**Project Milestone**

For your milestone, we expect you to have:

* Acquired
  + Link to Github repository - <https://github.com/davidwsant/now-u-c-me-datascience-final-project/tree/master/Output_Tables_and_Results>
  + Explain the source of the data
    - Summary of outlined details in the proposal
  + This project is using data about hydroxymetylcytosine (5hmC) and transcription. It is well known that changes in transcription rate (the amount of protein made from a given gene in a given time) can lead to differences that determine cell type. Additionally, this is how the cell can react to external stimuli and this is often dysregulated in disease. How these external stimuli affect transcription is not completely understood, but one of the methods is through covalent modifications to the DNA within the cell.
  + Covalent modifications to DNA include methylation of cytosines at position 5 (5mC) and hydroxymethylation of cytosine at position 5 (5hmC). Covalent modification means that the changes are stable and cannot generally be changed without the help of enzymes. These enzymes respond to changes in the environment, which in turn means that the DNA modifications respond to changes in the environment as well. These changes in DNA methylation have been shown to affect the rate of transcription and imbalances in methylation are known to contribute to disease. Although it is known that DNA methylation affects transcription, the relationship between the changes in DNA methylation (especially 5hmC) and transcription is not very well understood.
  + From data generated in the Gaofeng Wang Lab, we hope to determine if we can use a model to predict if genes will change in transcription levels in response to addition of vitamin C to the culture media (to enhance global 5hmC generation).

DEVIATION: Our proposal initially outlined three datasets that we would perform analyses on. However, only the third cell line (ARPE-19 cells treated with or without 50 micromolar Vitamin C) will be used for the remainder of the data science project. This is due to the 30+ hours it took to clean the singular data set and in order to keep up with the project timeline we initially set (see gantt chart included with project proposal), we will utilize this cleaned data for the outlined analyses from here on out.

* Cleaned
  + Explain what we (Dave) did to clean the data
    - The initial cleaning step that was performed involved a normalization to the hMeDIP count files to now show a read count per million (RCPM). No outliers were accounted for as of yet. The columns added by htseq-count were not removed either.
    - The second cleaning step involved accounting for outliers. This was done by ignoring for now the htseq columns and running a linear regression of the averages per treatment type (Vitamin C and control). A Cook’s cutoff value was used because it takes both X and Y axes into consideration. The points with a high Cooks value were removed for the normalization.
    - The third cleaning step is to divide by the length and multiply by 1000 to get “Fragments per Kilobase per Million,” also known as FPKM.
    - The fourth cleaning step is to determine which values increased, decreased, or experienced no change. A 1.5X fold change cutoff was used, as well as a corrected P-value from the edgeR cutoff of 0.05.
    - The fifth cleaning step involved coordinating each peak in the region analysis with a gene or region of a gene.
    - The sixth cleaning step was to sum the counts per gene for both location and change to get the counts of peaks within each region for each annotated gene.
    - These same normalization techniques were applied to the RNA-seq data similar to the way we performed the hMeDIP-seq normalization. The difference here is that the htseq-count lines will be removed prior to the first RCPM.
    - The next step for normalizing the RNA-seq involved removing genes that are comparatively low in expression because they would just add noise.
  + It took so long that we decided to limit the project to one dataset (but wow it’s shiny)
* and explored your dataset.
  + Descriptive stats!
    - Summarize Dave’s beginning stat analysis
    - Run some more?

You should also explain in more detail what will go into your final analysis.

* Copy segment from the proposal & extrapolate

Explain deviations from your initial project plan.

* Mention the feedback we got from others here

**Names of Critiquers (And the Oscar Goes to...):**

Ryzen Benson

Chrissy Neff

Ram Siripuram

**Notes for our own team:**

* + It was suggested that our team look for directionality between cAMP and Vitamin C as well as between the human retinal and Schwann cells.
    - This is no longer applicable as the datasets we originally planned to examine will now focus primarily on ARPE-19 cells (human retinal cells). See “deviations” listed above for more details.
  + Our peers also suggested that we perform a classification task to involve the up-regulated, down-regulated, and unchanging gene expressions.
    - This was taken under advisement and will likely be included in the primary analysis of our final submission.
  + The peer reviewers asked for a clarification on where/how we were obtaining our data. We responded with details similar to those that can be found under the “Data” header outlined above.
  + Our reviewers wanted more details as to what the significance of this project would be, what applications currently exist, and why we are interested in this topic. We answered these questions more fully in our submitted project proposal and a summarized version is detailed in this notebook under the “background(???)” header.
    - During this part of our peer review, it was also recommended that we describe the topic background in a way that a reader with a non-biological background might still easily understand. We attempted to do this in our project proposal and will maintain this goal in our final submission as well.
      * Relating to the above: it was suggested that an infographic or commonplace metaphor to represent the underlying workings of the process we are examining might be useful to readers/graders. This is currently being worked on.
  + It was suggested that regression analyses/models be applied to our data as well. We have taken this under advisement and plan to utilize this technique in our final submission.
  + There was some initial confusion about the number of classification categories we were utilizing (the reviewers initially understood 15,000 classification categories!). However, this was clarified in that there will be approximately 15,000 *genes* with 6-9 classification categories utilized in this project.

*If you are uncertain about the scope, please contact the staff member responsible for your project.*

Dave’s Final Summary:

Exploration of the data has shown that there are about twice as many upregulated genes as there are downregulated genes in the ARPE dataset, but this is still only about 1/3 of the expressed genes (14,599). Over 2/3 of the expressed genes are protein coding genes (11,228), even though they make up a relatively small portion of the annotated genome. This is consistent with other RNA-seq studies. It looks like the highest average peak counts are genebody upregulated (1.9) followed by genebody nondifferential (1.6). We now have many dimensions and this should be helpful for determining if we can classify genes based on the genomic info and hMeDIP peak information (data about DNA methylation).