Computer Science, Biology ID Major - 2016

Automated Data Analysis Pipeline for Gene Expression Studies

Anton Morozov

Chris Fernandes (CS Dept.) and Steve Horton (Biology Dept.), Advisors

In collaboration with Columbia University, New York

PROJECT

Background

- In a living cell, only some genes are active (expressed)
- Active gene → RNA
- Identify all RNA's = identify active genes
- Cell types differ by their active genes
- Different types of neurons communicate → learning and memory

⇩

Importance of studying differences in active genes

Research Question:

Which genes have different activity in motor *vs* sensory neurons?

Model organism -- sea mollusk *Aplysia*

- large (up to 0.5mm) neurons
- genome has ~30,000 genes
- genome and genes are poorly studied

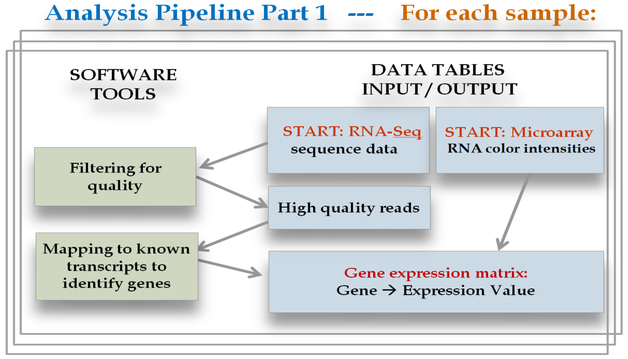
Problem

No available software for gene expression analysis in poorly studied genomes that are easily customizable and capable of processing various sources of RNA data (RNA-Seq and Microarrays).

APPROACH

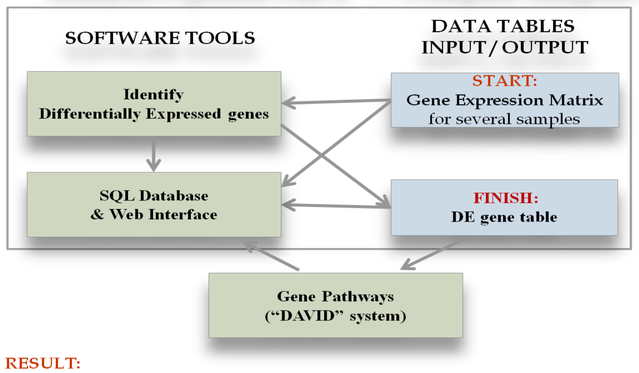
Analysis Pipeline

Analysis Pipeline Part 1 --- For each sample:



RESULT: A gene expression table (matrix) for each sample

Analysis Pipeline Part 2 --- Compare samples:



RESULT:
List of Gene Pathways (functions) that change activity from sample to sample

Pipeline Validation

Processing results of the same test data
Compared: *Pipeline* *vs* *semi-manual* (Columbia U Lab)

✓ **RNA-Seq processing**

High correlation between the "new" and "old" approaches:

Sample:	C1_old	C2_old	C3_old	C4_old
C1_new	0.9994	0.9434	0.9317	0.9398
C2_new	0.9564	0.9994	0.9940	0.9873
C3_new	0.9460	0.9948	0.9994	0.9887
C4_new	0.9522	0.9854	0.9861	0.9990

Correlations of gene activity levels calculated for 4 *Aplysia* samples (C1-C4).

Comparisons relevant to Pipeline validation are marked by red squares.

✓ **Microarrray processing**

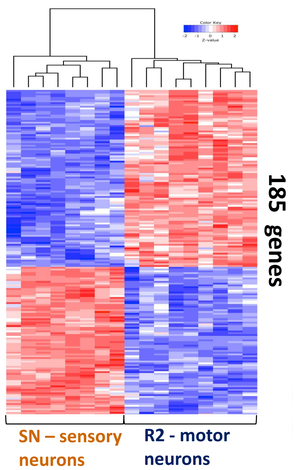
100% match of the "old" and "new" lists of genes that have different activity levels in motor *vs* sensory neurons.

("*Sample*" = RNA data from a certain type of cells)

RESULTS

Comparing Active Genes in motor and sensory neurons

Unsupervised hierarchical clustering of the data samples reflects their identity



Red=high activity
Blue=low activity

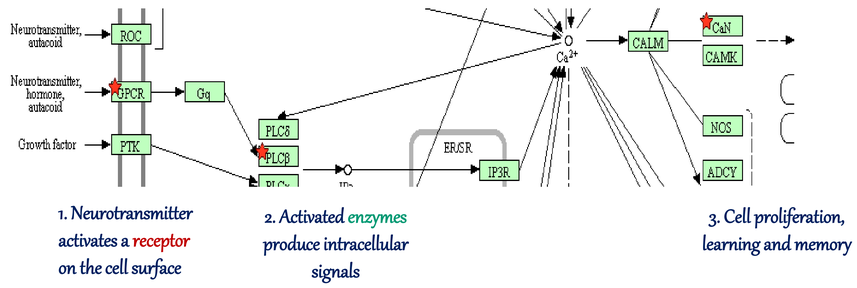
~10,000 genes analyzed, and 185 genes found to be associated with neuronal type

Gene Pathways

Major functional gene functions (gene pathways) that differ between motor and sensory neurons:

Cell Interaction group – 5 pathways
Cell Signaling group – 7 pathways
Cell Receptor group – 2 pathways

Example of Cell Signaling Pathway -- Calcium signaling



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

The same pathway is performing in different ways in SN and R2 neurons

The pathway is important to learning and long-term memory formation

CONCLUSIONS

The developed Pipeline:

- combines several analysis tools in a single system
- single data repository for several projects
- universal: applicable to any gene expression study
- accepts both RNA-Seq and Microarray data

Biological Research:

- Motor and sensory neurons have different sets of active genes.
- Identified 185 genes and 24 pathways that are associated with neuronal cell type (neuronal identity).

Acknowledgments

Dr. Steve Horton and Dr. Chris Fernandes – for guidance.
Dr. Sergey Kalachikov (Columbia University) – for helpful discussions and providing the data.

References

1.Kandel E. The Molecular Biology of Memory Storage: A Dialogue Between Genes and Synapses. *Science*, 2001. 294 (5544): 1030-1038.
2.Mortazavi A., Williams B. A., McCue K., Schaeffer L., and Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods,* 5(7):621–628, 2008.
3. Efroni I, Ip PL, Nawy T, Mello A and Birnbaum KD. Quantification of cell identity from single-cell gene expression profiles. *Genome Biology*, 16:9, 2015
4. Werner T. Bioinformatics applications for pathway analysis of microarray data. *Current Opinions in biotechnology*, 19(1):50–54, 2008.