ENM375: Midterm Project 1

Dr. Jennifer Phillips-Cremins* (jcremins@seas.upenn.edu)
TAs: Nicole Chiou (nicchiou@seas.upenn.edu); Ryan Nguyen (ryannguy@seas.upenn.edu)

Section 1: Project Description

The mammalian brain is the most complex organ in the body. It controls all aspects of human function and interprets the world around us through our senses. It enables us to plan for the future and defines our memories. Understanding how the brain is wired and how this is regulated through genetic and epigenetic mechanisms represents an exciting area of scientific inquiry. Moreover, understanding how neurons work in the healthy brain can give new insight into how neuronal function goes bad in disorders of the brain.

One of the most devastating human conditions is Alzheimer's disease (AD). AD is a neurodegenerative disorder that progressively disrupts cognitive function and is predicted to affect more than 16 million Americans by 2050[1,2]. The global health care cost for patients with Alzheimer's and dementia is estimated to be more than $600 billion in 2016[3]. Significant advances have been made in recent years toward understanding the genetic basis of FAD. Highly penetrant, rare coding variants underlying familial AD (FAD) have been discovered in the genes encoding amyloid precursor protein (*APP*), presenilin1 (*PSEN1*), and presenilin2 (*PSEN2*)[1-4]. Beyond inherited FAD, the large majority (>95%) of AD is non-inherited and known as sporadic (SAD), but the genetic basis of SAD is still poorly understood. The mechanism by which FAD mutations lead to pathologic features in the brain of the 5% of cases with inherited AD, and the stage in development in which they occur is completely unknown. Thus, we have large gaps of knowledge regarding the causal drivers for the majority of brain defects in AD. Importantly, no clinical trial based on our mechanistic knowledge to date has convincingly demonstrated the long-term attenuation or prevention of neurodegenerative disease onset and progression. Overall, there is tremendous opportunity to use new technologies to understand AD.

In this project, you will use your growing knowledge of biostatistics principles to analyze data acquired from healthy human induced pluripotent stem cells (iPS cells) and iPS cells containing FAD mutations at *APP* and *PSEN1* (**Figure 1**). We have assessed gene expression (total mRNA transcript levels per gene) genome-wide in the iPS lines with a new technique called RNA-seq. RNA-seq allows you to leverage next generation sequencing to assess gene expression across all genes in the human genome in one assay.

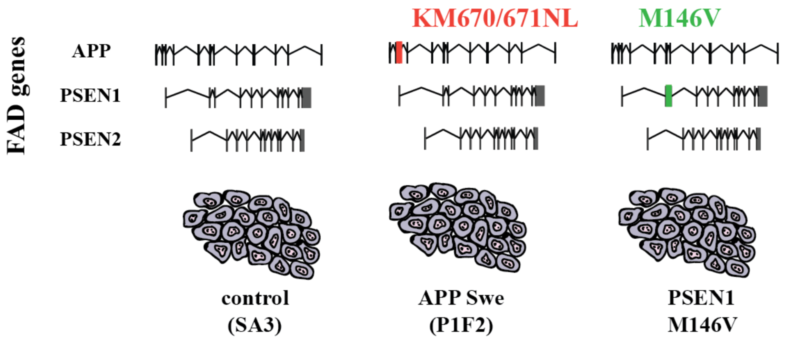

Figure 1. Schematic of the healthy and FAD human iPS lines analyzed in this study with RNA-seq.

Table 1. Experimental Conditions

Sample	**Description**
SA3_rep1	Healthy iPS Replicate 1
SA3_rep2	Healthy iPS Replicate 2
SA3_rep3	Healthy iPS Replicate 3
P1F2_rep1	iPS APP FAD mutation Replicate 1
P1F2_rep2	iPS APP FAD mutation Replicate 2
P1F2_rep3	iPS APP FAD mutation Replicate 3
PSEN1_rep1	iPS PSEN1 FAD mutation Replicate 1
PSEN1_rep2	iPS PSEN1 FAD mutation Replicate 2
PSEN1_rep3	iPS PSEN1 FAD mutation Replicate 3

Section 2: What Data Do I Have?

You have been provided with a .csv file which contains the RNA-Seq dataset you will be working with for this project. The dataset contains the gene names found in the first column, the replicate names in the first row (the header), and the rest of the cells are data points that represent the normalized RNA-seq counts for each genotype (n=3) and for each of the replicates (n=3) (**Table 1**).

Section 3: Preparing the Data for Analysis

1. **Getting to Know the Data:** As you may notice later in your academic/professional careers, working with real data is always messy. Before being able to properly analyze most results, the obtained data must be pre-processed. However, in order to pre-process the data, you must know what your data looks like and what to do in order to pre-process it accordingly.

   Data Science Portion: Import the data from the .csv file and create the following data structures:

   - A list containing all the gene names
   - A list containing all the replicate names
   - An array containing the counts for each gene across all replicates, with each row being a different gene

   Essay Portion: How many genes in total does your dataset include? How many cellular conditions and replicates per condition are provided (i.e. does this measure up to what Sections1-2 tell you should be in the data)?

2. **Visualizing the Raw Data:** Now that you know what your data contains, the next step is to see what it looks like.

   Data Science Portion: Create histograms for every replicate and every condition for all the counts in a given condition (include the values from all genes in each histogram). The number of bins you use is up to you.

Essay Portion: Describe the histograms. Are there differences between replicates in the same condition or between different conditions? Is there anything odd about the distributions? Is it easy to see the trend of the data? (HINT: If you are having trouble seeing any bars, zoom into the histogram).

3. **Fixing the Messy Data:** As you may have noticed from your histogram, the data is very messy and extremely hard to visualize with the way it currently is formatted. This is clearly due to the disproportion between the amount of high values and low values in the data. As a rigorous data scientist, you realize that a log transformation might alter the distribution of the data.

   Data Science Portion: Write code that filters out all the rows from your matrix in which all columns of a specific gene have values of a defined minimal threshold or lower. For example, if you choose a threshold of 0, then if a gene has a value of 0 at all of the experimental conditions, you should be deleting the entire row (all values of that gene across all conditions) from your data set. Once you have your filtered data set, add a pseudocount of +1 to each gene count across all conditions in your data (i.e. take the current number and add a +1 to it) and then perform the log2-transform of the data.

   Essay Portion: How many genes did you end up deleting from your data set that included 0 values across all samples? How many genes are left in your total data set? Explain why a log-transform after adding a pseudocount can make this data easier to work with.

Section 4: Data Analysis

1. **The Histogram (Ch. 2)** – After performing the pre-processing on your original real and messy data, you want to get a feel for what your new filtered and log-transformed data set looks like.

   Data Science Portion: Create histograms for every replicate and every condition for all the counts in a given condition (include the values from all genes in each histogram). Then, for your favorite condition, draw histograms of all genes in the column but try10, 100 and 500 intervals.

   Essay Portion: Describe the shapes of these histograms and how they have changed after transformation. Suggest your preferred number of intervals.

2. **The Scatter Plot (Ch. 2)** – Curious how the different treatment conditions affected gene expression, you take an interest in comparing the RNASeq values between the different conditions.

   Data Science Portion: Construct 4 scatter plots showing the relationship between (1) healthy iPS Replicate 1 and healthy iPS Replicate 2, (2) healthy iPS Replicate 1 and FAD iPS APP mutation Replicate 1, (3) healthy iPS Replicate 1 and FAD iPS PSEN1 mutation

Replicate 1, and (4) FAD iPS APP mutation Replicate 1and FAD iPS PSEN1 mutation Replicate 1. Make sure to label all axes properly.

Essay Portion: Describe the degree of similarity between the gene expression profiles of the replicates and conditions in each of the scatter plots. Is there any sort of clear correlation between the two replicates of the same condition? Is there any clear difference between replicates of two different cellular conditions? What might these plots suggest in terms of whether or not FAD mutations in APP or PSEN1 affect gene expression globally?

3. **The Box Plot (Ch. 3)** You notice that there are specific genes that change their expression levels based on the mutation. Some seem to upregulate during APP FAD mutation, whereas some downregulate compared to the healthy iPS cells with no mutation, and some do not change at all. You want to compare and contrast the gene expression values of these different genes between healthy iPS cells replicate 1 versus FAD APP replicate 1.

   Data Science Portion: Draw box plots in a single figure for all replicates and conditions showing the distribution of RNA-seq values for the following transcripts: ENST00000259915, ENST00000396861, ENST00000369080, ENST00000503395, ENST00000216117. There should be 5 total figures.

   All replicates for a single condition should be included in that boxplot and boxplots for all conditions should be shown for each gene. For each gene, the figure should contain a boxplot for each of conditions (WT, FAD APP, FAD PSEN1). Make a completely separate plot for each gene.

   Essay Portion: Explain what the different parts of the box plot are: the box itself, the line in the middle of the box, the whiskers, and any data points not included in the box. From these results can you point out which genes might exhibit a change in gene expression based on the treatment conditions? Describe your observations.

4. **Mean & Variance (Ch. 3)** You already noted that different genes show different expression levels across the treatment conditions. In fact, you even showed that some genes express at higher/lower levels depending on the treatment. However, to quantify this you will want to visualize the distribution of the means, variances, and coefficients of variation of the RNA-seq values.

   Data Science Portion: For the five genes in question 3 (for which you made the scatter plots) – compute the sample mean, the sample variance, the coefficient of variation and the population variance across replicates in each of the three conditions.

   Essay Portion: Describe the shape of these histograms. Why are they different? From these data, can you deduce whether genes with the highest RNA-seq values have all high RNA-seq values (low variance) or some high and some low RNA-seq values (high variance)?

5. **Mean vs. Median (Chp. 3)** – Hypothesis confirmed: there is definitely variation between genes and testing conditions. Looking more closely, you can now use the tools you've developed to compare distributions of RNA-seq values for the different cellular conditions.

   Data Science Portion: For the five genes in question 3 (for which you made the box plots) – compare the sample mean to the median across replicates in each of the three conditions.

   Essay Portion: Given the shape of the distributions, argue for whether the mean or the median is the most appropriate descriptor and why.

6. **Sampling (Ch. 4)** – Technology can be fickle. While you are able to work with the data set, perhaps your lab partner was unlucky and only recovered 100 random rows from the genome-wide RNA-seq data set. To help, you want to show them how such a sample would have affected your own results.

   Data Science Portion: For one specific condition (for example: a column representing wild type human iPS Replicate 1) create: (i) a histogram displaying the distribution of counts of all genes, (ii) the distribution of 100 genes chosen uniformly at random from the column, (iii) the mean sampling distribution by taking 1,000 samples of n=100 genes and computing the mean from every sample. And finally, (iv) compute the mean from all genes in the column from part (i) and compare to the mean of the sampling distribution of the means.

   Essay Portion: Describe how the 3 distributions differ. Explain why the second and third distributions differ from the first, and why the second distribution differs from the third. How many samples (of 100 values chosen uniformly at random from the data set) would be necessary to estimate the true mean RNA-seq value across the column with the precision of 2 decimal places?

7. **Error (Chp. 4)** – Your boss took a look at your results using the mean values as an estimate in problem 4 but is not swayed. For total transparency, you need to show the uncertainty in this estimation.

   Data Science Portion: For one specific condition (i.e. column representing wild type untreated healthy iPS cells replicate 1), take the first 10 samples of 100 genes chosen uniformly at random from the column, and for each sample compute the (i) mean, (ii) the standard error of the mean, (iii) the sample standard deviation. Plot a bar plot of means for each of the 10 samples, showing error bars as (i) +/- 1 SEM, (ii) +/- 2 SEM, (iii) 95% confidence intervals, and (iv) +/- 1 sample standard deviation.

   Essay Portion: Discuss the advantages and disadvantages of the different types of error bars in describing the data.

8. **Probability (Ch. 5)** – The highest gene expression values each serve an important part in the cellular state of iPS cells. Therefore, you become interest in how likely the highly expressed genes are in each of your conditions.

   Data Science Portion: For one specific condition (i.e. column representing wild type human iPS FAD APP Replicate 1), estimate (i) the probability that any gene expression value in the column for human iPS FAD APP Replicate 1 would be between 3 and 5 and (ii) the probability that any gene expression value in the column for human iPS FAD APP Replicate 1 would be >=7.5.

   Essay portion: Describe how you came to this estimate. How could you come to a better estimate? Provide relative frequency histograms if they help you make your argument.

9. **Hypothesis Testing (Ch. 6-8)** – Now that you have a thorough understanding of the data from different viewpoints, you can begin to formulate and test hypotheses about FAD mutations APP and Psen1 compared to the wild type iPS cells. We will focus on several key genes to test our hypotheses:

Gene Symbol	**ENSEMBL Gene ID**	**Transcript ID**	**Relevance**
GAPDH	ENSG00000111640	ENST00000396861	housekeeping gene
POU5F1	ENSG00000204531	ENST00000259915	iPS/ES cell marker
APP	ENSG00000142192	ENST00000357903	FAD gene
PSEN1	ENSG00000080815	ENST00000357710	FAD gene
INPP5F	ENSG00000198825	ENST00000369080	
ZNF732	ENSG00000186777	ENST00000419098	
SYT11	ENSG00000132718	ENST00000368324	

Using the healthy iPS cells as the control, assume the null hypothesis that the means of the replicates for each gene between healthy vs. FAD APP or healthy vs. FAD Psen1 do not change. You will perform a hypothesis test for each of the 7 genes above (i.e. there is a separate test for each gene, testing the difference in expression counts between the healthy control and APP FAD and then between healthy control and PSEN1 FAD. This means that there will be 7 total tests.)

Data Science Portion: Write out the null and alternative hypothesis for each gene before starting the statistical test. Determine your test statistic (for example, the means between the replicates of a condition). Plot strip charts of expression values of the replicates for iPS cells (n=3) vs. replicates of FAD APP (n=3), showing the median expression value as a horizontal line in the middle of the strip charts. For each gene, use the Poisson distribution to create the null model, using the mean of the healthy iPS samples as the lambda parameter for the null model. Plot the null distribution for each gene modeled as the Poisson distribution with the lambda computed from the null healthy iPS condition. Then compute the p-value by comparing the probability of the average expression value from the FAD APP samples to the healthy iPS null distribution. Plot a vertical line representing the location of the mean FAD APP value compared to the healthy iPS null on the plotted Poisson distribution. Show the plots and the p-values for every gene. Finally, discuss whether you will reject the null hypothesis or choose not to reject the null hypothesis. Based on your selected alternative hypothesis, state whether you chose a one-sided or two-sided test.

Essay Portion: What significance level did you use for your statistical inference and what role does this choice play in controlling for hypothesis testing error? Interpret what this result might mean for neuron activity and gene expression.