Coordination of gene expression during Nervous System Development of *Drosophila melanogaster* evidenced by K-means clustering

Flavio Pazos, Ramón Alvarez, Gustavo Guerberoff, Rafael Cantera
1 - Departamento de Biología del Neurodesarrollo del Instituo de Investigaciones Biológicas Clemente Estable , 2 - Instituto de Estadística de la Facultad de Ciencias Económicas, 3 - Instituto de Matemática y Estadística de la Facultad de Ingeniería

INTRODUCTION

In recent years there have been great advances in DNA and RNA sequencing methods and the number of available temporal series of genomic expression data is rapidly increasing (1).

This kind of data can be used to characterize the function of specific genes, relationships between genes, its regulation, coordination and even clinic implications of differential expression (2). Some studies have suggested the coordinated expression of genes during the development of *Drosophila melanogaster* (3, 4, 5, 6). Nevertheless they limit to very general aspects of body morphogenesis.

Here we show the results of applying the K-means algorithm to cluster the developmental transcriptome of Drosophila melanogaster and a functional characterization of the obtained clusters. We are searching coherent waves of genetic expression with special interest in the development of the nervous system. Our hypothesis is that there is a temporal correspondence between the stages of nervous system development and the expression of clusters of genes during each of these stages. Our final objective is to obtain a catalogue of genes with biological relevance to the synapse assembly.

DATA

We used RNA-seq poliA (the RNA that will be translated to proteins) data from Graveley at al. 2011 (7). The used data covers 30 time points along embryonic, larval, pupal and adult stages. Each sample consists of RNA isolated from 30 whole animals. The level of expression for each gene is reported as a RPKM value, a method of quantifying gene expression from RNA sequencing data by normalizing for total read length and the number of sequencing reads.

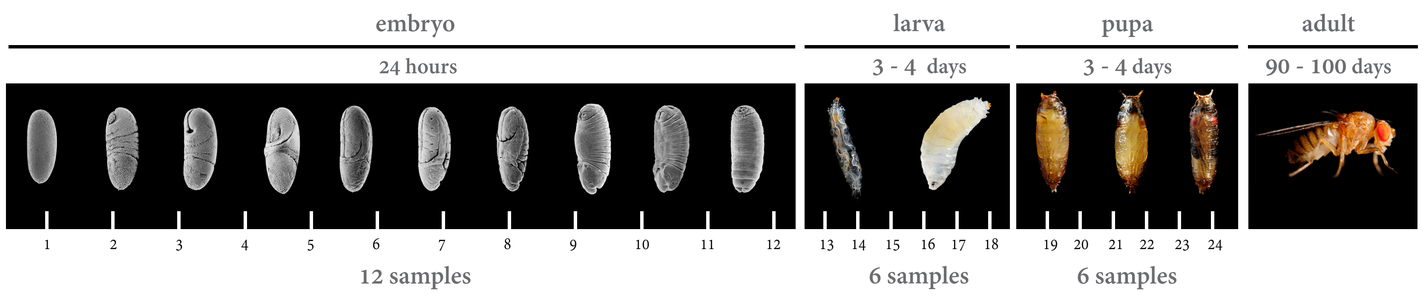
METHODS

We set aside the genes that have no expression along the three developmental stages (genes that are translated only in the adult animal) so our final data set consists in 27 successive expression values for 13,767 genes. Because we are more interested in the way genes change their levels of expression along time than in the absolute values of expression, we devided each value of expression by the maximum value that the gene reach through the considered time lapse.

Then we applied hierarchical (WARD) and non-hierarchical (K-means, Pam, Clara, Fanny) clustering methods using R packages. To compare the obtained results we used Silueta. The more homogenous clusters were found to be those obtained applying the K-means algorithm , clustering the data in 5 groups.

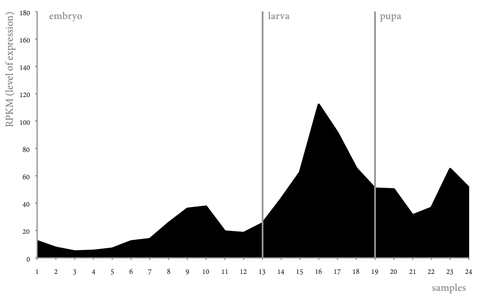
To characterize the clusters we performed functional enrichment analysis of Gene Ontology terms using the GORILLA platform (8). Enrichment is defined as (b/n) / (B/N), where N is the total number of genes, B is the total number of genes associated with a specific GO term, n is the number of genes in the analysed list and b is the intersection between n and B. The 'P-value' is the enrichment p-value computed according to the HG model.

We also labeled each gene as being or not being over-expressed in five different tissues; central nervous system, fat bodies, midgut, salival glands and the carcass, and then studied the distribution of the labels. We labeled a gene as being "over-expressed" in a tissue when the level of expression of that gene in that tissue is at least 5 times higher than the mean level of expression of the gene in the whole organism. We used data of tissue specific expression during third larval stage, available for 10.000 genes in FlyAtlas, the Drosophila gene expression atlas (9).



RESULTS

mean expression of 2.987 genes in **CLUSTER 1**

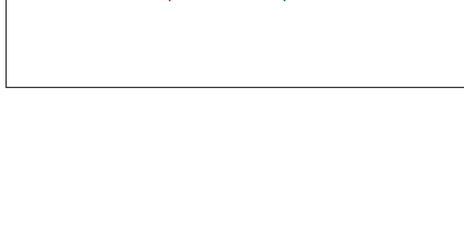


Total enriched GO temrs: 70

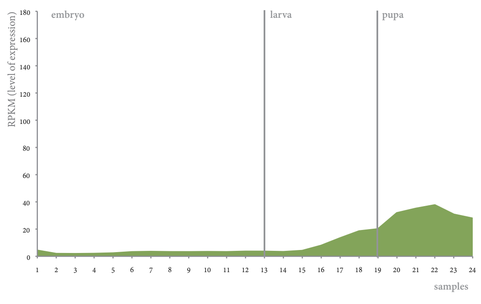
Selected terms	P-value	E	
reproduction	2.12E-37	4.60	General processes
body morphogenesis	1.7E-11	4.66	
red-ox process	6.33E-10	1.57	
defense response	5.9E-9	1.93	
None			Related to Nervous System

Cluster 1 is enriched in general biological processes as "body morphogenesis", "oxidation-reduction processes" and "defense response". Interestingly, its expression profile shows a peak in each of the three stages of development.
There's no enrichment in any processes directly related to the nervous system.

Number of over expressed genes by tissue



mean expression of 2.400 genes in **CLUSTER 2**

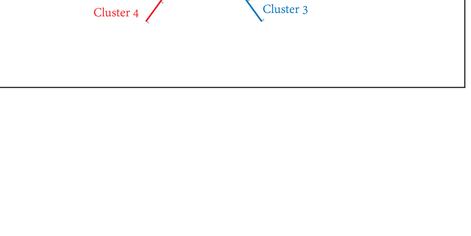


Total enriched GO temrs: 23

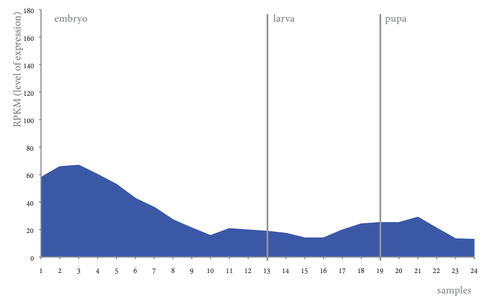
Selected terms	P-value	E	
microtubule-based movement	8.59E-8	2.36	General processes
male gamete generation	1.85E-6	2.43	
detection of chemical stimulus	1.47E-6	2.31	
glycolysis	4.12E-6	3.61	
None			Related to Nervous System

Cluster 2 is the less enriched cluster. It seems to gather genes that begin to express only in the final larval stages and that don't reach high levels of expression. It's enriched in a few very general biological processes such as "microtubule based movement" and "glycolysis".
There's no enrichment in any processes directly related to the nervous system.

Number of over expressed genes by tissue



mean expression of 3.061 genes in **CLUSTER 3**

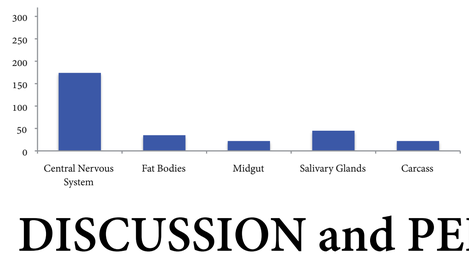


Total enriched GO temrs: 548

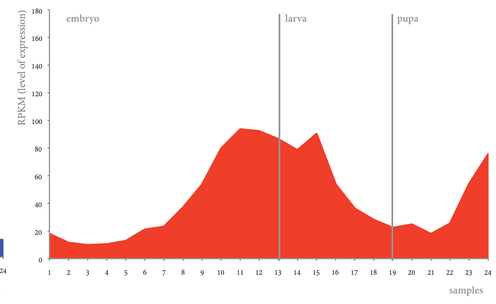
Selected terms	P-value	E	
microtubule-based movement	8.59E-8	2.36	General processes
male gamete generation	1.85E-6	2.43	
detection of chemical stimulus	1.47E-6	2.31	
glycolysis	4.12E-6	3.61	
neurogenesis	1.07E-60	2.2	Related to Nervous System
neuron projection morphogenesis	4.38E-12	1.97	
neuron differentiation	6.25E-9	2.56	
nervous system development	1.19E-8	2.06	
peripheral nerv. system develop.	2.69E-8	2.18	
neuroblast division	3.63E-8	2.93	
central nerv. system develop.	2.23E-6	2.00	
axon guidance	2.25E-6	1.63	
regulation of neurogenesis	2.63E-6	1.87	
neuron fate commitment	9.28E-5	2.09	

Cluster 3 features a "maternal" expression profile, with high levels of expression at the very beginning of embryonic life. This is characteristic for genes whose mRNA is inherited by the egg from the mother. The cluster is enriched in general biological processes as "regulation of gene expression" and "cell differentiation". Regarding nervous system is enriched in various processes that take place at the very beginning of its development, such as "neurogenesis", "neuron differentiation" and "neuroblast division".

Number of over expressed genes by tissue



mean expression of 2.385 genes in **CLUSTER 4**

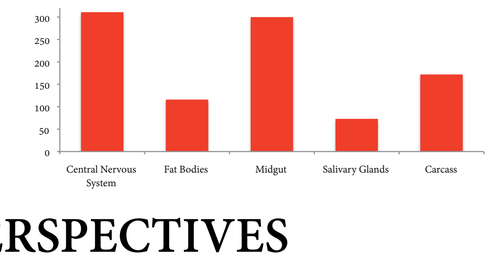


Total enriched GO temrs: 115

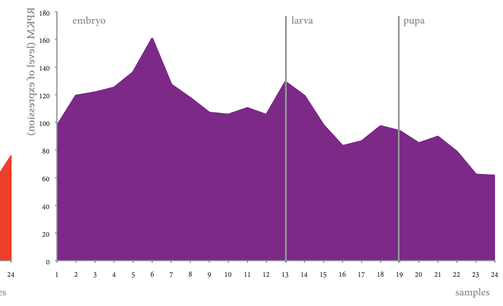
Selected terms	P-value	E	
transmembrane transport	1.12E-41	2.59	General processes
ion transport	4.54E-28	2.47	
electron transport chain	3.33E-21	3.98	
signal transduction	6.93E-6	1.32	
K ion transmembrane transport	1.33E-6	4.73	Related to Nervous System
behavior	2.26E-4	1.39	
synaptic transmission	4.67E-4	1.94	
reg. of neurotrans. levels	4.95E-4	1.78	
reg. of neurotrans. transport	5.11E-4	3.49	
reg. of neurotrans. secretion	5.11E-4	3.49	
larval locomotory behavior	5.57E-4	2.98	
synaptic vesicle exocytosis	9.39E-4	2.84	
reg. of calcium ion transport via voltage-gated Ca channel activity	9.65E-4	5.67	

Cluster 4 is enriched in general processes such as "transmembrane transport", "ion transport" and "signal transduction". Accordingly, is enriched in terms related to the functioning of the assembled synapse, such as "neurotransmitter transport", "synaptic transmission" and "synaptic vesicle exocytosis". Its expression profile, that has low levels at the beginning of embryonic life, rises coherently during the hours in which is already well described that the first synapses begin the transmission of axon potentials.

Number of over expressed genes by tissue



mean expression of 2.849 genes in **CLUSTER 5**

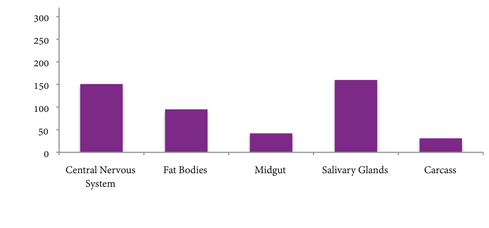


Total enriched GO temrs: 302

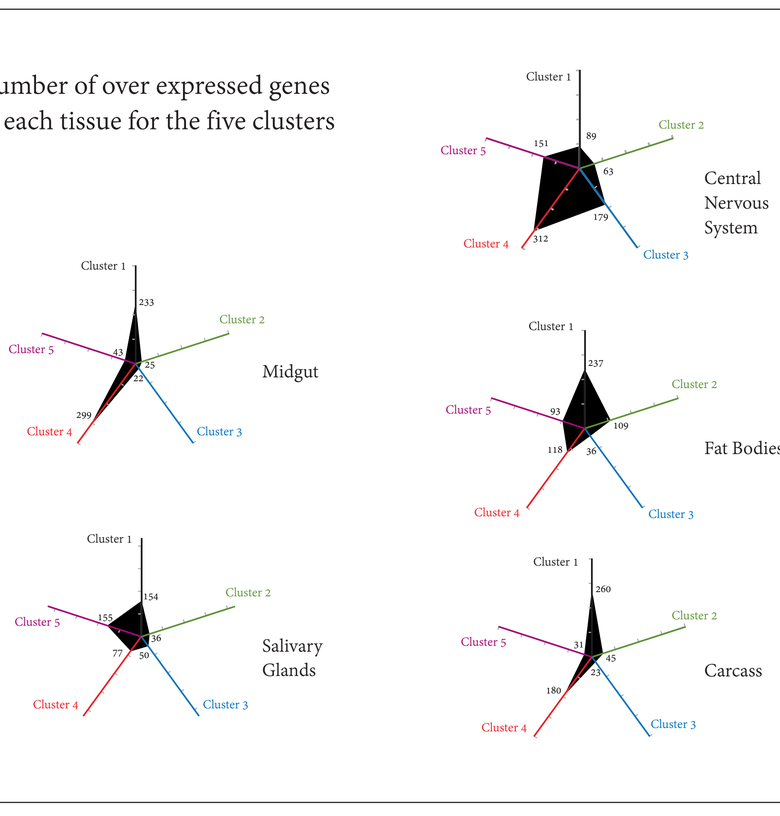
Selected terms	P-value	E	
cellular component organization	3.27E-24	1.44	General processes
organelle organization	2.01E-15	1.46	
cytoskeleton organization	2.75E-13	1.62	
intracellular transport	3.12E-24	2.10	
neurotransmitter secretion	6.46E-13	2.46	Related to Nervous System
reg. of neurotransmitter levels	1.76E-11	2.32	
reg. of dendrite development	8.03E-10	3.58	
vesicle organization	1.47E-10	3.04	
neurotransmitter transport	3.99E-8	1.97	
reg. of dendrite morphogenesis	3.68E-8	3.48	
reg. of synapse organization	8.27E-7	2.42	
neuron recognition	1.73E-5	2.11	
axonogenesis	3.34E-4	1.95	

Cluster 5 is enriched in general processes related to the spatial organization of the cell and its internal organization, such as "cytoskeleton organization" and "organelle organization". Regarding nervous system, its enriched in processes that take place after neuronal differentiation but before the synapses are ready, such as "regulation of dendrite morphogenesis" and "axonogenesis". Is also enriched in processes related to the localization and transport of the synaptic vesicles.

Number of over expressed genes by tissue



Number of over expressed genes of each tissue for the five clusters



DISCUSSION and PERSPECTIVES

Each cluster is enriched in terms associated to several biological processes, showing that clustering genes by its temporal expression profiles results in functionally enriched clusters. This results support the hypothesis that there's a relationship between the temporal expression profile of a gene and its biological function and that it could be possible to predict a gene's function from its temporal expression profile. Our next steps include further clustering of each group of genes and improve the biological characterization of the clusters using the increasing amount of available experimental information.

REFERENCES

1 - Barrett et al. NCBI GEO: archive for functional genomics data sets 10 years on. Nucleic Acids Res, 2011.
2 - Bar-Joseph et al. Studying and modelling dynamic biological processes using time-series gene expression data. Nature Reviews Genetics, 2012.
3 - Ferreiro et al. Whole transcriptome analysis of a reversible neurodegenerative process in Drosophila reveals potential neuroprotective genes. BMC Genomics, 2012.
4 - Arbeitman et al. Gene expression during the life cycle of Drosophila melanogaster. Science, 2002.
5 - Hooper et al. Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. Mol Syst Biol , 2007.
6 - Papatsenko et al. Temporal waves of coherent gene expression during Drosophila embryogenesis. Bioinformatic, 2010.
7 - Graveley et al. The developmental transcriptome of Drosophila melanogaster. Nature, 2011
8 - Eden et al. GOrilla: A Tool For Discovery And Visualization of Enriched GO Terms in Ranked Gene Lists. BMC Bioinformatics, 2009.
9 – Chintapalli et al. Using FlyAtlas to identify better Drosophila models of human disease. Nature Genetics, 2007.