MB590-012 Microbiome Analysis
Final Project Guide

Description

In this assignment you will select a published paper for re-analysis. You must select a paper that has publicly available data, i.e., raw sequences and sample data in a public database, but NO R CODE. You can choose to fully replicate what the authors have done or, if you disagree with their approach, you can explain your disagreement and use a different approach (e.g., OTU vs. ASV definitions, GLM design, etc.). Papers cannot be from your current lab, any previous lab in which you have worked, or similar.

This assignment will build on the concepts and approaches you learned in class, while allowing you to focus more deeply on a topic of interest. In addition, you will gain experience with finding and reading scientific literature focused on analytical or computational methods. All topics must be approved in advance.

Approval and Submission Deadlines

Item	Due Date	Points
Paper/data proposal	Feb 9	25
Data successfully downloaded from SRA	Feb 22	5
Optional use for in-class practicum days	Mar 9, Apr 6	--
Final presentation	Apr 20	40
Final Rmd and knitted file	Apr 27	50

Proposal Materials to Submit

For your project proposal, create a GitHub folder called "Final Project Proposal" inside your repo. Submit the following items in that folder:

1. The original paper (pdf)
2. Your proposal (1-2 page pdf) including information on:
   a. Header with your name, date, and email
   b. Why this paper is a good choice
   c. Your plans for re-analyzing the data and how that differs from the original
   d. Confirmation that the sequences and metadata are available for you to carry out the re-analysis
      i. Include SRA accession numbers for the sequence data
      ii. Provide the link to the metadata if available (e.g., dryad, zenodo) or include the metadata file (.csv or .xlsx) in the GitHub folder
   e. Confirmation regarding R code
      i. Either confirm that NO code is available from the authors
      ii. Or confirm that your re-analysis plan is so different that the available R code is unusable
      iii. If (ii) is true, provide a link to the R code (e.g., dryad, zenodo) or include the .R or .Rmd file from the authors in your folder and refer to it here

Final Materials to Submit

As part of your final project, you should include the following in a GitHub folder called "Final Project" inside your repo:

1. The original PDF and any supplemental materials (e.g., metadata files)
2. R markdown ("YourLastName_Final.Rmd") and companion knitted HTML or PDF of the same name that includes the following (in order):
    a. Summary of data sources. If you used data from public databases such as the NCBI SRA, GitHub, Dryad, Zenodo, etc. provide those accession numbers and/or links.
    b. Summary of your original re-analysis plan and any changes, including any changes made due to feedback on your presentation.
    c. High-level summary of your findings (1-3 sentences)
    d. The step-by-step re-analysis with annotation to explain what you did and interpretation of the new results.
    e. Discussion (1-3 paragraphs) of your results in the context of the original published results, including whether/how and why your results did or did not differ from the authors' published results. If you used an entirely different approach, what did your approach add that the original authors missed?

Presentation

The last day of class will be used for final project presentations. Each presentation slot will be 15 min, with 13 min for the presentation and 2 min for questions. Your presentation should include a description of the original paper and why you selected it, an overview of your approach and how it differed from the original, and a comparison of your results with the original results. Save your presentation as a pdf file and upload it to your GitHub "Final Project" folder before class. Note that any feedback provided during the presentation can be incorporated into the final project you turn in a week later.

Grading

Your final project is worth 120 points. For each late day on approval or intermediate steps, 2 points will be deducted. For each day late on final submission, 5 points will be deducted. Exceptions will be made if an extension was requested and approved in advance. If you are discovered having obtained R code from the authors or through other avenues, you will have an automatic zero grade. The final project will be graded on content, accuracy, clarity, degree of difficulty, organization, and interpretation.

Dowloading data from the SRA

For this project, you are likely to download date from the NCBI Short Read Archive, which is regularly synchronized with the EBI European Nucleotide Archive and the DDBJ Sequence Read Archive. The following links will help you with these efforts.

Instructions for downloading from the SRA:
https://www.ncbi.nlm.nih.gov/sra/docs/sradownload/
https://www.ncbi.nlm.nih.gov/books/NBK47534/?report=reader

The SRA Toolkit:
https://github.com/ncbi/sra-tools/wiki
https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

You might also find the SRAdb R package useful:
https://www.bioconductor.org/packages/release/bioc/html/SRAdb.html
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-19