BMI206: Statistical Methods for Bioinformatics
FINAL PROJECT

For the project, you will work with one of the research papers that we read and discuss during the quarter. You will conduct your own analyses of the published data, possibly using other publicly available data or data from your own research. 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

The graded portions of the project will be (1) Write-up of data summary, (2) Two write-ups of analysis plans, (3) Two oral presentations, and (4) Project report. Part of your project will involve working with a small group (~4 people). The novel analysis part of the project will be done individually, though you are welcome to get input from classmates.

<u>Part 1</u> – Get the data (GROUP): Choose one of the papers and select a figure from the paper that you will aim to recreate and test for sensitivity to analysis choices. 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 BEFORE CLASS TUESDAY OCT 22: Each group will turn in one paragraph plus a table or figure summarizing their 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 the TAs and kpollard@gladstone.ucsf.edu.

<u>Part 2</u> – Reanalysis (GROUP): You will evaluate your own understanding and the reliability of the published results by (i) attempting to recreate a figure from the publication on your own from the original published data and (ii) testing the sensitivity of this result to analysis choices (e.g., different method, different parameters, or different software settings). The goal is to learn how a result is obtained from primary data and to evaluate how conclusions may depend on analysis choices.

You may deviate from the precise software used in the paper, but keeping the methods as close as possible for recreating the figure will help you to determine the source of any disagreements. In terms of visualization, the emphasis should be on whether or not someone looking at your figure would come to the same conclusions. It is not critical that your figure is publication quality or that it uses exactly the same colors/symbols. Your analysis should involve some processing of primary data. 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 BEFORE CLASS TUESDAY OCT 29: Each group will turn in two paragraphs describing their specific analysis plan for Step 2. These should include the data to be used, the figure to be re-created and sensitivity tested, statistical methods to be applied and software/tools to be employed. *Include the names of all team members*. Email to the TAs and kpollard@gladstone.ucsf.edu.

DUE NOV 25-27: 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 and interpretation, 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. *Please plan to attend class Monday – Wednesday this week and start any Thanksgiving travel after class on Wednesday.*

Part 3 – Novel Analysis (INDIVIDUAL): The goal of the individual project is to add a figure to the manuscript that you used in the group project. You will design, implement (in code) and conduct your own analysis of the published data. You are also very welcome to incorporate publicly available data from other sources, data from one of the other papers discussed in the course, or data from your own research. Your analysis should apply methods from the course, or related methods, to this data to address at least one novel question not covered by the authors. You will need to write some of your own code and produce a figure along with the supporting text and code for that figure.

DUE BEFORE CLASS TUESDAY NOV 12: Each person should turn in two paragraphs describing the plan for conducting their novel analysis for Step 3. These should include the data to be used, the hypothesis to be tested (or question investigated), statistical methods to be applied and software/tools to be employed. *Include your name.* Email to the TAs and kpollard@gladstone.ucsf.edu.

DUE DEC 12-14: Each person will present their new figure to the class and answer questions from classmates. You should have one slide with your data/methods and one slide with the figure you made. 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.

DUE MIDNIGHT DEC 13: Each person will submit a project report describing their novel analysis that adds a figure to the manuscript. The report should include the following sections describing your findings: Title, Hypothesis / Problem statement, Results, Methods, Discussion / Conclusions, and Figure with Figure Caption. Imagine that you are co-author of the manuscript and your responsibility is to add a new results section to the manuscript that includes a new figure. The text should be a concise summary <1,000 words. You should also include a separate file or repo link to your code. You do not need to include everything you did or tried. Instead focus on your main findings. If you were part of the team that wrote the manuscript you would need to consolidate your explorations into a compelling narrative! Email your individual write up to kpollard@gladstone.ucsf.edu.