281/zenodo.836113.

[10] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. nature methods., 2017.

[11] M. D. Robinson and G. K. McCarthy, Davis J & Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.