BMI206: Statistical Methods for Bioinformatics
FINAL PROJECT

For the project, you will work with one of the papers assigned during the quarter and conduct your own analyses of their data. The goals of this exercise are

- Learn to critically read bioinformatics papers from a statistical perspective
- Obtain primary data from a publication
- Practice selecting appropriate analysis methods
- Practice making figures and other data displays
- Compare and evaluate different bioinformatics and statistical approaches to answering a scientific question
- Learn to evaluate the sensitivity of results to analysis choices

Your project will involve working with a small group. The graded portions of the project will be (1) Write-up of data summary, (2) Write-up of analysis plan, (3) Oral presentation of re-analysis of data for one figure, (4) Oral presentation of a novel analysis with published data, (5) Project report. The final report will be done individually, and the other four components will be done as a group.

Step 1 – Get the data: You will need to obtain primary data through a website or from the authors. Note: this can be easy or might be the hardest part of the project, so you should start it early. If the paper includes many data sets, you may focus on just one of them if it allows you to accomplish the goals of the project. If the data set is very large, you are encouraged to use high performance computing to analyze the full data to facilitate direct comparisons with the published results. Students can request a cluster account from the instructor. Again, starting early will be helpful, because compute time can be a bottleneck. If you can meet the goals of the project by only analyzing part of the data (e.g., a random subset of the observations or variables), this is also acceptable.

DUE MONDAY OCT 10: Each group should turn in one paragraph plus a table or figure summarizing your data. Describe what data you obtained and how you got it. Calculate some basic summary statistics on the data, including but not limited to (i) how many variables? (ii) how many observations? (iii) how many missing data points? (iv) what type of data (e.g., continuous, counts, categorical, binary)? Include the names of all team members. Email to kpollard@gladstone.ucsf.edu.

Step 2 – Analysis plan: The goal of this step of the project is to plan Steps 3 and 4. You will select (i) a figure from the manuscript to re-create, (ii) a result to test for sensitivity to analysis choices (can be same as figure in (i)), and (iii) a question that goes beyond the original results of the manuscript. See Steps 3 and 4 for details. You do not need to use all the data in the manuscript, but a substantial amount of the published data should be used. You are welcome to use data, tools, and methods beyond those in the manuscript.
**** Your ultimate analysis plan can evolve after this date. This is a starting point.***

DUE MONDAY OCT 17: Each group should turn in two paragraphs describing their specific analysis plans for Steps 3 and 4. These should include the data to be used, statistical methods to be applied and software/tools you to be employed. Include the names of all team members. Email to kpollard@gladstone.ucsf.edu.

Step 3 – Reanalysis: You will evaluate your own understanding and the reliability of the published results by attempting to make a figure from the publication on your own from the original published data. You may deviate from the precise software used in the paper, but keeping the methods as close as possible will help you to determine the source of any disagreements. Your analysis should involve some processing of primary data. For example, if you aim to recreate a heat map of pairwise distances between gene expression profiles, it is not sufficient to get the distance matrix from the authors and then give that as input to a plotting function. You should try to compute the matrix yourself from the gene expression profiles. You might go back to raw read counts and do a full RNA-seq quantification. At a minimum you should start from the normalized expression values for each gene in each sample. In terms of visualization, the emphasis should be on whether or not some one looking at your figure would come to the same conclusions or not. It is not critical that your figure is publication quality or that it uses exactly the same colors/symbols.

DUE WEEK OF NOV 7: Each group will give an oral presentation describing the results of their reanalysis. You should use slides to describe what you did and what you found. Be sure to also highlight the challenges and lessons learned. Grading will be based on approach not similarity of the plots per se. Classmates will be expected to ask questions during oral presentations. You will be graded on participation, both asking questions and answering them during your presentation. There will be one talk per group.

Step 4 – Novel Analysis: You will design, implement (in code) and conduct your own analysis of the published data. Your analysis should apply methods from the course to the published data to do both of the following:

1. **Sensitivity:** Explore the sensitivity of at least one main finding to the methods employed. Do you come to the same conclusions when you analyze their data with a different method, different parameters, or different software settings?
2. **New Hypothesis:** Address at least one novel question not covered by the authors using their data.

Your analyses may incorporate data from other sources, such as browsers, databases, or other publications. It may be possible to accomplish your goals by sticking to methods introduced in the labs. You may also explore methods from the lectures by finding or writing your own code. Techniques not covered in the course can also be used, but you will need to give a brief explanation of the method and your reason for using it. It is encouraged to employ methods from a week other than the one in which your paper was assigned; most of the methods in the course have applications to many different bioinformatics problems.

DUE WEEK OF NOV 28: Each group will present their novel analysis in an oral presentation to the class. You should use slides to describe what questions you investigated, how you did it, and what you found. There will be one talk per group.

DUE MIDNIGHT DEC 12: Each student will submit a project report. The report should include the following sections: introduction, methods, results, and discussion. The results section should contain three subsections corresponding to reanalysis, sensitivity, and new analysis. The text should be ~1000 words. You may include up to four figures or tables. You do not need to detail everything you did or tried. Instead focus on your main findings. Email your individual write up to kpollard@gladstone.ucsf.edu.