BIOM200 Module1 Group Project

Phase 1 – Generate your question:

* Assigned Friday 9/23, due to Emily by Sunday 9/25 at 9PM

* Work together on your teams to generate an interesting biological question that can be answered with data that is available on the ENCODE website. Your solution must incorporate data from two different assays, one of which must be RNA-Seq, and the second must be something other than RNA-Seq. Explain why you chose the datasets you did and how they will be used to answer your question. What particular features of that type of assay make it the best choice in order to answer your particular question? Write up a proposal (half page, single spaced) explaining your question and computational approach. Feel free to use the literature to draw inspiration for your proposal.
* Turn in by Sunday 9/25 at 9PM to Emily by email. Once you have turned in your proposal, you can sign up for a meeting time for Monday evening for your group to meet with Emily and discuss the proposal. Meetings will be Monday night from 4 – 7 in 30 minute intervals (location TBD). She will provide feedback and some direction with tools that will help you carry out the proposal. A final assignment of what you are expected to complete for your project will be given by the beginning of class on 9/28.

Phase 2 – Answer your question!

* Assigned Wednesday 9/28, due to Emily by Monday 10/3 at 9AM
* For RNA-Seq Datasets:
* Run featureCounts to quantify gene expression
* Calculate RPKM and make a clustermap using log2(RPKM). Use the spearman correlation coefficient.
* Make a PCA plot based on gene expression
* Use DESeq2 to quantify differential expression, select a significance cutoff and explain how you arrived at that cutoff. Output a csv file containing a list of significantly differentially expressed genes along with their p-value and fold change (this is the output directly from DESeq2 that you will filter for significant genes)
* Make a MA plot based on the significance cutoff you chose
* Include the full path of the notebooks used for calculating RPKM, clustering, making the PCA plot and performing differential expression (One notebook per group is fine, each group member does not need to provide a notebook).
* Copy all figures into a word document and attach them to your report along with a 1-2 sentence interpretation of each one.
* For other dataset:
* Use at least one bedtools command to analyze a bed file. This analysis must be integrated with some feature of your RNA-Seq data. (Example – are the genes that are differentially expressed also bound by my transcription factor).
* Include all of the bedtools commands that you used as a supplement to your report.
* Write up a final paper describing your project (up to 2 pages single spaced)
* Introduction/Background – Make if brief, one paragraph is fine.
* Question/Hypothesis – Introduce the specific question you are able to address with your datasets. Introduce what datasets you are using and how you will use them to answer your question.
* Results – Discuss what you learned from the differential expression and combining your results with the other dataset.
* Discussion/Future directions – Are the results what you expected? What else could you learn from the data you have? What other information would you like to have in order to answer your question more thoroughly? What experiments would you conduct to follow up on your results?
* Make sure to properly cite all sources. There are resources on the website to help you with this.