# Assignment 3: Mapping cDNA reads

**Due date: Wednesday, 2/10 10 a.m.**

You have just struck up a collaboration with the Scott Lab and offered them assistance with analysis. They have data from an RNA-seq experiment performed by the McKinley Lab to test levels of gene expression in the mouse brain under several conditions and need to know what genes their Illumina sequences map back to. Your first task is to map back the reads from a single Illumina run to the coding sequences of the 10 genes they are most interested in.

## Part 0

*Setup your assignment 2 directories*
Change directories to your work directory

```
$ cd assignment3/work/
```

Copy the template `README.txt` and the Python "starter script" `map_sequence_starter.py` to your work directory

```
$ cp /home/assignments/assignment3/README.txt .
$ cp /home/assignments/assignment3/map_sequence_starter.py .
```

## Part 1

*Finish coding the Python "starter script" `map_sequence_starter.py` to map the RNA-seq reads to a set of mouse cDNAs.*

You will use two data files located in the class server:

1. 10 mouse cDNAs (`/home/assignments/assignment3/scott_mouse_cDNAs.fa`)
2. An Illumina read file (`/home/assignments/assignment3/mckinley_raw_reads.txt`)

You don't need to copy these data files to your work directory. Instead, you can (and should) access the files by specifying the path to them.

To get started, open both the Illumina file (`mckinley_raw_reads.txt`), which contains 1 million RNA-Seq reads, and the coding sequence (`scott_mouse_cDNAs.fa`) file with the '`more`' command in the terminal to see what each file looks like.

`map_sequence_starter.py` first reads the cDNA file and stores the data in a dictionary (key = gene name, value = sequence). The script then takes this dictionary and creates another dictionary that uses 25mer sequences from the 10 genes as keys and the corresponding gene name as the value. Finally, the script reads in the Illumina reads and checks to see if the reads match anything in the 25mer dictionary. You'll notice that the sections of script you need to write are already commented for you. This provides a framework within which you should write your code. Please complete this script as indicated by the accompanying comments.

The usage of `map_sequence_starter.py` is:

```
$ python3 map_sequence_starter.py <cDNA_file> <seq_reads_file>
```

The script prints a table of the gene name, reads per gene, and normalized reads per basepair.

## Part 2

*You'll notice that the section of code we provided you does not contain much in the way of commenting.* ***Please provide comments for each line of code indicating in English what the code is telling Python to do****. Specifically, make sure you look up and define all of the built-in and user-defined functions used.*

### Question 1

Run `map_sequence_starter.py` on the Scott cDNAs and the McKinley sequencing reads to align the sequencing reads. Paste the output in your `README.txt`. What genes are highly expressed in this sample? Which are lowly or not expressed? What is the function of each of these genes, and explain why do you think it is highly or lowly expressed in the sample. (Hint: NCBI, ENSEMBL, and MGI have websites with a gene database that will be helpful for this question.)

### Question 2

The Scott Lab has further analyzed the expression of an additional 300 *Mus musculus* genes in this experiment. Of the 300 genes they analyzed, 30 are annotated as having brain-specific expression. Of the top 20 most expressed of these genes, 10 are annotated as having brain-specific expression. Is there an enrichment of genes annotated as having "brain specific" expression in their list of the top 20 most expressed genes? What statistical test did you use, and why? Was the test one-tailed or two-tailed? What was the exact p-value? (Do not state $p < X$). **Show your work to get partial credit.**

### Question 3

In `map_sequence_starter.py` why did we calculate hits as the raw number of counts divided by length of the gene, rather than just using raw number of counts?

### Question 4

If we were trying to map RNA-seq data for the whole transcriptome rather than just 10 genes, what would you do differently? Name at least three limitations of the hash table alignment approach that we have followed in this assignment.

## What to turn in

- Your modified `map_sequence_starter.py`
- A completed `README.txt` file with the first line with how to run your code followed by answers to the above questions.

These two files should be in your `assignment3/submission` folder.

Note: To copy your work files to your submission folder, type
`$ cp <file_name> ~/assignment2/submission/`
where `<file_name>` is the name of the file you want to copy.