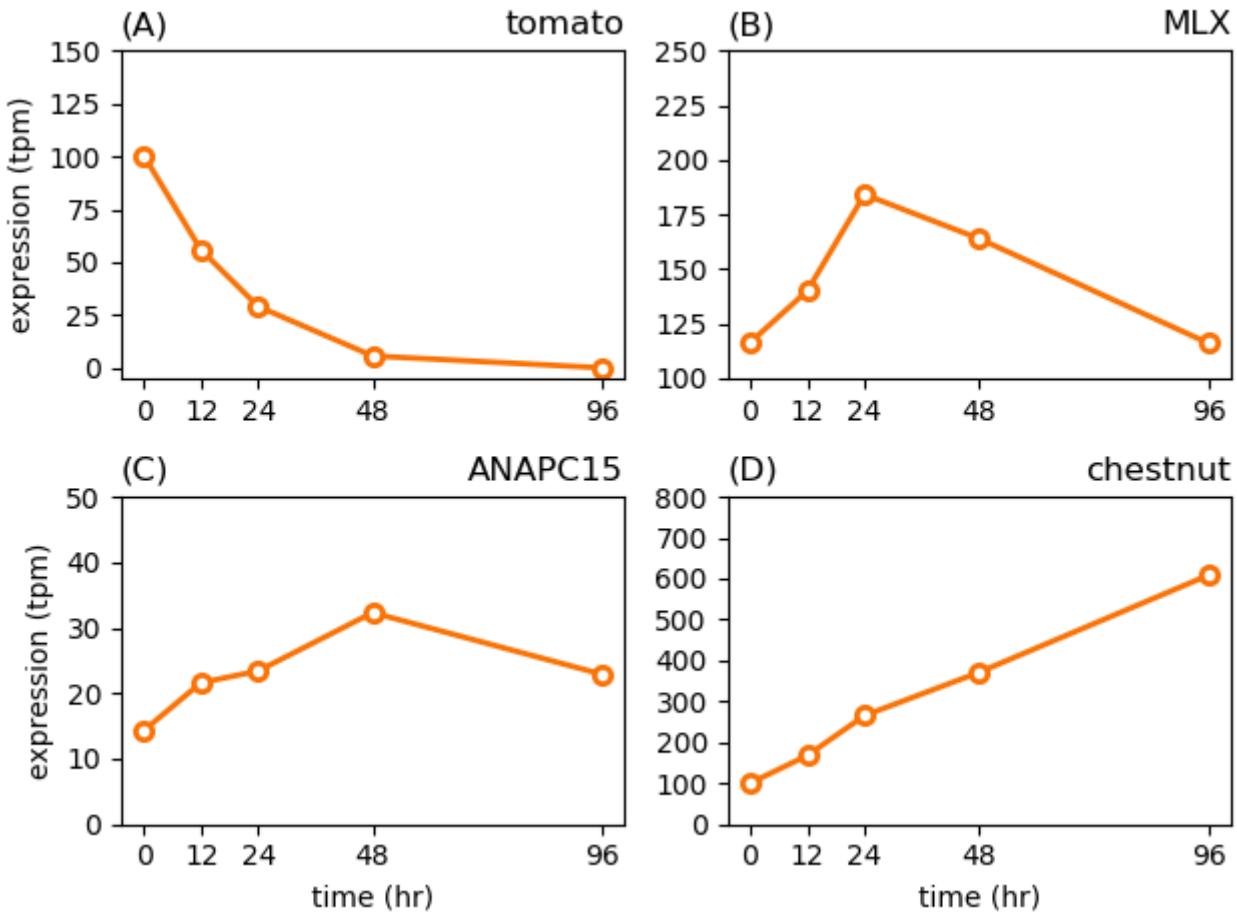
Problem Set 1: The case of the dead sand mouse

You've just joined the Holmes lab. A senior postdoc in the group, Moriarty, recently published a paper that described what he calls the "thanatotranscriptome" of the sand mouse. Moriarty et al. says that a large number of genes in the prefrontal cortex of the sand mouse are differentially upregulated after the mouse dies.

The experiment

Moriarty's methods section says that he instantly kills a sand mouse at time t=0. At t=0, 12, 24, 48, and 96 hours, he dissects out the prefrontal cortex, prepared poly-A+ mRNA, and does RNA-seq. An example of the results is shown in Figure 1 from Moriarty et al.: tomato goes down after death as you might expect from being dead... but MLX comes up at 24hrs then goes down, ANAPC15 peaks at 48h, and chestnut is still coming up at 96h after death:



In his discussion, Moriarty says that these data provide evidence for an ancient program of cortical gene expression that causes the sand mouse's life to flash before its eyes (very slowly).

Well, maybe. But you suspect an artifact of some kind. You decide to look into the result. After carefully reading the paper, it looks like there aren't any obvious problems with the RNA-seq experiment itself. Each data point is an average over several dead sand mice, and the variation seems small, so it's not that this is just experimental noise. The dissections seem carefully done, so it's not that he mistakenly collected anatomically different chunks of brain (which could easily give gene expression differences). He also did some controls to make sure that the relative proportions of different cell types in the dissected tissue hadn't changed much (a common problem in differential RNA-seq analysis of tissues composed of mixtures of cell types: you measure a population average, so if cell type composition changes, it can look like gene expression changes.) Everything suggests to you that the expression level (TPM) calls are reproducible and robust, so if there's a problem, it's in his interpretation of the expression levels.

The data

The expression data for the paper are available in his Supplementary Data Table 1: Moriarty_SuppTable1.

It happens that you've also just been reading another paper from Irene Adler's lab that seems relevant. Adler et al. systematically measured mRNA synthesis rates (in mRNA/hr) and mRNA halflife (in hr), also in the prefrontal cortex, for every gene in the sand mouse. Those data are available in their Supplementary Data Table 2: Adler_SuppTable2.

Both data files are in the problem set zip file, and you'll need them for the exercises below.

1. Check that the gene names match

In principle, the gene names in the two data files ought to match, but (sigh) you know that all kinds of things can go wrong in practice: people use different names for the same gene, spell them with different capitalization or punctuation or added spaces, and so on.

Write a Python script to compare the gene names in the two data files. Output the names that appear in Moriarty_SuppTable1 but not Adler_SuppTable2, if any.

You're especially suspicious of the Moriarty et al. data table because he mentions in his paper that he exported the supplementary methods tables from Microsoft Excel, and you've run across people on Twitter discussing how Excel has systematically corrupted the supplementary data files in many published articles.

If there's a difference - why?

(Use plain Python for this pset. Don't use Pandas, if you already happen to know it. One of the points of the pset is to be able to ingest poorly-formatted data files, by writing your own parsing code.)

2. Explore the data

Write Python code to:

- output the five genes with the highest mRNA synthesis rate. (i.e. in Adler_SuppTable2)
- output the five genes with the longest mRNA halflife. (i.e. in Adler_SuppTable2)
- output the five genes that have the highest ratio of expression at t=96 hours post-mortem vs. t=0 (i.e. in Moriarty_SuppTable1) (You might want to explore the data in other ways too, on your own, as you're doing the next part.)

3. Figure out what happened

This is partly an exercise in manipulating data files line-by-line in Python, but it's also designed to give you a big clue as to what happened in the dead sand mouse experiment.

Write a Python script that merges the two data files, line by line, merging them on gene name. That is, for each line in file 1 for gene X, find the corresponding line for gene X in file 2; we're going to write a single output file with one line per gene. The genes are in different orders in the files, so this merge isn't entirely trivial. For any gene name X that isn't found in both files (Excel corruption) just skip it. For each gene name that is found in both files, output one whitespace-delimited, column-justified data line consisting of 7 fields per line:

- gene name
- Four expression ratios relative to t=0: i.e. tpm[12h]/tpm[0], tpm[24h]/tpm[0], tpm[48h]/tpm[0], tpm[96h]/tpm[0], by processing the TPM data in Moriarty_SuppTable1
- DNA synthesis rate (in mRNA/hr) and mRNA decay halflife (in hr) from Adler_SuppTable2 Save your merged dataset to a file.

Merging data sets, even crudely like this, is often useful as a step towards exploring data and looking for problems, outliers, and correlations.

Explore the data, however you want, for example by looking at the genes with the highest expression ratio t=96/t=0. What do you think is the real explanation for what happened in the dead sand mouse experiment?

Turning in your work

Upload your jupyter `.ipynb` file to sakai, making sure that the notebook can run from start to finish if the required data files are in the same directory as the notebook (i.e., don't use any custom or absolute paths that would prevent the graders from running the notebook).

Hints

- The python script that Moriarty et al. used to make Figure 1 is available in his Supplementary Methods: figure1.py. You could use it as a starting point for parsing the Moriarty_SuppTable1 file. It also gives you an advance sneak peek at using matplotlib to plot data.
- The first hundred sand mouse gene names in Moriarty_SuppTable1 are fictitious (vegetable and fruit names) but the other 19,931 are the names of all the protein-coding human gene names in the GRCh38 human genome annotation. If you want, you can get that annotation from ftp://ftp.ensembl.org/pub/release-85/gtf/homo_sapiens/Homo_sapiens.GRCh38.85.gtf.gz for a GRCh38 human genome assembly as a gzip'ed GTF (gene transfer format) file. It's fun to play with. As a starting point, we've given also given you the python script we used to pull the gene names out (genenames_extract.py).