Gene expression in NSM neurons of C. elegans

Abstract

The brain of C. elegans is one of the more studied brains of any organism. However as data collection methods have improved so has the amount of data available. This paper provides an example of a way to parse large data sets on gene expression into coherent conclusions about the most interesting data points.

Introduction

# The worm C. elegans is one of the more intensely studied organisms on the planet. In fact its entire 302 neuron connectome has been completely mapped (Portegys, 2015.). Fortunately for C. elegans researchers however this connectome is not the whole story. While the connectome can be used for simulation projects such as the one located on <http://www.openworm.org/> , however it is not just the connections between neurons that matter, but the makeup of these neurons. In the paper Isolation of Specific Neurons from C. elegans Larvae for Gene Expression Profiling the authors demonstrated the effectiveness of sequencing a single neuron from a larval worm (Spencer, 2014). In this paper I will analyze Spencer et al’s dataset to identify genes with large differences in expression between NSM serotonergic and non-serotonergic neurons in C. elegans.

Methods

The dataset was acquired from Spencer et al and then trimmed down to just the gene identifiers and the RNA-seq values for the reference and NSM neurons. Log2 fold changes were then calculated for each gene with the equation log2(NSM) – log2(ref). The results were then filtered to pull out the most extreme log2 fold change values. Values above 9 and below -6 were used in order to keep both lists relatively short. All coding was done in Python 3.4 in an ipython notebook before it was migrated to bash shell scripts for the purposes of automation and replicability. A link to the github repo can be found here: https://github.com/milesing/Final\_Project

Results

This study found eleven genes with a log2 fold change above 9 and twenty genes with a log2 fold change of less than -6. Figure 1 shows that the majority of the data falls between \_\_\_\_\_.

Figure 1. Genes with high log2 fold change values.

|  | value\_1 | value\_2 | Log2\_Fold |
| --- | --- | --- | --- |
| gene |  |  |  |  |
| C27A12.12 | 16181 | 0.051161 | 34.41880 | 9.393923 |
| clec-179 | 18208 | 0.069678 | 67.56400 | 9.921325 |
| D2007.6 | 19002 | 0.097317 | 81.40110 | 9.708145 |
| F36A2.11 | 22031 | 0.004513 | 4.37214 | 9.920184 |
| F49E11.2 | 23177 | 0.319903 | 188.16300 | 9.200133 |
| fbxa-224 | 24443 | 0.007189 | 14010.80000 | 20.894321 |
| mir-52 | 28132 | 0.127244 | 193.86800 | 10.573261 |
| sls-2.1 | 31081 | 0.098528 | 96.20990 | 9.931429 |
| Y6G8.5 | 38141 | 0.242319 | 171.95300 | 9.470891 |
| ZK177.1 | 39051 | 0.040644 | 22.00700 | 9.080715 |
| ZK783.8 | 39443 | 0.385796 | 543.53000 | 10.460306 |

Figure 1.

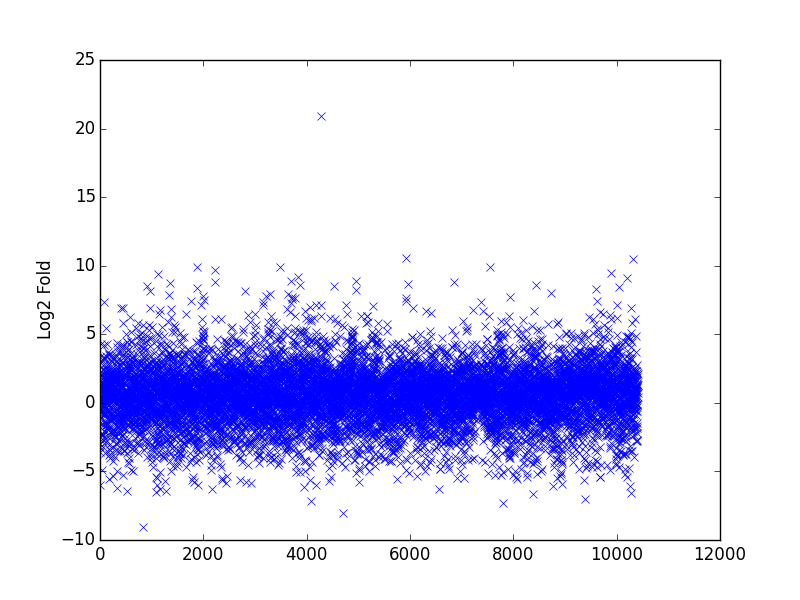
These genes had a log2 fold change value of 9 or higher, distinguishing them as the genes with the greatest increase in expression from the reference to the NSM neurons.

Figure 2. Genes with low log2 fold change values.

| Unnamed: 0 | value\_1 | value\_2 | Log2\_Fold |
| --- | --- | --- | --- |
| gene |  |  |  |  |
| 2L52.1 | 13348 | 12.73480 | 0.195868 | -6.022751 |
| B0222.5 | 13908 | 18.55230 | 0.246669 | -6.232878 |
| C01B10.6 | 14402 | 156.59700 | 1.840340 | -6.410940 |
| C12D12.1 | 15354 | 96.31820 | 0.184189 | -9.030478 |
| C25F9.2 | 16103 | 4.62872 | 0.052299 | -6.467689 |
| C26B9.5 | 16140 | 73.80010 | 1.150950 | -6.002726 |
| C29F3.7 | 16309 | 40.85260 | 0.519527 | -6.297085 |
| C34C6.7 | 16639 | 32.88400 | 0.374726 | -6.455406 |
| cyp-34A10 | 18855 | 12.92170 | 0.166725 | -6.276182 |
| F54C9.9 | 23516 | 7.33118 | 0.103820 | -6.141889 |
| F56D5.6 | 23861 | 24.91570 | 0.177469 | -7.133344 |
| F59B1.2 | 24202 | 204.95100 | 3.198350 | -6.001807 |
| gtl-1 | 25264 | 11.89400 | 0.044780 | -8.053151 |
| pbo-4 | 29311 | 16.41260 | 0.208723 | -6.297070 |
| sto-1 | 32492 | 30.21630 | 0.194596 | -7.278701 |
| T23F4.3 | 34179 | 25.37930 | 0.254125 | -6.641970 |
| tsp-10 | 34960 | 36.31690 | 0.549482 | -6.046425 |
| Y32F6A.5 | 36495 | 50.74580 | 0.402430 | -6.978407 |
| ZK550.6 | 39283 | 19.85400 | 0.303243 | -6.032812 |
| ZK6.8 | 39303 | 24.23710 | 0.254500 | -6.573408 |

Fig. 2. These genes had a log2 fold change value below -6. This indicates that they are the genes with the greatest expression in the reference neurons compared to the NSM neurons.

Figure 3. Distribution of log2 fold change values.



Discussion

Spencer et al used a quite broad sequencing method, sequencing tens of thousands of genes, not all of them are well known and studied, or even lightly known and studied. Of the genes with the most extreme log2 fold change values only the genes discussed in the rest of this section had information present about them from NCBI.

# The first interesting gene is mir-52, a member of a class of microRNAs that are expressed in the development of C. elegans(NC,2001). While it is not particularly well studied its greater presence in serotoninergic neurons suggests that they are either performing a regulatory function in the adult neuron, or like other microRNAs in their class, it may be involved in the differentiation and development of these neurons. The next noteworthy gene is cyp-34A10, which is not as present in NSM neurons as in the reference. This makes sense, as the Cytochrome P450 family is primarily involved in drug metabolism in the liver. Next we look at gtl-1 a Ca+ ion channel. Gtl-1 is also expressed more in the reference neuron than the NSM neuron. While this is an ion channel it is primarily associated with electrolyte homeostasis in the C. elegans intestine, which is coherent with its low expression in the NSM neurons (Teramoto, 2005). Lastly we have pbo-4, which is also expressed more in the reference than the NSM. This too makes sense as pbo-4, an ion pump, is primarily associated with muscle contraction.

Conclusion

This paper demonstrates one approach to deal with large gene expression datasets. A comprehensive gene expression analysis necessarily contains a mind boggling number of genes. However the use of python and similar tools allows a researcher to perform calculations on massive amounts of data, as well as allowing them to filter that data very specifically. Filtering the data can bring 36,000 data points down to the 30 most different. This type of data analysis is a powerful tool to understand the complex interactions of countless genes that go into even simpler brains.

References.

Thomas E. Portegys, Training sensory–motor behavior in the connectome of an artificial C. elegans, Neurocomputing, Volume 168, 30 November 2015, Pages 128-134, ISSN 0925-2312, http://dx.doi.org/10.1016/j.neucom.2015.06.007.

Teramoto, T., Lambie, E., & Iwasaki, K. (n.d.). Differential regulation of TRPM channels governs electrolyte homeostasis in the C. elegans intestine. *Cell Metabolism,* 343-354.

Lau, N. (2001). An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans. *Science,* 858-862.

Beg, A., Ernstrom, G., Nix, P., Davis, M., & Jorgensen, E. (2008). Protons Act as a Transmitter for Muscle Contraction in C. elegans. *Cell,* 149-160.