Automated Data Analysis Pipeline for Gene Expression Studies

Anton Morozov

March 14, 2016

Project advisors: Chris Fernandes (CS) and Steve Horton (Biology)
In collaboration with Columbia University, New York

Abstract

In any particular tissue cell type, only some genes of an organism's genome are active (expressed) at any given time. Identifying these active genes can help us understand what defines the cells identity. This is especially important in brain tissue, where neurons of different types have to interact in particular ways. Existing computational tools for gene expression analysis cannot effectively analyze data from poorly studied genomes, such as that of *Aplysia*, a model organism used in many neurobiological studies. To address this challenge and to find active genes that are associated with neuronal types, I have developed a computer pipeline suitable for processing both RNA-Seq and microarray gene expression data. The pipeline run requires little input from the user, and the results are presented via a web-based interface that accesses the projects database. Using this pipeline, I compared gene expression data derived from *Aplysia* motor and sensory neurons and found 185 genes and 24 gene pathways with significantly different steady-state levels. Many of these genes code for proteins that participate in cell interactions and signaling and reflect the functionality of the corresponding neurons. The developed pipeline can be used in a wide variety of projects dealing with gene expression analysis.

Contents

List of Figures

List of Tables

Terminology used in this paper

Aplysia Transcriptome at NCBI - a set of *Aplysia* transcripts predicted by the NCBI and Broad Institute bioinformatics teams based on the *Aplysia* genomic sequence and known genes.
Differentially expressed (DE) gene - a gene that has been found to have different activity levels in different samples.
GEM - gene expression matrix holding expression values for each gene in each sample.
Gene - a portion of the genome (with a particular location and boundaries) that codes for a product (usually a protein) that performs a particular function in a cell.
Gene expression - the level of gene activity measured by the frequency (abundance levels) of the corresponding transcript (RNA) found in the cell sample.
Gene Expression Microarray - A method for determining the relative abundance of various RNAs in a cell, using a different method from RNA-Seq. The RNA molecules are converted to DNA, and then again to RNA and are labeled by attaching a fluorescent dye. After that, RNA are hybridized to gene-specific probes attached to a glass slide. The hybridized RNAs are called targets. Measured color intensity is proportional to RNA abundance. Coordinates of the gene-specific areas on the microarray are known in advance.
Gene pathway - a network of interconnected genes whose products perform a particular shared function or role.
Genome - genetic information of an organism, stored as DNA.
Read mapping - uses sequence similarity analysis to determine which corresponding gene an RNA-Seq sequence read belongs to. The goal of read mapping is to determine the identity of the corresponding genes.
RNA Library - a set of amplified RNA fragments corresponding to the sample.
RNA-Seq - relatively new sequencing approach specifically developed to determine both the identity (corresponding gene) and the relative abundance of the RNA molecules found in a sample of cells. The read counts (numbers) in the RNA-Seq output are proportional to the original transcript abundances in the sample. Therefore, read counts can be used to determine transcript abundance and, hence estimate the transcriptional activity levels of the corresponding genes.
RPKM value a measure of a relative gene expression level. If RNA-Seq used as a source of RNA data, RPKM corresponds to Reads Per Kilobase per Million.
Sample - RNA extracted from a piece of tissue (many cells).
Sequence read - individual sequence produced from experiment using a sequencing machine, ranging from 40 to 1,500 nucleotides in length, depending upon the sequencing method.
Transcript - RNA sequence encoded by a gene (transcribed from the gene), may be a protein-coding (mRNA) or non-coding RNA.
Transcriptome - a collection of transcripts found in the cells of an organism.

1 Introduction

1.1 Importance of Gene Expression Studies in Neurons

Understanding how the nervous system performs its function is one of the greatest challenges in 21th century biology. Functioning of the brain's neuronal cells and the way those neurons communicate is influenced by the genes that are active (expressed) in the cells at any given moment. Currently not much is known about gene expression in neurons. By studying the differences in gene expression in various neuronal types we can obtain important insight into what determines the identity of a neuronal cell and makes it react in a certain way to drug treatments, diseases and signals from other neurons [1, 2, 3].

1.2 *Aplysia californica* as a Model Organism

The sea slug *Aplysia californica* is a popular model organism in many neurobiological studies due to its well-defined and studied neurons and its ability to demonstrate major neurological traits of interest, such as memory formation and learning [4]. The *Aplysia's* nervous system is relatively simple and contains only 10^4 nerve cells compared to about 10^{11} neurons of a mammalian brain [5]. *Aplysia's* neuronal cells are large (up to 0.5mm in diameter) and are easily identifiable for sample preparation.

Neurons of different types perform different functions and communicate by forming cell-to-cell connections (e.g. a sensory neuron linked to a motor neuron). This process is well-studied in *Aplysia*, therefore samples from two different *Aplysia* neuronal cell types, motor and sensory, were used in this study. Despite its status as a model organism, *Aplysia* has a poorly-studied genome and transcriptome (the set of known and predicted functional genes transcribed into RNAs) [6, 7, 5].

1.3 Methods of Generating RNA Data for Gene Expression Analysis

To study gene expression, RNA is first extracted from the appropriate cells, then identified and quantified. The fact that a particular RNA molecule was transcribed from a gene means that the corresponding gene that coded for that RNA is deemed active (expressed) in the cell under the studied conditions (Figure 1). Inactive genes on the other hand do not get transcribed and the RNA molecules encoded by them are not produced. The sequence of an RNA molecule corresponds to the sequence of the gene from which that RNA was transcribed. Thus, by identifying all of the various RNAs in a cell, one can determine which genes were active (and to what degree) in that cell at that time. A higher amount of RNA transcribed from a particular gene reflects the higher level of expression of that gene [8].

Currently, RNA-Seq and gene expression microarrays are the two most popular experimental approaches for obtaining RNA data for gene expression analysis. Both methods start with RNA extracted from a cell sample and produce data that can be further used by computer software to analyze gene expression in the sample.

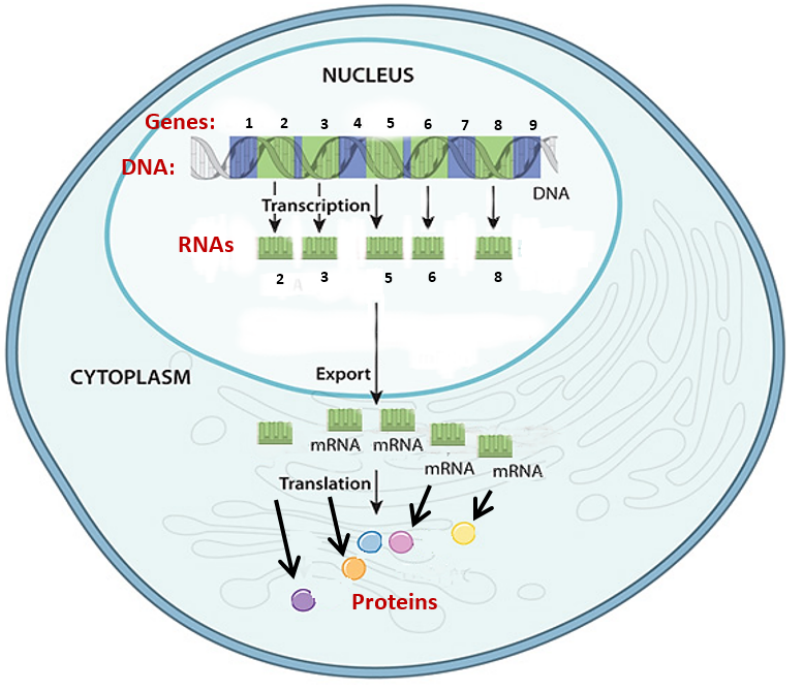


Figure 1: Gene Activity in Eukaryotic Cells. In any given cell, only some genes are active (shown in green) and are transcribed into mRNA which eventually are translated into proteins. The set of functional proteins (and therefore derived from the set of active genes) reflect the cells function and identity. Inactive genes are shown in blue.

RNA-Seq approach

In RNA-Seq, all RNAs extracted from the sample are sequenced producing millions of relatively short sequence reads ranging from 40 to 1500 nucleotides (nt) in length. In addition, RNA-Seq provides quality information for every read which allows estimating and filtering out of probable sequencing errors. After removing low quality reads, all of the remaining sequences need to be annotated (assigned gene names that they correspond to). If the organism's genome is well-studied, this can be easily done by comparing read sequences to a database of known genes that have been experimentally verified. However, such databases do not exist for *Aplysia*. In this case, gene name assignments are done by mapping the reads to predicted reference genes or RNAs transcribed from these genes (transcripts) using sequence homology search tools such as BOWTIE and BLAST [9, 10]. The mapping step significantly depends on the quality of the reference genome or transcriptome. The number of sequence reads that can be mapped to a particular gene corresponds to the gene's activity level and is used as a relative gene expression level value (RPKM value). Values of RPKM for multiple samples are usually presented in a table known as the Gene Expression Matrix (GEM). The GEM table contains values that characterize gene expression levels in each sample under study. This table is the starting point for further analysis to compare gene activity in several samples and identify those genes that have different activity levels in different samples (Differentially Expressed, or DE genes).

Gene expression microarrays

In microarrays, 60-nucleotide probes (which are complementary to specific RNAs) are attached to a glass slide. The corresponding RNAs are called targets. Each spot on the slide is 25 microns in diameter and contains several thousand probes specific to the same RNA. Since the probes are designed to match the sequence of a particular known gene, there is no need to map any sequences to the transcriptome, which simplifies the analysis. All RNAs extracted from a cell are used as templates to synthesize corresponding DNA, which is converted back to RNA, and then labeled with a fluorescent dye, so if a labeled molecule binds to a complementary probe on the array, it is read by the detector in the machine after the microarray is subjected to excitations by multiple lasers. Due to the array design, it is known which gene each fluorescent spot corresponds to. The spot intensity is then measured by the microarray scanner and is corrected for background noise. The instrument's software outputs the list of the corrected intensity values for each gene probe. The resulting normalized signal intensity for each gene reflects the quantity of the corresponding RNA in the sample and can then be used as RPKM values.

Both RNA-Seq and microarrays have some advantages and disadvantages [11]. In particular, RNA-Seq is considered more sensitive and suitable for finding genes with extremely low activity levels and for the discovery of new genes, which cannot be done by microarrays. On the other hand, the microarray approach is much cheaper and since it is an older method, its errors and biases are better understood compared to RNA-Seq. For most gene expression studies, microarray sensitivity is considered high enough to be usable. When choosing the method, it also should be considered that RNA-Seq data is harder to analyze than microarray data [11]. However, both methods are valid for use in gene expression studies and the choice depends on the available financial resources and on whether it is important for the study to identify genes with very low expression levels.

The developed Pipeline is able to accept both RNA-Seq and microarray datasets as sources of RNA data.

From gene expression values to pathways and function

Gene expression analysis starts with evaluating the expression levels of individual genes obtained by the processing the RNA-Seq or microarray data described above [12]. It then focuses on groups and networks of genes that are involved in the same function (gene pathways). If an identified pathway contains many DE genes, chances are that such a DE pathway is especially important for the cell function(s) under study. Recently many methods and tools for analysis of gene pathways have become available [1, 13, 14], including the DAVID system (https://david.ncifcrf.gov/). The DAVID functional annotation tool is one of the most useful pathway analysis and visualization systems since it collects information from several pathway databases, has a convenient interface and allows the user to change the pathway search parameters as needed [15]. The Pipeline presented in this report uses DAVID to search for pathways associated with the identified DE genes and to create the visualization of the found pathways to aid in interpreting the results of the study.

Aplysia Genome and Transcriptome

The current *Aplysia* genomic sequence (`http://www.ncbi.nlm.nih.gov/assembly/GCF_000002075.1`) is incomplete and consists of 164,545 pieces with many un-sequenced gaps of unknown length between them. The exact gene number and boundaries of many genes in the genome are not known [6, 7].

To address the high interest of the scientific community in *Aplysia* genetic information, in 2015 its genomic sequences were annotated by the NCBI bioinformatics team using their Eukaryotic Genome Annotation Pipeline, an automated system that annotates (identifies and assigns names to) genes, transcripts and proteins derived from genomes. The pipeline gene prediction software used the genomic sequences and took into account the predicted homology to known genes and proteins in other organisms. It also took into account existing experimental evidence for genes in that organism. These efforts resulted in the creation of the *Aplysia transcriptome* dataset which contains predicted gene sequences along with their names [16]. NCBI *Aplysia* genome annotation results are presented in the annotation report at `http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Aplysia_californica/101/`.Thus, the transcriptome from NCBI is the product of an in-depth bioinformatics analysis of the existing genomic sequence and experimental data for Aplysia and is the most comprehensive collection of *Aplysia* transcripts currently available. This was the main reason for selecting the *Aplysia* transcriptome and not genomic sequence as a reference to map RNA-Seq reads and obtain the corresponding gene names in this project.

Mapping to a transcriptome instead of a genomic sequences has additional advantages due to the transcriptomes lower complexity. Unlike a genomic sequence, a transcriptome does not contain long non-coding sequences between genes (intergenic regions) and within genes (introns) which can comprise up to 90% of a genomic sequence [17, 10]. As a result, less computational time is needed to map millions of sequence reads produced by RNA-Seq.

It is understood that, since *Aplysia* transcriptome dataset constructed by NCBI contains mostly predicted genes, some genes may be under-predicted (they exist in the genome but absent from the transcriptome). If this projects task were to discover all genes that are expressed but not yet characterized in *Aplysia*, the genomic sequence would have been the first choice to map the reads. But, for such a study, the experimental portion of the project (sample preparation for sequencing and sequencing approaches) would have to be quite different compared to those utilized in the study described in this report. The sequencing would have been done to produce longer reads with higher sequencing depth (or coverage, the number of reads covering the same nucleotide position in the genome). Then, since the noncoding sequences comprise over 90% of an animal genome [17, 10], mapping millions of reads to a genomic sequence would take much more computational time, even if only those reads that were not found in the transcriptome were mapped to the genome. But the real challenge would be to interpret the mapping results. Since the time when the initial genomic sequence was obtained in 2006 [18], more than 8-year-long efforts of the Broad Institute (MIT) and NCBI bioinformatics teams could not find any sequence homology to known genes in other organisms for genomic sequences other than those included in the current *Aplysia transcriptome*. Considering this, the most information that would be possible to derive from such mapping to the genome would be that a read matches some sequence in a genomic contig $\#X$. Any interpretation of this fact that would help understand the importance of having this presumable gene expressed more in one neuronal type vs another would still be not possible. This project's task was to examine gene expression, accepting the fact that the results would be limited to the list of those genes in *Aplysia* for which there exist some information on their functions and pathways they are involved in. Therefore, the *Aplysia* transcriptome was sufficient to use as a reference for mapping. In the future, when more genes are annotated in the *Aplysia* genome and included into the transcriptome database, it will be still possible to use the same RNA-Seq dataset and re-run the Pipeline, to obtain more information on the differentially expressed genes.

1.4 Project Rationale and Objectives

There exist several software suites and pipelines for gene expression analysis in well sequenced and studied organisms. For such organisms there exist well-established and annotated genome and transcriptome databases incorporated into commercially available analysis software. The situation is different for less-studied genomes such as that of *Aplysia*. There is currently a lack of appropriate tools that are capable of efficiently processing data from less annotated genomes, [1, 2]. The databases for these organisms are constantly changing and the new ones that appear often have unique formats.

Considering the large interest in studying gene expression in *Aplysia*, there is a need for a software tool that can perform gene expression analysis even when a comprehensive gene annotation database is lacking and available reference databases are incomplete and contain many imperfect sequences. This tool should be easily customizable to accommodate a variety of changing databases and user preferences and would replace the need for many manual steps in analyzing gene expression. My pipeline was designed to address the challenge of analyzing gene expression in poorly annotated genomes in a user-friendly and easily customizable manner, combining several analysis tools in a single system.

The Project also had a research objective – to analyze biological data of gene activity in *Aplysia* neurons. The objectives were to:

- identify active genes in two types of neurons;
- identify DE genes that are associated with the neuronal cells identity by comparing active genes in the samples;
- identify DE gene pathways to find out which functions differ in motor vs sensory neurons.

The hypothesis was that gene expression patterns are different in various types of neuronal cells and the set of expressed genes reflects the cells identity (neuronal identity). The main biological question to be addressed in the project was: Which gene activities correlate with identity of the specific types of neuronal cells? The results of the analysis should be helpful for understanding the functioning of *Aplysia*s nervous system.

2 Project Design

2.1 Design Requirements

The pipeline should be able to:

1. accept both RNA-Seq and gene expression microarray data;
2. efficiently process poorly annotated transcriptome data;
3. create a single SQL Database repository for multiple projects;
4. provide a user-friendly web interface for the data;
5. require little manual input from the user.

2.2 Pipeline Overview

The Pipeline was developed as a set of tools connected and run in a chain-like manner, where the output of the previous step was the input of the next. The Pipeline significantly facilitated computations by eliminating the need of manual input at each step of the process. The results were stored in an SQL Database. The stored information included cell and tissue source information, treatment types and sequence data along with the results of the analysis pipeline. Access to the Database was provided through the web interface utilizing HTML, CSS, PHP, SQL and Python scripts. The web interface was used to upload, process, organize and access the data. Graphical visualization of the results was also provided.

2.2.1 Pipeline Part 1

The major goal of Part 1 of the Pipeline was to process individual sample data and prepare gene expression values for sample comparison which would be done by Part 2 of the Pipeline. Pipeline Part 1 processed RNA-Seq and microarray data differently (Figure 2). Unlike microarray tables that could be used as input from the start, RNA-Seq data had to be processed, involved several steps that are presented below. This is necessary when analysis is done for organisms with poorly annotated genomes.

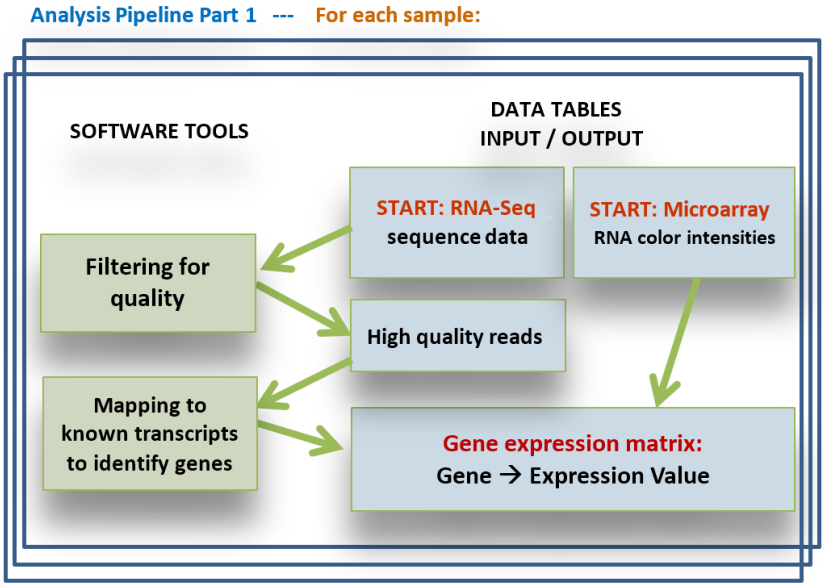


Figure 2: Analysis Pipeline Part 1. The main output of Part 1 is the gene expression table (matrix) for each sample.

Step 1. RNA-Seq input data format and sequence quality information:

The FASTQ format is the most widely used output format in high-throughput sequencers such as the Illumina HiSeq. FASTQ files are text files and have four lines for each sequence read in the file:

```
line1: @ Sequence name (the line always starts with @ symbol)
line2:  Sequence
line 3: + Optional line  (the line always starts with + symbol)
line 4:  Quality codes   (one symbol for every nucleotide)

Shown below is an example of a FASTQ file that has two reads in it:

@HWI-ST1383:51:D1LFRACXX:4:1101:1142:2126 1:N:0:
GCCTTAATTTGTGTGTTAGTTCCCTTTGTGAAGGTTATTTTTGTAACGGTCGTTCAGATTTGTCCTTC
+
FHHHHBHCFHHFFFCCGEIFFHIGC<CBGGBGHGBFAEH<FDDD>>?B;7;FABDHB>C)7;?B3??@
@HWI-ST1383:51:D1LFRACXX:4:1101:1227:2136 1:N:0:
CTAAACGCTTTTTGTGGAGCAGCGAGAAAATTAGACACGGGTGGAGAAAAAACTTCATTCATTACCCC
+
FHHHHFIIBGGEGGHGDHIEIIIIIFFGGIIEDBFGGHGIH@BEFBC>CABBB3;;AACDDA>@A###
```

A quality score for each nucleotide is determined during the sequencing run and corresponds to the probability of having an error in that particular position in the sequence. In the Illumina 1.8 sequencing machine output file, sequence quality at each nucleotide position is encoded by ASCII symbols from # to J (Supplementary Table S1). This quality score can be used to calculate the number of expected errors in a given read as the sum of the probabilities of an error for each nucleotide as described in [19] and implemented in the pipeline. While not being absolutely accurate, this quality measure provides a good and simple estimation of the number of errors in a read. A more exact calculation of probabilities of errors would require much more intensive computations, since every sequence run produces millions of reads. Then reads with more than 2 expected errors were eliminated based on published recommendations [20].

Step 2. Adapter trimming:

Adapters are artificial sequences attached to every RNA fragment during sample preparation for sequencing. These adapter sequences are part of the sequence read produced by the sequencer and have to be identified and removed before further analysis can be done. Adapters may be of variable lengths and may contain sequencing errors, which makes it a challenge to properly recognize and remove them. The CUTADAPT program [21] is a widely used software tool which serves this exact purpose. CUTADAPT is an open source program written in Python and can be easily adjusted to accommodate changes in sequencing technologies. CUTADAPT finds and cuts the known adapter sequences off the sequence read.

Regardless of the level of sophistication of the adapter trimming software, if the sequenced portion of the adapter within a read is shorter than 3 nt, the program cannot recognize it as an adapter and fails to trim it off. This causes false nucleotides to be included in the final sequence. This will be accounted for during the read mapping procedure.

Step 3. Read mapping to the *Aplysia* reference transcriptome:

The RNA-Seq reads were mapped to the NCBI transcriptome used as a reference [22]. The current *Aplysia* transcriptome database (version of July 8, 2015) contains 28,849 transcripts ranging in length from 99 to 42,283 nucleotides with an average of 2964 and a mode of 1086 nucleotides (Figure 3). The transcriptome is available via NCBI FTP site: ftp://ftp.ncbi.nih.gov/genomes/Aplysia_californica.

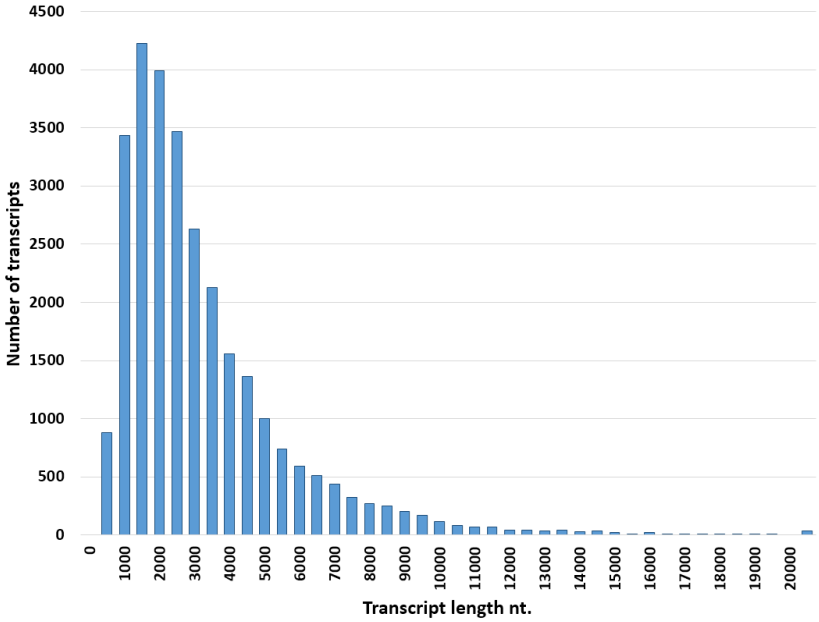


Figure 3: Distribution of the *Aplysia californica* transcripts by length. Data source: [22].

Reads were aligned to the reference Aplysia transcriptome using the BOWTIE tool [23]. BOWTIE was designed to perform fast and memory-efficient mapping of large numbers of short nucleotide sequences to a reference sequence (such as a genome or transcriptome). The reference sequence has to be formatted and indexed before the search. The necessary indexes were created by BOWTIE-build using Burrow-Wheeler transformation (BWT) and FM index.

Burrow-Wheeler transformation of a given string T (called BWT(T)) is done by rotating the strings elements as illustrated in Figure 4 (a). Simultaneously, BOWTIE-build creates the FM index which contains the positional information for BWT(T). All sorted permutations of T are called the Burrow-Wheeler Matrix (BWM), and the last column in the BWM corresponds to the BWT transformation, or BWT(T). The first (F) and last columns (L) of the BWM are used to perform a fast search of a small pattern within the reference sequences, as shown in Figure 4(c). This procedure is also known as an FL search. There is no need to keep any columns other than F and L, and because the first column is sorted, it can be stored as just the counts of each symbol. In this case, the memory needed is of the size of the alphabet (which is 4 for nucleotide sequences). The last column in many cases can be efficiently compressed by replacing the stretches of repeating symbols by the symbol itself and the length of the repeat, making memory use very efficient. The number of search steps equals to the length of the

pattern to search which in our case, is a sequence read and therefore is rather short. Indexing of the reference transcriptome, combined with the relatively small length of the reads that have to be mapped to the transcriptome, helps to make the mapping procedure fast and efficient.

The result of the FL search is a range of lines in the BWM, starting with the pattern (a read) we are searching for, but not the position of the found pattern in the original reference sequence T. This location information is provided by the FM index. When we know the position of the pattern in the BWM, we immediately know all of the mapping positions from the FM index.

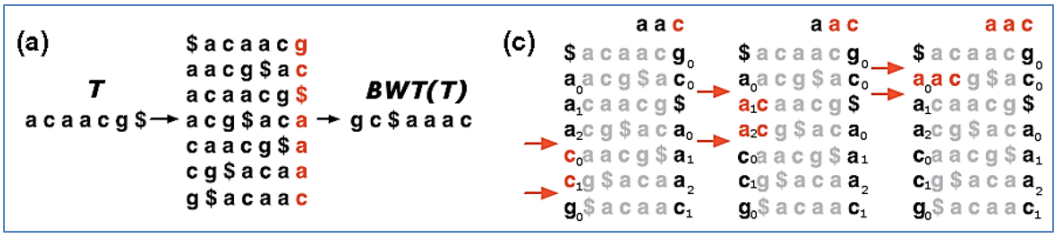


Figure 4: (a) BWT transformation, (c) FL search. (Figure adapted from Langmead et al. [23].

Computational complexity of the BOWTIE algorithm is $O(n)$, where n is number of reads. Since FL search is using the Burrow-Wheeler transformation, it does not depend on the size of the reference genome, and its complexity is $O(n)$ [20]. Memory requirements of mapping reads to a transcriptome are also $O(n)$, which is relatively low.

Unlike mapping, the procedure of indexing the reference transcriptome (which must be done before mapping can begin) has a much higher complexity of $O(Lg * La * N)$, where Lg is the total transcriptome size, La is the average size of reference sequences and N is the number of reference sequences in the transcriptome [20]. Despite high complexity, indexing will not slow down the pipeline since indexing needs to be done only once when preparing the reference transcriptome. Therefore read mapping to an indexed reference transcriptome is a relatively fast procedure.

In the study, all libraries were mapped to the reference transcriptome with a maximum of two allowed mismatches [20]. In addition to assigning a gene name, the mapping procedure eliminated sequences of poor quality that were missed at the sequence quality control step, as well as adapter fragments that had not been eliminated by the adapter trimming procedure. The percentage of reads lost at the mapping step due to unremoved adapter is very small (~0.5%).

Step 4. Calculating RPKM and GEM:

The most popular measure for gene expression is an RPKM value (Reads Per Kilobase per Million), calculated according to Mortazavi [24]:

$$RPKM = 10^9 * N/(L * S)$$

where N is number of reads mapped to particular transcript, L is the length of same transcript in base pairs and S is number of reads in the sample; sample; coefficient 10^9 was used, since the length should be in kilobytes and the number of reads should be in millions.

Values of RPKM for multiple samples are usually presented in a table known as the Gene Expression Matrix (GEM) (see also section 1.3). The GEM table contains values that characterize gene expression levels in each sample under study. This table is the starting point for further analysis of differential gene expression presented below. Calculating RPKM values and constructing the GEM table concluded the processing of the RNA-Seq data by Pipeline Part 1.

Unlike in RNA-Seq, the identity (the corresponding gene name) of each RNA captured on the microarray is known. In case of microarray data, the only processing that had to be done by Pipeline Part 1 was to read the data containing corrected intensity values for each gene. The output result of Part 1 was a gene expression table for each of the samples that was then stored in the SQL database. The subsequent analysis (done by Pipeline Part 2) was the same regardless of whether the gene expression data originated from RNA-Seq or microarrays.

2.2.2 Pipeline Part 2

The goal of this portion of the Pipeline was to compare samples in order to obtain the list of genes that are differentially expressed (DE) (Figure 5). For this purpose, gene expression matrices for all of the samples were combined into the GEM table. The GEM table contained all the RPKM values for all the genes (as rows) in all the samples (as columns) without any averaging of the values. We will now explain how we utilized the GEM table to compare the datasets using the t-test followed by the Bonferroni correction according to [25].

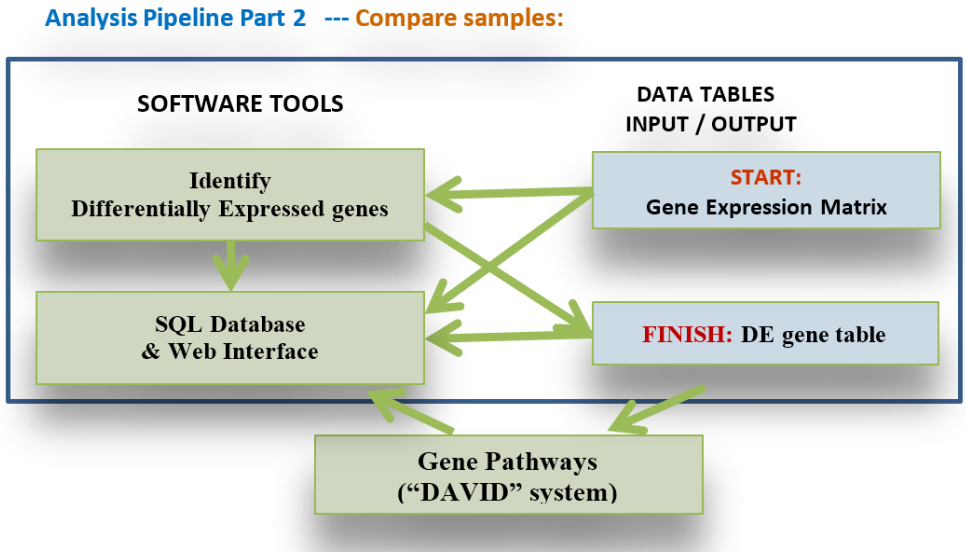


Figure 5: Analysis Pipeline Part 2. The result is a list of DE genes and pathways that change activity from sample to sample.

A t-test is a statistical method that helps evaluate the probability that the means of the two sets of data are the same. The method estimates a probability of getting the observed ratio between the-difference-of-the-means and their-standard-deviations. The smaller the probability, the less likely it is that the two sets of measurements are the same. In this study, the t-test was

applied to each gene separately. For each gene, two sets of values were compared: Set 1 (for example, genes expression in sample replicates collected from motor neurons) vs Set 2 (e.g., those from sensory neurons). Since each genes expression can change between the neurons in both directions (positive or negative), the p-values were estimated for two-tailed t-test.

The t-test was done for each gene with available microarray data, to compare this genes expression in the two types on neurons. But to evaluate statistical significance of the differences in expression for a particular gene, one needs to also take into account the fact that this difference was observed on the background of other genes also changing their expression between the two neurons. In this study, to compare the neuronal types, the t-test was done as many times as there were gene probes on the microarray ($\sim 10,000$). It is known that such multiple applications of the t-tests in the same analysis affect the calculations in such a way that the chosen probability threshold (the p-value cut-off) has to be corrected using Bonferroni approach explained below.

When the difference between two features is claimed to be statistically significant, this implies that the probability (p-value) of getting the observed result (the t-statistic in this particular case) simply by chance (under the assumption that there is no difference between the means) is low. This probability can be set as 0.05 when comparing datasets based on a single gene, but when two genes are compared in the same analysis, the chances of observing the differences between the datasets increase with each additional gene (test). If expression of the genes is independent from each other (meaning that some expression level of a gene is not a prerequisite for a certain expression level of another gene, which is correct in most cases, e.g., 8% of human genes code for factors regulating transcription [26]), the probability of seeing a difference between the two datasets using two gene pairs will increase two times, for three genes three times and so on. To maintain the low p-value in multiple t-tests, the probability cut-off of 0.05 is divided by the number of tests (genes in this case). This approach is known as Bonferroni correction for multiple tests.

As a result, the Pipeline Part 2 produced the list of DE genes for the two datasets. To interpret the significance of these genes having different activity levels in motor and sensory neurons, one needs to find which cellular functions these genes are involved in. Databases for functional pathways do not exist for *Aplysia* but available for humans, therefore, the list of human genes that correspond to the identified *Aplysia* DE genes was submitted to the DAVID system to search for pathways.

2.2.3 SQL Database, User Interface and Visualization

The Database contains information on the project data, both input by the user (such as tissue sample sources and their treatments) and produced by the Pipeline (sample sequence data, gene expression values and lists of DE genes and pathways) (Figure 6).

The website affords the uploading of either RNA-Seq or microarray data files (Supplementary Figure S1). The data file is then sent to the server where it is added to the table containing the sample data (the Sample Table). Since the Sample Table has information pertaining to the study description, dates and names, the user is able to document all of that data before the Submit button is pressed for file submission. After the file submission, the user will have the option of conducting DE analysis on the samples that are present in the Database.

The results of DE analysis are presented as tables and a heatmap used to visualize and analyze similarities in gene expression for multiple samples. A heatmap is a table in which each row corresponds to a single gene and columns represent different samples. The cells in the table are colored according to the corresponding z-values (or z-scores) in the matrix. Z-score value for a particular gene in a particular sample is calculated as the difference between the

average expression value for that gene in all the samples and the value for the gene in that particular sample, divided by the standard deviation.

In this project, the heatmap was constructed by the R function *heatmap*.2. This function used the GEM table as an input, based on which it performed hierarchical clustering of the samples. Hierarchical clustering grouped the samples in such a way that overall similarity in gene expression within each cluster was higher than between different clusters. The programs output contained the resulting dendrogram (a tree-like structure) in which closely positioned branches represented similar samples. The produced heatmap showed the entire data matrix rearranged according to the clustering. Since hierarchical clustering was done based solely on the expression data, the dendrogram can be used to make inferences about the samples.

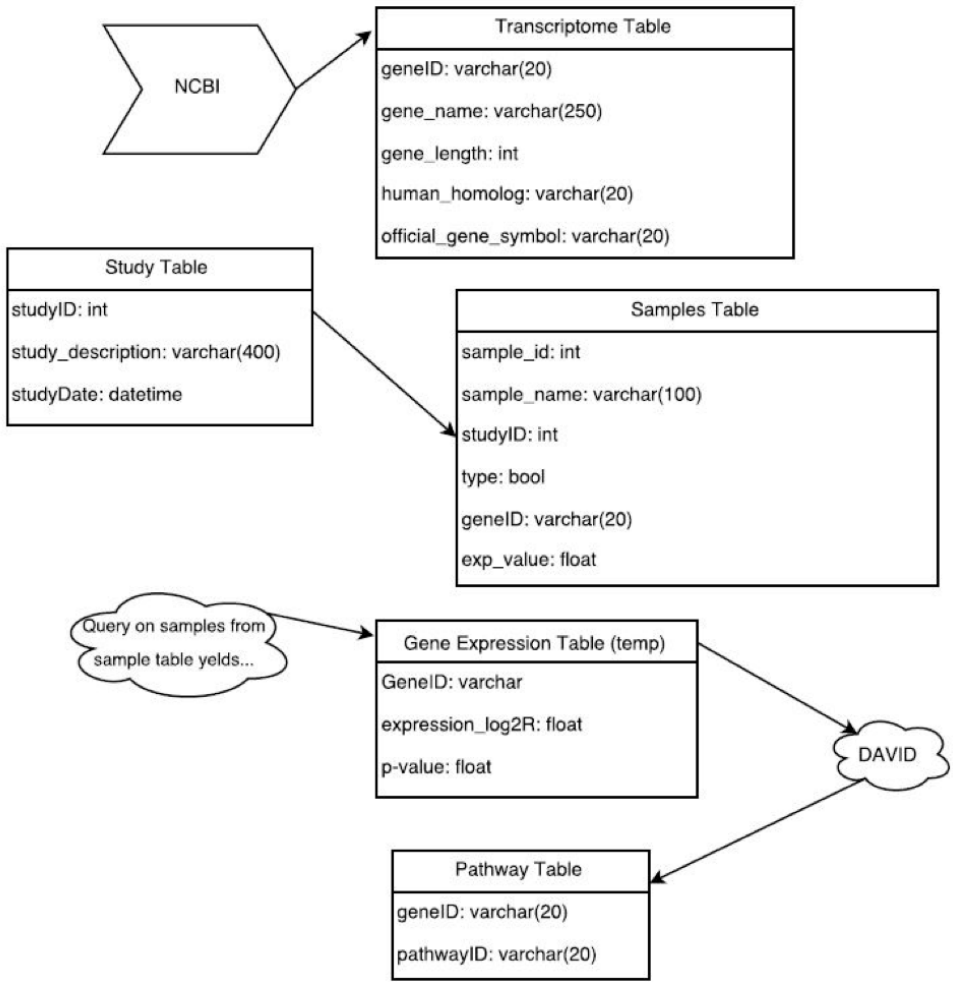


Figure 6: The Database is composed of four persistent tables: Sample Table, Study Table, Transcriptome Table and Pathway Table.

2.3 Project Data

The RNA-Seq and microarray data used in this project were provided by Columbia University. The data was derived from *Aplysia* neuronal samples.

2.3.1 RNA-Seq Data

The RNA-Seq data for 4 libraries (samples C1-C4) was used to validate the Pipeline's portion that is processing the RNA-Seq sequence reads. The libraries were prepared from homogenized neuronal tissues of *Aplysia* and sequenced using the Illumina sequencing machine (Illumina, Inc.). The data was received in the form of zipped FASTQ files, each of which contained sequences and nucleotide quality values for 25-50 million reads.

2.3.2 Microarray Data

The microarray data was used to both validate the Pipeline and answer the research question, to find the list of genes that are expressed differently in motor and sensory neurons. The microarray data was produced using a custom-designed microarray made by Agilent Technologies. The array contained 10,170 probes to known *Aplysia* genes and was used to probe RNA extracted from *Aplysia* R2 and SN neurons:

R2: a major motor neuron that controls mucus release by the animals skin and a major component of the animals defense reflexes.
SN: a mechanosensory neuron (touch receptor neuron); a part of a memory forming network.

The neurons were extracted from live animals, and each sample consisted of 2 to 5 neuronal cells from different animals, pooled together. It has been previously shown that identifying differentially expressed genes is not affected by sample pooling [3, 27]. Even though *Aplysia* neurons are large, the pooling of several cells together was needed to obtain enough material (RNA) to improve the sensitivity of the experiment. Pooling and using samples in replicates ensured that biological diversity (natural variations between different animals) was accounted for and averaged [27, 28].

For statistical purposes, 17 samples were analyzed: 9 replicates from motor (R2 cells) and 8 from sensory neurons (SN). The data consisted of a table of color intensities for each probe that have been corrected for the background noise, along with *Aplysia* gene ID and corresponding known human gene names.

3 Results

3.1 Pipeline Implementation

The current implementation of the pipeline is made with Python and is installed on a server at Columbia University running Centos 7.1 Linux operating system. The pipeline implements CUTADAPT and BOWTIE software along with multiple custom Python scripts. R, a programing language and environment, was used for the statistical analysis of data, where tabulated spreadsheets for gene expression data are analyzed and heatmaps are generated. In addition, R uses the gplot library for visualization. Python scripts utilize SQLdb, mdb and cgi modules.

3.2 Pipeline Validation

The computational Pipeline was validated by comparing the results produced by its parts with test data analysis performed by Columbia University labs established (a.k.a. old) procedure for DE analysis using the same datasets.

Validation of Pipeline Part 1:

The main task of Part 1 of the Pipeline is processing input and producing RPKM tables for each sample. Since microarray data already contained known gene names, Part 1 validation for microarray processing only involved ensuring that the data tables were correctly interpreted as input.

To validate RNA-Seq processing by Pipeline Part 1, the results produced by the Pipeline were compared with those obtained using the old manual approach analyzing the same 4 RNA-Seq datasets. First, the sequence data was screened for quality. The number of reads with more than two estimated incorrect nucleotides per read accounted for about 25% of the reads. 75% of the data was used in subsequent analysis (Figure 7).

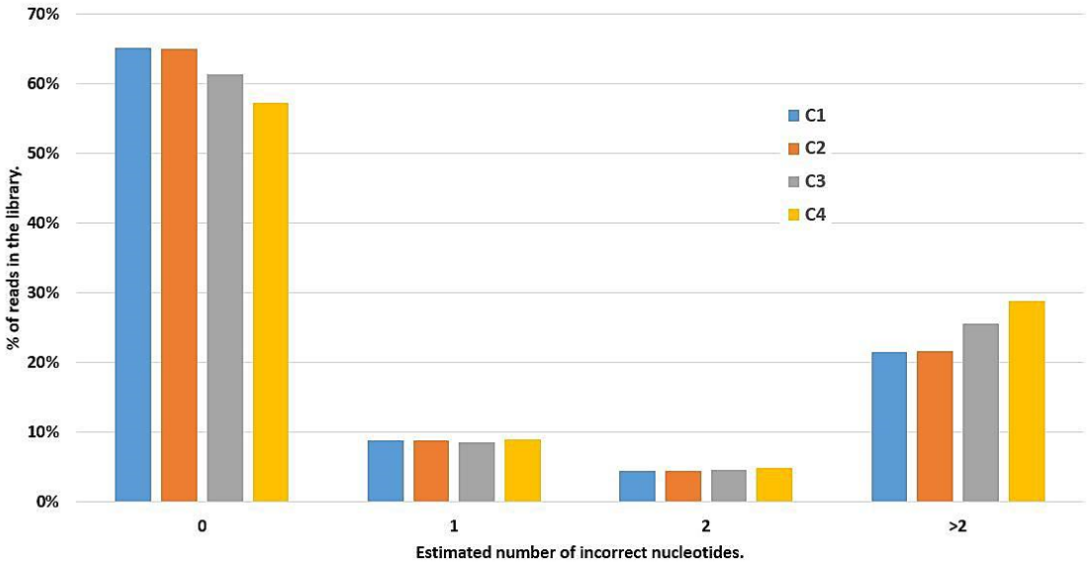


Figure 7: Distribution of the *Aplysia* reads from the tested samples by the number of errors per read

Reads of good quality were then mapped to the reference transcriptome. In the four sample datasets, from 20% to 35% of the reads were successfully mapped and were used to calculate RPKM values.

Finally, the RPKM tables generated by the old procedure and the current Pipeline for the same 4 samples were compared and correlated (Figure 8). The correlation plots were done using the R function. As can be seen, the correlation coefficients between the old and new

expression values for the same samples (marked by red squares in Figure 8) were very high (0.9990-0.9994).

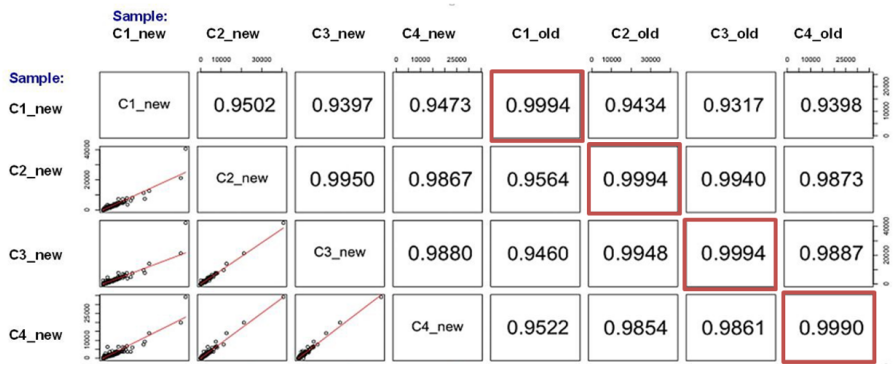


Figure 8: Correlation coefficients in Pipeline validation: (top-right) and scatter plots (bottom-left) for pairwise comparisons of RPKM values calculated for 4 *Aplysia* samples (C1-C4). Comparisons relevant to the Pipeline validation are marked by red squares.

Validation of Pipeline Part 2:

For software validation, the list of DE genes produced by the Pipeline was compared with Columbia University labs established semi-manual approach. The two approaches used the same set of microarray data for the motor and sensory neuron samples. The comparison of the two gene sets demonstrated a 100% match between the two approaches, thus supporting the validity of Part 2 of the developed Pipeline.

3.3 Differential Gene Expression in Motor and Sensory Neurons

The analysis showed that motor and sensory neurons had significantly different sets of active genes and gene pathways. I identified 185 DE genes and 24 DE gene pathways at a confidence level of 5%. These DE genes and pathways correlate with the identity and function of the two specific types of neurons.

In Figure 9, the DE genes are presented as a heatmap, where rows correspond to 185 DE genes and columns are 17 individual analyzed samples. Hierarchical clustering of the samples resulted in good separation of the R2 samples from SN samples. In Figure 9, DE genes shown in the top half have statistically lower activity in the SN neuronal samples (blue color), while in R2 samples, the same genes have high expression levels. The bottom portion of the DE genes shown in Figure 9 have opposite expression patterns – these genes are less expressed in R2 compared to the SN samples.

The full list of the DE genes is shown in the Supplementary Table 2. As can be seen from the table, the expression levels of these genes differed in motor and sensory neurons up to 1,000 times.

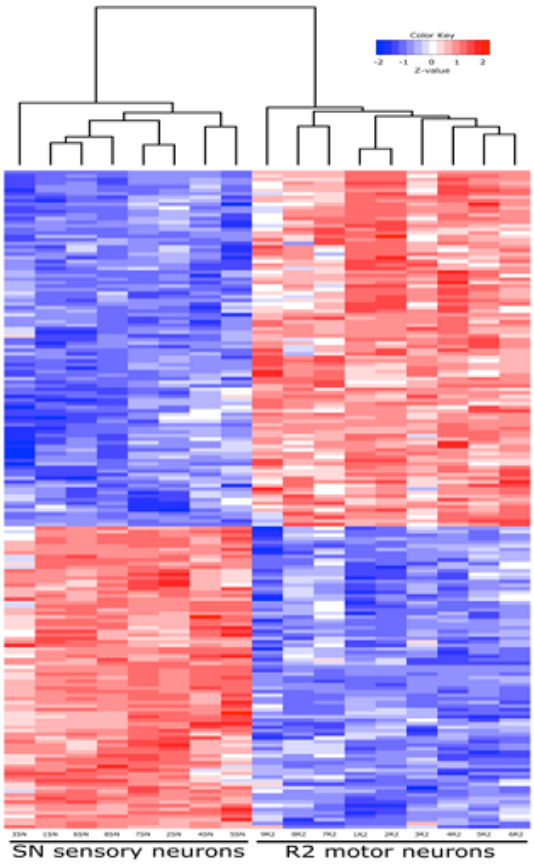


Figure 9: Heatmap of DE genes in R2 and SN neurons and hierarchical sample clustering based on gene expression microarray analysis. Genes exhibiting high activity are shown in red; those with low activity are in blue. The list of the corresponding genes is shown in Supplementary Table S2. See section 2.2.3 for explanation of clustering. Z-value is explained in section 2.2.3.

Table 1: Examples of identified DE pathways grouped by their functions.

Pathway	DE Genes
Cell Interaction pathway group	
hsa04360: Axon guidance	DCC, PPP3CB, ABL1
hsa04720: Long-term potentiation	PPP3CB, PLCB1
hsa04730: Long-term depression	GRIA3, PLCB1
hsa04510: Focal adhesion	TNXB, TNXA, TNR, COMP
hsa04540: Gap junction	TUBA3C, TUBA3D, PLCB1
Cell Signaling pathway group	
hsa04020: Calcium signaling	TACR1, TRHR, PPP3CB, PLCB1
hsa04910: Insulin signaling	HK2P1, IRS2, HK2, PCK1
hsa04330: Notch signaling	JAG2, NUMBL
hsa04310: Wnt signaling	FZD8, PPP3CB, PLCB1
hsa04722: Neurotrophin signaling	IRS2, ABL1
hsa04010: MAPK signaling	PPP3CB, CACNG4
hsa03320: PPAR signaling	CYP27A1, SCP2, PCK1
Cell Receptor pathway group	
hsa04512: ECM-receptor interaction	TNXB, TNXA, TNR, COMP
hsa04080: Neuroactive ligand-receptor interaction	HTR1A, GABRA3, TACR1, TRHR, GRIA3

Table 2: Molecules active in R2 and SN neurons in the Calcium signaling pathway.

Cell type	**Active Receptor**	**Active Enzyme**
SN	TACR1	Phospholipase C
R2	TRHR	Phosphatase 3

Many of the identified DE pathways (Table 1) are known to be important in neurological processes. For example, the Calcium signaling pathway (Figure 10) transmits the signal from the outside to inside of the neuronal cell. In this pathway, a cell surface receptor (exemplified by GPCR marked with a star in Figure 10) detects a signal from a neurotransmitter and passes it to an enzyme (PLCβ in Figure 10). The activated enzyme then starts a chain of intracellular enzymatic reactions. The resulting effect has been shown to be involved in learning and memory formation [29].

This study showed that the Calcium signaling pathway is active in both motor and sensory neurons, but the active genes corresponding to the proteins participating in this pathway are different: the role of the receptor that accepts the outside signal is performed by TRHR in R2 cells as opposed to TACR1 in SN cells (Table 2). Similarly, the active intracellular enzyme in R2 is Phosphatase 3, but in SN it is Phospholipase C. As a result, the same neurotransmitter signal received by the cell is interpreted and carried out by R2 and SN neurons in different ways, thus ensuring different functions and reflecting the identities of these two neuronal types.

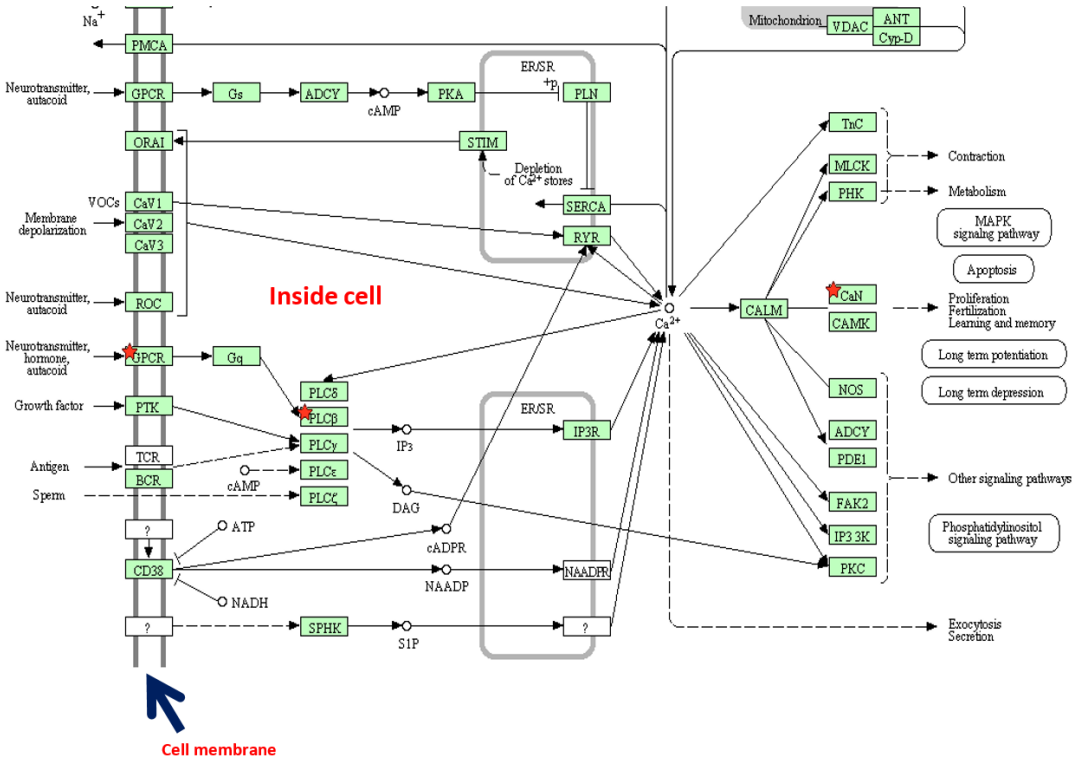


Figure 10: Calcium signaling pathway. Functional proteins are shown as green rectangles, with arrows showing connections between them. Proteins marked with red stars correspond to the DE genes identified in this study. The diagram was produced by the DAVID system (see section 1.3).

4 Discussion

4.1 The Pipeline

The pipeline developed in this project was designed to characterize and compare gene expression in tissue samples while being able to handle data derived from organisms with poorly sequenced and annotated genomes and transcriptomes. Therefore, special attention was given to the proper mapping of the sequence reads produced by RNA-Seq to the available transcriptome sequences done by Part 1 of the Pipeline which was the most challenging portion of the project. At the same time, this study showed that despite of poor annotation and many genes missing in the currently predicted *Aplysia* transcriptome, the transcriptome can still be used in gene expression studies as a reference for mapping. The Pipeline was able to use both RNA-Seq and gene expression microarray data as an input and generated a table with RPKM values that can be used to characterize the gene expression levels in the tissue samples.

Comparison of the Pipeline results for RNA-Seq and microarray data with those obtained independently by alternative analysis approaches demonstrated the validity of this newly developed software. The web interface proved to be sufficient for managing the data and its analysis.

4.2 Identified Differentially Expressed Genes and Pathways

This study compared gene expression in 17 microarray datasets derived from the samples of motor and sensory neurons. Hierarchical clustering revealed that the R2 and SN samples are grouped according to the corresponding neuronal types. Therefore, we can infer that all of the R2 cells have similar expression patterns, just like all of the SN samples have similar expression. The gene expression in R2 neurons, however, is significantly different from that of SN cells, which supports the initial hypothesis. In addition, the fact that the R2 samples cluster together among themselves and away from the SN samples shows that the pooling of the neuronal cells from several animals into one sample was justified as it did not alter the separation between R2 and SN types. This means that the differences in gene expression in motor vs sensory neurons are greater than the differences between individual animals, justifying the use of the pooling procedure.

To answer the main question of this study, I identified 185 genes that have different expression levels in R2 compared to SN. These differences can be associated with the different types of the neuronal cells. These genes belong to gene pathways many of which have been shown to be involved in important neurological processes, such as cell-to-cell signaling and cell interactions. This further supports the notion that DE genes are related to neuronal cell functionality.

The Calcium signaling pathway for example, is active in both motor and sensory neurons, but this function is carried on in different ways due to different activated genes resulting in different outcomes. This is just one of the examples of how the identified genes and pathways are associated with the identity of the two cell types. Among the identified DE genes are several genes that have been previously connected to neurological disorders such as Alzheimer and Huntingtons diseases.

Over the last 50 years of research, *Aplysia* as a model organism has provided important insights into the fundamental organization of neuronal functions, especially learning and memory [4]. Studying the differences in gene functioning in various interacting components of nervous system, such as motor and sensory neurons, will aid in furthering the understanding of the principles behind long-term memory formation.

4.3 Limitations of the Study

In this project, only the transcriptional level of gene activity was considered. In a cell however, proteins perform the major portion of a cells functions and the level of transcriptional activity does not necessarily corresponds to protein activity. It has been shown that transcriptional activity accounts for only 40% of protein activity levels [30]. Nevertheless, since protein activity is significantly harder to study, gene expression analysis at the transcriptional level is usually used to understand the cells functions. This limits the interpretation of the inferences of any gene expression analysis.

If a gene expression microarray is used as the source of information on gene transcritional activity, the gene list in the study is limited to those genes that have been put on the array as probes. The microarray used in this project had probes corresponding to $\sim 10,000$ *Aplysia* genes out of the total predicted number of $\sim 28,000$. This somewhat limits the discovery power of the study.

But even if the RNA-Seq approach (which is less limited in detecting RNA types compared to microarrays) is used, the expression study's ability to interpret the results will be limited. This is because the collection of known *Aplysia* genes is still incomplete, with many genes missing from the databases. In addition, many of the predicted genes have unknown functions with no analogs in other organisms [6].

Also, in addition to transcriptionally active genes, environmental conditions may influence the cells identity. Furthermore, even if an RNA corresponding to a particular gene is present, this does not necessarily mean that the gene's function is carried on properly, since that RNA may not be translated into any functional protein.

4.4 Future Work

All the projects limitations listed above should be addressed in future studies. This can be achieved by developing new approaches of producing samples for gene expression analysis. In addition to analyzing RNA samples, proteins can be extracted and processed to further our understanding of cellular functions.

Since this project revealed interesting genes and pathways with the potential to explain the differences in the functioning of various neuronal cells, the study should be repeated using gene expression microarrays with additional gene probes on it. Ideally, the array should contain probes for all of the potential genes in the genome. A better annotation of the *Aplysia* genome will help with the results interpretation regardless of the method used to obtain RNA data – gene expression microarrays or RNA-Seq.

In this study, only several gene pathways were examined in detail. Further analysis and interpretation of all the identified pathways will provide additional insight into the differences in the function of motor and sensory neurons.

The software development portion of the project should also be addressed in future work. A Web interface usability analysis should be conducted and appropriate modifications introduced. Data management would benefit from adding the option to delete samples, organize samples into groups and move samples between groups. An option to build a heatmap based only on a subset of genes and samples, defined by user, can also be added.

Pathway analysis can be simplified by incorporating the DAVID system into the pipeline by obtaining a license for a local installation from LHRI (Leidos Biomedical Research, Inc.).

4.5 Conclusions

The Project has both biological and bioinformatics results and implications.

The Bioinformatics portion resulted in a software pipeline tailored for analysis of differential gene expression in various tissue samples, especially belonging to poorly annotated genomes. The corresponding database, visualization system and a user-friendly web-based interface were also developed.

The system has a wide range of applicability, since it can be used in any gene expression study, such as those analyzing effects of drug treatments or diseases on gene expression, or detecting the differences between cancer and normal cells.

The study supported the hypothesis that motor and sensory neurons have different sets of active genes. Genes and pathways associated with the neuronal identity have been identified. The major involved pathway groups are: Cell interactions, Signal transduction and Cell receptors. By knowing which genes and pathways correlate with the identity of the neuronal types, we can better understand the nature of the connections between neurons that are created as a result of learning and memory formation.

5 Supplementary Materials

Supplementary Table S1. ASCII symbols for FASTQ file quality line, quality values and associated probability of having an error in a particular position (for Illumina 1.8 sequencer).

ASCII code (decimal)	Symbol	Probability of an error
35	#	0.6309600
36	$	0.5011900
37	%	0.3981000
38	&	0.3162300
39	’	0.2511900
40	(	0.1995300
41	)	0.1584900
42	∗	0.1258900
43	+	0.1000000
44	,	0.0794300
45	−	0.0631000
46	.	0.0501200
47	/	0.0398100
48	0	0.0316200
49	1	0.0251200
50	2	0.0199500
51	3	0.0158500
52	4	0.0125900
53	5	0.0010000
54	6	0.0079400
55	7	0.0063100
56	8	0.0050100
57	9	0.0039800
58	:	0.0031600
59	;	0.0025100
60	<	0.0020000
61	=	0.0015800
62	>	0.0012600
63	?	0.0010000
64	@	0.0007900
65	A	0.0006300
66	B	0.0005000
67	C	0.0004000
68	D	0.0003200
69	E	0.0002500
70	F	0.0002000
71	G	0.0001600
72	H	0.0001300
73	I	0.0001000
74	J	0.0008000

Supplementary Table S2. List of the DE genes identified in this study.

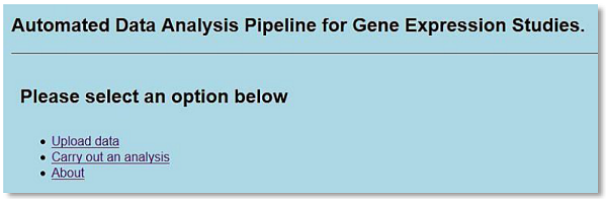
Aplysia GeneID	Log2R	p-value	Human Gene Symbol
XM_005097670.1	10.3455709962744	0.00063721	PGS1
XM_005104695.1	7.40773751045426	0.025527308	none
XM_005097948.1	6.50000131812082	0.0000000379617	none
XM_005089968.1	5.72018528567664	0.0000182498	none
XM_005096989.1	5.6347255030728	0.004096095	none
NM_001204612.1	5.16045098790086	0.00000000549149	Gria3
NM_001204685.1	4.55705415932579	0.0000566777	none
NM_001204551.1	4.49871227248746	0.00095914	EIF5
XM_005099399.1	4.44517484266375	0.041208666	Pdia4
XM_005111786.1	4.18695844701631	0.000674441	none
XM_005107015.1	4.03740009212454	0.0000338158	plekhg4b
XM_005105726.1	3.7892096885152	0.00000244526	none
NM_001204673.1	3.53706096940149	0.011354286	ENPEP
XM_005098739.1	3.51612058326539	0.001705971	Ninj1
XM_005099880.1	3.32686251525645	0.00100573	none
XM_005111036.1	3.16690600277998	0.007107571	TNR
XM_005089362.1	3.13716807111911	0.021086555	SORL1
XM_005113056.1	3.10359807575977	0.002538648	none
XM_005089545.1	2.98239557445281	0.018244824	wfs1
XM_005095712.1	2.89149957638341	0.000523994	LSM5
XR_220520.1	2.84727916204345	0.000282808	UPRT
XM_005097439.1	2.83462003773705	0.010958189	none
XM_005092393.1	2.82893096234879	0.014770474	Fzd8
XM_005101729.1	2.80885177285732	0.003246673	MLL2
XM_005097255.1	2.80211659070925	0.006647027	Abcc2
XM_005096740.1	2.74390306637557	0.008190189	RANBP9
XM_005098459.1	2.73455595081007	0.000263899	Chrna4
XM_005091719.1	2.5604620014997	0.000000372024	comP
XM_005093163.1	2.53852348915774	0.013768557	none
XM_005107873.1	2.42893831335313	0.000101389	none
XM_005108761.1	2.41176602965095	0.001330335	WBSCR16
XM_005105811.1	2.39220606027384	0.000105112	chrna1
XM_005107439.1	2.39043676897376	0.024467026	none
XM_005105399.1	2.38624103610015	0.009221975	vil1
XM_005101009.1	2.37332652027261	0.000599082	Moxd2
XM_005090541.1	2.36557926349437	0.00334022	znf593
XR_220619.1	2.35482182293577	0.009996134	none
XM_005108023.1	2.33701378850773	0.021880501	NUPL2
XM_005098057.1	2.2505345375742	0.021587305	none
XM_005107805.1	2.24789021419973	0.000143248	TPRXL
XM_005105965.1	2.19698373686394	0.026227672	none
XM_005090020.1	2.17733394192006	0.007804765	none
XM_005091714.1	2.12550914648286	0.010095092	METTL2B
XM_005098458.1	2.10844861863603	0.00029881	Chrna4
XM_005092450.1	2.06343504326686	0.0000910289	PPP3CB
XM_005091448.1	2.02755515561874	0.004755814	Tubgcp2
XM_005090073.1	2.00292999694874	0.015805031	none
XM_005109638.1	1.98866748046586	0.021025386	RPA1
XM_005096955.1	1.94687608676619	0.002364871	NT5C1B
XM_005100939.1	1.88904073578646	0.010447905	YIPF4
XM_005107305.1	1.87819840337496	0.001018784	TUBA3D

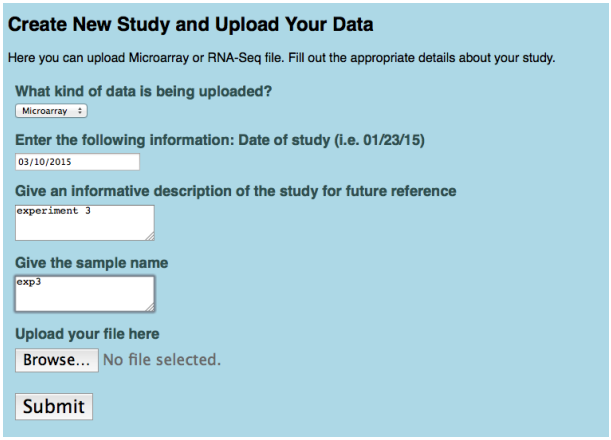
XM_005100413.1	1.85420863451401	0.000700893	DCTPP1
XM_005092769.1	1.7786526281066	0.006932642	none
XM_005112881.1	1.7776316713288	0.00484576	none
XM_005095059.1	1.7672259826512	0.00107365	OS9
XM_005095384.1	1.76412891490823	0.000171506	ARHGAP11B
XM_005093768.1	1.75897139759276	0.0000712221	DCXR
XM_005107440.1	1.69766375954153	0.000000262143	RABEP1
XM_005093406.1	1.66304070357657	0.013154451	Cyp27a1
XM_005092813.1	1.66068193119276	0.000479626	none
XM_005092784.1	1.63868033846269	0.043142313	Fbn2
XM_005108270.1	1.6120272001783	0.000093582	Duox2
XM_005108746.1	1.57826134352416	0.030160743	PM20D2
XM_005090976.1	1.50651974590576	0.041327911	RAB26
XM_005098150.1	1.50198698714645	0.008777464	RAB34
XM_005102973.1	1.49216783261311	0.038896186	Mesdc1
XM_005090509.1	1.48986868482202	0.002631345	RRAD
XM_005096641.1	1.41885993087926	0.032613678	none
XM_005098192.1	1.41423511621862	0.036683444	SLC22A13
XM_005099831.1	1.41126556280412	0.0000201092	none
XM_005102911.1	1.39484447539751	0.014814901	irs2
XM_005112912.1	1.38911082465927	0.003230919	none
XM_005090050.1	1.36104315671431	0.000727712	DPH5
XM_005111124.1	1.32774628027549	0.041104806	none
XM_005109842.1	1.32184091025613	0.035259822	none
XM_005106906.1	1.31449849369348	0.000965899	TRHR
XM_005094083.1	1.31295137766936	0.011670224	Pdcd11
NM_001204578.1	1.30704492635891	0.000369805	tgfbi
XM_005090147.1	1.26794080226341	0.006780028	maeA
XM_005099574.1	1.26617224928554	0.01192577	PET112L
XM_005111697.1	1.22532135575921	0.03085365	none
XM_005089230.1	1.21649365373966	0.0277136	none
XM_005107031.1	1.12220149587106	0.010506553	none
XR_220082.1	1.09852457619749	0.021358393	MXD1
XM_005097672.1	1.08650486673541	0.012002562	none
XM_005101825.1	1.08532045397249	0.038211781	none
XM_005107701.1	1.07626761725704	0.000721869	POLR3C
XM_005093342.1	1.0449294291349	0.004412048	SUMO3
XM_005105348.1	0.996209109997	0.009022193	none
XM_005090043.1	0.98998462275809	0.041610789	SCP2
XM_005092416.1	0.97195080424268	0.02508255	Rsl1d1
XM_005098561.1	0.87649728842583	0.013305365	snx6
XM_005107912.1	0.87393947013696	0.01014497	slc16a13
XM_005089373.1	0.84676994017371	0.003101753	none
XM_005097674.1	0.83771900397696	0.002741348	ELAC2
XM_005097787.1	0.83492648554779	0.041865436	rg9mtd1
XM_005110282.1	0.82028647617273	0.020138334	hscB
XM_005109929.1	0.72337979641967	0.033924292	KIAA1109
XM_005091914.1	0.68444051820329	0.04702778	CCDC40
XM_005089417.1	0.58891094664601	0.041543903	none
XM_005097744.1	-0.62407112585917	0.049553086	none
XM_005095347.1	-0.70619136854781	0.008366715	none
XM_005112816.1	-0.71084519241952	0.049288352	none
XM_005112734.1	-0.7791626028501	0.008248344	none
XM_005097457.1	-0.80198706118859	0.030363301	Cct7
XM_005101645.1	-1.0496938508624	0.033764364	none

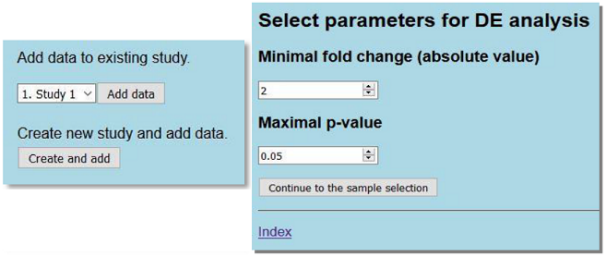
XM_005107639.1	-1.12581360594068	0.000267488	none
XM_005108801.1	-1.17707670849789	0.002415959	none
XM_005096839.1	-1.17720798984426	0.00543623	none
XM_005103254.1	-1.18798274932405	0.029314387	DCC
XM_005098554.1	-1.22983899846174	0.02507342	none
XM_005098755.1	-1.23392160956009	0.007795776	none
XM_005093577.1	-1.23575273045183	0.0000238577	CD82
XM_005095188.1	-1.25118323398237	0.005807683	PNPLA6
XM_005100824.1	-1.39353684693613	0.016458593	PLCB1
XM_005100261.1	-1.39496708847061	0.011748629	none
XM_005101679.1	-1.40600309793594	0.021164745	JAG2
XM_005100134.1	-1.43087009305392	0.005412432	none
XM_005110266.1	-1.50591294174333	0.00292772	PPWD1
XM_005097825.1	-1.60067862094412	0.00040798	none
XM_005108738.1	-1.62807737791997	0.04425206	Tacr1
XM_005109742.1	-1.69034993885379	0.001917383	ROS1
XM_005105359.1	-1.69972649784352	0.000829421	sip1
XM_005112221.1	-1.7133600479062	0.001651417	taf3
XM_005105710.1	-1.80913897873581	0.0000553356	PTPN4
XM_005099881.1	-1.82836568622896	0.008161502	keap1
XM_005098713.1	-1.82970405385707	0.003011369	tprg1
XM_005096302.1	-1.86620022055073	0.011556477	none
XM_005113377.1	-1.88506701384035	0.008331623	HIST2H3A
XM_005100973.1	-1.92496959928283	0.01667112	USP54
XM_005097958.1	-1.92929798287473	0.036899696	PROM1
XM_005107399.1	-1.95095873697681	0.013949197	TSPAN7
XM_005110175.1	-1.97274402581889	0.000757811	PIPOX
XM_005107421.1	-1.98694751223535	0.039249068	none
XM_005095946.1	-1.98942871352843	0.0000206811	PRDM5
XM_005097581.1	-1.99889031661056	0.00000264257	FABP9
XM_005105406.1	-2.00908424240257	0.000727793	none
XM_005092873.1	-2.05429857917269	0.021833985	none
XM_005097398.1	-2.10727731039968	0.00075902	loxl2
XM_005103079.1	-2.15213294878497	0.008692372	BOLL
XM_005095323.1	-2.17801863156933	0.000324453	GABRA3
XM_005095735.1	-2.19104296410575	0.043980277	none
XM_005097184.1	-2.21351617857521	0.028315006	Numbl
XM_005093081.1	-2.25992055903794	0.002009882	ptgds
XM_005089937.1	-2.29280381984249	0.036289517	none
NM_001204708.1	-2.3130467539965	0.0000823064	none
XM_005095057.1	-2.32838433893034	0.012990414	OPA3
XM_005101939.1	-2.39707447207003	0.030542606	TTLL12
XM_005090194.1	-2.47635146193382	0.004050886	ubfd1
XM_005111831.1	-2.53349253204205	0.03553821	RPAP2
XM_005095062.1	-2.53498683739364	0.004139931	ADAMTS9
XM_005103720.1	-2.57152684233217	0.019774294	Arrdc2
XM_005104524.1	-2.58728386102624	0.008795851	plaa
XM_005112789.1	-2.67406608238188	0.001132647	ABL1
XM_005111968.1	-2.75409457406328	0.00080466	LHX4
XM_005099413.1	-2.78949539664283	0.000353245	C2orf60
XM_005108384.1	-2.89246501626644	0.002735361	none
XM_005104569.1	-2.94821220404151	0.003596882	none
XM_005090316.1	-2.95575373124931	0.013265347	hormad1
XR_220671.1	-3.03056717764591	0.0000408997	HTR1A
XM_005108382.1	-3.16130994157745	0.00243688	none

$XM\_005112238.1$	-3.47672782988525	0.002912873	LPPR3
$XM\_005094382.1$	-3.49155613658257	0.000127097	Cops3
$XM\_005089428.1$	-3.54158550677349	0.000355963	YKT6
$XM\_005096860.1$	-3.58859047301298	0.030659274	POLR2E
$XM\_005102183.1$	-3.79816988096311	0.00000109931	setmar
$XM\_005101528.1$	-3.85440382550522	0.00000100784	C16orf48
$XM\_005106005.1$	-3.88589606723315	0.000415655	none
$XM\_005106559.1$	-3.94512792335215	0.000405903	RIT2
$XM\_005105746.1$	-3.99164947999038	0.00000590048	none
$XM\_005089033.1$	-4.00894379924564	0.03019938	none
$XM\_005090895.1$	-4.07348365955564	0.00000000587487	PIN1
$XM\_005098353.1$	-4.13233565179922	0.00000503232	CDKL4
$XM\_005103069.1$	-4.13667036657425	0.0000012531	ABCG1
$XM\_005111344.1$	-4.19290355108379	0.0000016885	none
$NM\_001204563.1$	-4.34346206722373	0.000000119213	BMP1
$XM\_005099337.1$	-4.36112249868699	0.0000000321569	TSPAN3
$XM\_005101273.1$	-4.74022790159443	0.000481624	FRAS1
$XM\_005095980.1$	-4.90649151412736	0.000127652	Amz2
$NM\_001204668.1$	-5.13598849929156	0.000000426908	GPHB5
$XM\_005098296.1$	-5.76330733121353	0.0000000725074	none
$XM\_005105831.1$	-5.95948654049992	0.0000000013989	TNXB
$XM\_005092166.1$	-6.3241796080326	0.00000000157077	none
$XM\_005106551.1$	-6.74128636269914	0.0000000375729	none
$NM\_001204654.1$	-7.54209900547857	0.000123999	none

Supplementary Figure S1. Snapshots of selected web pages.







Select samples for DE analysis

You can select either a set of microarray samples or a set of RNA-Seq samples.

You must select two groups of samples to compare to each other.

Note that groups should not intersect, i.e. no one sample should be included in both groups.

Microarray samples

Group 1

1. Sample neuro_1R2
2. Sample neuro_2R2
3. Sample neuro_3R2
4. Sample neuro_4R2
5. Sample neuro_5R2
6. Sample neuro_6R2
7. Sample neuro_7R2
8. Sample neuro_8R2
9. Sample neuro_9R2
10. Sample neuro_1SN

Group 2

7. Sample neuro_7R2
8. Sample neuro_8R2
9. Sample neuro_9R2
10. Sample neuro_1SN
11. Sample neuro_2SN
12. Sample neuro_3SN
13. Sample neuro_4SN
14. Sample neuro_5SN
15. Sample neuro_6SN

Submit

RNA-Seq samples

Group 1

1. Sample neuro_C1
2. Sample neuro_C2
3. Sample neuro_C3

Group 2

1. Sample neuro_C1
2. Sample neuro_C2
3. Sample neuro_C3

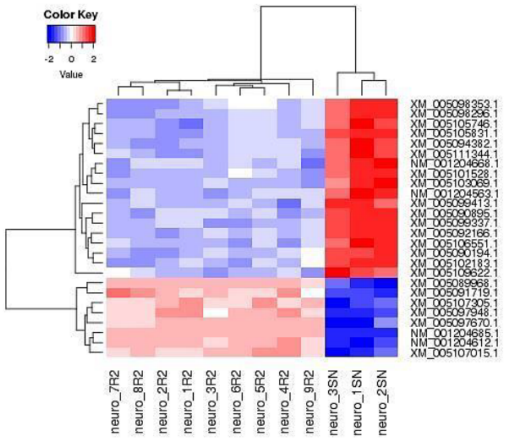
Submit

Index

DE results:

26 genes were differentially expressed.

Heat map for DE genes



View DE table

Continue to pathway analysis with DAVID

References

[1] P. Khatri, M. Sirota, and A. Butte, "Ten years of pathway analysis: Current approaches and outstanding challenges," *PLOS Computatuional Biology*, vol. 8(2), pp. 1–10, 2012.

[2] O. Naumova, M. Lee, S. Rychkov, N. Vlasova, and E. Grigorenko, "Gene expression in the human brain: The current state of the study of specificity and spatio-temporal dynamics," *Child Development*, vol. 84(1), pp. 76–88, 2013.

[3] I. Efroni, P. Ip, T. Nawy, A. Mello, and K. Birnbaum, "Quantification of cell identity from single-cell gene expression profiles," *Genome Biology*, vol. 16, p. 9, 2015.

[4] E. Kandel, "The molecular biology of memory storage: A dialogue between genes and synapses," *Science*, vol. 294(5544), pp. 1030–1038, 2001.

[5] L. Moroz, J. Edwards, and S. P. et al., "Neuronal transcriptome of aplysia: Neuronal compartments and circuitry," *Cell*, vol. 127(7), pp. 1453–1467, 2006.

[6] L. Moroz, J. Ju, J. J. Russo, S. Puthanveett, A. Kohn, M. Medina, P. Walsh, B. Birren, E. Lander, and E. Kandel.

[7] L. Moroz and A. Kohn, "Do different neurons age differently? direct genome-wide analysis of aging in single identified cholinergic neurons," *Frontiers in Aging Neuroscience*, vol. 2, p. 6, 2010.

[8] J. Perkel, "Transcriptome analysis: Microarrays, qpcr and rna-seq." http://www.biocompare.com/Editorial-Articles/137520-Transcriptome-Analysis-Microarrays-qPCR-and-RNA-Seq/, 2012. Accessed: 2015-05-04.

[9] L. Stein, "Genome annotation: from sequence to biology," *Nature Review Genetics*, vol. 2, pp. 493–503, 2001.

[10] A. Rust, E. Mongin, and E. Birney, "Genome annotation techniques: new approaches and challenges," *Drug Discovery Today*, vol. 7(11) Suppl., pp. s70–76, 2002.

[11] W. Zhang *et al.*, "Comparison of rna-seq and microarray-based models for clinical endpoint prediction," *Genome Biology*, vol. 16, p. 133, 2015.

[12] L. Stein, "An introduction to the informatics of next-generation sequencing," *Current Protocols in Bioinformatics*, vol. 36, pp. 11.1.1–9, 2011.

[13] H. Ooi, G. Schneider, T. Lim, Y. Chan, B. Eisenhaber, and F. Eisenhaber, "Biomolecular pathway databases," *Methods in Molecular Biology*, vol. 609:Data Mining Techniques for the Life Sciences, pp. 129–144, 2010.

[14] D. Soh, D. Dong, Y. Guo, and L. Wong, "Consistency, comprehensiveness, and compatibility of pathway databases," *BMC Bioinformatics*, vol. 11, pp. 449–465, 2010.

[15] D. Huang, B. Sherman, and R. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources,"

[16] "Ncbi collection of sequenced genomes.genome of aplysia californica." "http://www.ncbi.nlm.nih.gov/genome/443", 2015.

[17] B. Raddatz, F. Hansmann, I. Spitzbarth, A. Kalkuhl, W. Baumgartner, and R. Ulrich, "Transcriptomic meta-analysis of multiple sclerosis and its experimental models," *PLOS*, vol. 9(1), pp. 1–9, 2014.

[18] "Aplysia genome project. broad institute.." "https://www.broadinstitute.org/science /projects/mammals-models/vertebrates-invertebrates/aplysia/aplysia-genome-sequencing-project", 2009. Accessed: Sep. 20 2015.

[19] "Usearch: Ultra-fast sequence analysis." "http://drive5.com/usearch/manual/exp_errs.html", 2015. Accessed: 2015-09-01.

[20] B. Berger, J. Peng, and M. Singh, "Computational solutions for omics data," *Nature Reviews Genetics*, vol. 14, p. 333346, 2013.

[21] "Cutadapt 1.9.." "https://pypi.python.org/pypi/cutadapt/", 2015. Accessed: 2015-09-01.

[22] "Ncbi, aplysis transcriptome database.." "ftp://ftp.ncbi.nlm.nih.gov/genomes/", 2015. Accessed: 2015-10-01.

[23] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome," *Genome Biol.*, vol. 10(3), p. R25, 2009.

[24] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mam- malian transcriptomes by rna-seq,"

[25] S. Draghici, *Data Analysis Tools for DNA Microarrays.* Chapman and Hall, CRC Press, 2003.

[26] J. Vaquerizas, S. Kummerfeld, S. Teichmann, and N. Luscombe

[27] C. Kendziorski, R. Irizarry, K. Chen, J. Haag, and M. Gould, "On the utility of pooling biological samples in microarray experiments," *Proc Natl Acad Sci USA*, 2005.

[28] M. Konczal, P. Koteja, M. Stuglik, J. Radwan, and W. Babik, "Accuracy of allele frequency estimation using pooled rna-seq," *Molecular Ecology Resources*, vol. 14, p. 381392, 2014.

[29] K. Fukunaga and E. Miyamoto, "A working model of cam kinase ii activity in hippocampal long-term potentiation and memory," *Neurosci. Res.*, vol. 38, pp. 3–17, 2000.

[30] C. Vogel and E. Marcotte, "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nature Reviews Genetics*, vol. 13(4), pp. 227–32, 2012.