BIOST 509: Exercise 9 (optional, not graded)

Instructor: Amy Willis, Biostatistics, UW

Due date: None!

Instructions

This is a free-form assignment to give you some practice with reading in data, cleaning it up, reorganising rows and columns, manipulating data, and visualising data.

This dataset is from the paper *Vaginal microbiome of reproductive-age women* by Ravel et al, published in 2011 in the Proceedings of the National Academy of Sciences. The relevant dataset is "bv-vaginal-microbiome.xlsx", available from "Files/Datasets" or "Module 9 Materials" on Canvas.

This "exercise" is longer than the time allotted; it is intended to give you things to practice after this course ends. Please don't stress out about completing it today! I invite you to return to it at any time that you would like to practice your data analysis.

Solutions will be posted at 3:30pm today.

Questions

0. *Please take a moment to complete your evaluation for this course* at https://uw.iasystem.org/survey/214294. *I really appreciate your feedback on the course!*
1. Read in the dataset and inspect the data. What do you see that you want to fix?
2. Fix up the issues you noted in Question 1. Feel free to ask us for help if you don't know how. You could consider:
    - Renaming columns
    - Deleting empty/superfluous rows
    - Changing columns that contain numeric data to being numeric
3. `Race` (or `Racea`) is coded as a numeric variable, and the mapping between numbers and racial groups is given in the first/second row. Change `Race` to a character that reflects the coding given in the spreadsheet.
4. All columns between `L.iners` and `Yersinia` should be numeric. That's 330 columns! Change them all to numeric. *Don't do this one-at-a-time!*
5. Calculate the relative abundance of "L.iners" out of total counts. That is, divide `L.iners` counts by `Total Read Counts`.
6. Make a plot that shows the relative abundance of L.iners over time where the color of the points/lines shows the Patient ID. Describe the patterns you see in your plot.
7. Practice restructuring your data from wide to long format: make each row a subject-taxon combination, rather than showing all taxa side-by-side. I'm specifically refering to the columns between `L.iners` and `Yersinia`.
8. Calculate the relative abundance of all taxa (`L.iners` to `Yersinia`). *There are many ways to do this!* You could do this using either long or wide format of your data.