**Objective: Analyze data assigned from RNA-seq 1 assignment and “healthy donor” HD dataset to identify differentially expressed genes between two datasets.**

**Step 1: Analyze an additional dataset.**

**Here, you will essentially repeat RNA-seq assignment 1 with the newly assigned HD data (**[**https://www.encodeproject.org/files/ENCFF875PQO/**](https://www.encodeproject.org/files/ENCFF875PQO/)**) and (**[**https://www.encodeproject.org/files/ENCFF979NLQ/**](https://www.encodeproject.org/files/ENCFF979NLQ/)**) onto Galaxy. You will run Salmon for each biological replicate as before (Use both files). Redo steps 2 and 4 (step 3 files should already be in your queue) for the above data.**

**You must also revert the original “Gene Quantification” files to tabular using the “DataType” tab.**

**At this point, you should have 4 “Gene Quantification” files all in “tabular” format: 2 from your original ENCODE ID and 2 from above.**

**Step 2: Prepare these files for differential gene expression analysis.**

**We need to set up these tabular files to input into DESeq2.**

1. **Use the “Text Manipulation > Compute on rows” tool to round the 5th column in each Gene Quantification file.**

**Note: The current default version of this tool on Galaxy is 2.0. Click the “Versions” icon on the top right to change to version 1.6 in order to follow the below instructions.**

A screenshot of a computer

Description automatically generated

* 1. **Under “Add Expression”, put c5.**
  2. **Under “as a new column”, select your Gene Quantification file**
  3. **Under “round result?”, select YES.**
  4. **Execute.**
  5. **Repeat these steps for the other 3 files.**
  6. **Watch 4 files get queued up on the history and wait for them to go green.**

1. **Use the “Text Manipulation > Cut columns from a table” tool to cut the 1st and 6th column out of these new files. (DO NOT CHOOSE Cut columns from a table (cut))**
   1. **Under “Cut columns” write c1,c6**
   2. **Under “From” select file generated in step 1**
   3. **Execute**
   4. **Repeat these step for the other 3 files from step 1.**

**Step 3**: **Run** **DESeq2 to do differential expression analysis.**

1. **Under Genomics Analysis: RNA-seq select “DESeq2”**
2. **Under “Specify a factor name, e.g. effects\_drug\_x or cancer\_markers”, put “CellType”.**
3. **Under the first Factor level, under “Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'”, put “HD”.**
4. **Ctrl-click (Command-click on macs) the first 2 “cut on data XX” files that originated from the HD ENCODE IDs.**
5. **Under the Second Factor level, under “Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'”, put the cell type of your first assignment’s ENCODE ID.**
6. **Ctrl-click (Command-click on macs) the second 2 “cut on data XX” files generated from your ENCODE ID.**
7. **Under “Files have header ?”, select “No”**
8. **Under “Output normalized counts table”, select “Yes”.**
9. **Hit Execute.**
10. **This should make 3 files on the history. Wait for these to go green.**

**Step 4: To remove genes below a .05 threshold value go to “Filter and Sort” > Filter data on any column using simple expressions**

1. **Under Filter, select the DESeq2 result file**
2. **Under “With following condition”, enter c7<.05**
3. **Hit Execute.**

**Step 5: Count the number of lines in the DESeq2 result file and this new “Filter” file using “Text Manipulation > Line/Word/Character count of a dataset” like before.**

**Step 6: Filter the new “Filter” file twice. Once for “c3>0” and once for “c3<0”. Those lines in “c3>0” are higher in Panc1 and those in “c3<0” are higher in your original sample. Count the lines in these files to answer question 2 and view them to answer question 3.**

**On your own**

**How many genes (lines) were in your original DESeq2 results file (These represent all of the genes in both conditions)? How many genes were called as differential (Lines in your Filtered file)?**

1. **How many genes are at a higher level in HD? How many genes are at a higher level in your original sample?**
2. **Select a gene that is higher in HD and one that is higher in your original sample. Do a bit of research on it like in the last assignment. Does it make sense that this gene is higher in its particular sample? Why?**
3. **Look over the plots generated by DESeq2 (send these to me along with the word document). The first plot is a 2D representation of all 4 datasets known as a PCA. There should be 2 points per sample which represent the different replicates. Which experiment was more reproducible (aka the points were closer together)?**