Environmental Bioinformatics Final Project

"An experiment is reproducible until another laboratory tries to repeat it."
-- Alexander Kohn, co-founder of Journal of Irreproducible Results

Open science and reproducible research are common goals across research communities. Openness and reproducibility facilitates scientific advancement and has the potential to accelerate discoveries. However, reproducibility is a complex issue -- even within the constraints of computational biology. New programs are developed and old ones are updated or abandoned. Computers change and improve, as does the technology used to produce HTC data. There is a recent push for transparency in research, with the sharing of raw data and of "in-house" scripts used for analysis increasingly encouraged or required by scientific journals. During this course you will be exposed to a survey of bioinformatic approaches to deal with a variety of data types. For your final project, you will attempt to reproduce the *bioinformatic analyses and findings* from a recent publication.

The purpose of this final project is severalfold:

1. To dive deeper into an area of bioinformatics that interests you
2. To practice structuring a full computational project
3. To get a sense of the level of detail in computational biology that helps (or hinders) reproducibility

Project overview:

Teams of two students will select a paper of common interest includes some variety of -omic analysis. Your team will attempt to completely replicate the data analyses using the study's publicly archived data and any/all details included in their materials and methods or supplementary materials. All projects must be cleared with the instructors ahead of time to ensure that the scope of the project is appropriate for the class (i.e. no, you cannot re-analyze the 40+ Tb of data from the Human Microbiome Project).

Project proposal:

Teams will submit a brief (less than 1/2 page) proposal including the paper title, proposed reanalysis, and primary raw data location and estimated size (in Gb). Please note: it is perfectly fine for teams to choose to analyze only a subset of the data within a publication (e.g. choose one pair of treatments rather than analyzing all 9 treatments presented). Depending on the topic of the proposed project, each team will be assigned a faculty mentor who will be their main point of contact for questions or help over the course of the project. Each team will be responsible for scheduling a series of meetings with their faculty mentor. At the first meeting, teams will discuss their plan of analysis and present a Gantt chart to show the projected division of labor across the team. At the second meeting, teams will update their faculty mentor on progress and highlight any issues they are running into.

Project introduction:

After meeting with their faculty mentor, teams will make a present a brief (5 minute) overview of the paper of choice and their proposed analyses for the class. This presentation should provide background on the paper and its findings, the type of data and number of samples proposed for analysis, any metadata associated with the samples, and the types of analyses to be performed. Additionally, and most importantly, the teams should outline how they will compare their results to those of the original paper. What metrics will they use to assess the "reproducibility" of the study? What figure or figures will they try to recreate?

Project products:

1. Each team will create a repository on GitHub for their project. This repository should house all scripts used for analysis and any data products that are under 50Mb in size. Additionally, this repository should be well organized and show a clear system of file management. Folders that must be present:
    a. envs/: a folder that contains yaml files that can be used to generate any conda environments used in the data analysis
    b. scripts/: a folder that contains any scripts, including slurm scripts, used in the analysis of the project
    c. logs/: a folder containing any log files from the analysis of the project (i.e. the output of any analyses run)
    d. jupyter-notebooks/: a folder containing any jupyter notebooks used for the analysis or visualization of the data
    e. data/: a folder that contains the raw data in the project. The contents of this folder that are larger than 50Mb in size should not be pushed to GitHub (and should be noted in a .gitignore file).
    f. output/: a folder that contains multiple sub-folders, which include the output of all analyses.
2. The repository of the project should have a README.md file that describes the project (i.e. the project proposal). Additionally, all files and directories within the repository should be described in the README.md and any instructions for how to run the analyses using the provided code should be provided.
3. A final jupyter notebook should be created called final-comparison.ipynb that performs a direct comparison of the original data and the re-analyzed data. This is one of your main outputs for the project, so make sure it is clear, well-annotated, and includes specific discussion of how your output compares with the published results, and why you think it does or doesn't match them.
4. Each team member will submit a publication-style contribution statement detailing who was responsible for each aspect of the project.

Final presentation:

Each team will make a 15-minute presentation on their final project. This presentation should describe the pipeline that they used, how they structured their project, and the results of their analysis. In particular, each team should describe any issues they had in performing their reanalysis (i.e. were version numbers not included for software packages, could you not find a program or get it to work on the HPC, was meta-data missing), and what choices they made to get around those issues. Finally, teams should present a comparison of their final results to those of the original study and make an assessment of how well they were able to reproduce the original results.

Timeline:

Note that project milestones must be **completed** by these dates. Please don't leave them until the last minute, especially the check-in meetings since these may take a little while to schedule.

17 September: Project teams and proposed papers must be identified and posted to Slack. Teams will write a short (half page) project proposal that identifies the title of the paper, proposed reanalysis, the number and type of samples, and the estimated size of the data in Gb.
20 September: Project teams must have held an initial meeting with their faculty advisor discussing the feasibility of their proposed project.
3 October: Teams must post to Slack the web address of their GitHub project repo and the location of their group folder on the HPC (project file structure should be set up within this folder). All raw data should be downloaded to the HPC.
17 October: In-class project introductions and work plans.
5 November: Project teams must have held a 2nd check-in meeting with their faculty mentor.
4 December: ALL materials due by midnight (inc. final project & presentation)
5 December: Final project presentations, Part 1
10 December: Final project presentations, Part 2

Project Evaluation:

Projects will be worth 100 points in total, divided as follows:

20 points: Clear README file delineating everything we need to know about your project & project files
20 points: Final Jupyter notebook with a thorough and thoughtful comparison of your results to the published results
10 points: Timeliness - intermediate project milestones are achieved well and on time
10 points: Tidy and appropriate repository structure
10 points: Well-commented scripts
10 points: Comprehensive supporting files (e.g. envs/, logs/, raw data, etc)
10 points: Jupyter notebook(s) used for analysis and/or visualization of data
10 points: Final presentations