Homework 2 - Article Data

Gaea Daniel

9/20/2017

#### Find a journal article in your field and on a topic you’re interested in that has been published recently (in the last 5 years).

Fettweis, J. M., Brooks, J. P., Serrano, M. G., Sheth, N. U., Girerd, P. H., Edwards, D. J., ... & Vaginal Microbiome Consortium. (2014). Differences in vaginal microbiome in African American women versus women of European ancestry. Microbiology, 160(10), 2272-2282. <http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.081034-0#tab2>

#### Read through the paper along with any supporting files (supplementary documents, code, data, etc). Your goal is to try to find as many “yes’s” as you can to these questions: [I realize some fields of study are likely to have more yes’s than others so I know this will be challenging – just find as many as you can. I’m hoping you’ll get at least 1 yes.]

1. Did the authors clearly define all of their data sources? YES
2. How many data sources were there? [this is not a yes/no question] TWO
3. Did the authors provide instructions on how to request their data? NO
4. Did the authors provide a link to their data either pointing to a cloud-based resource or even as supplementary material with their paper? YES
5. Did the authors provide (in your opinion) enough detail that if you had their original data, that you would be able to reproduce their analyses? NO
6. Did the authors provide (in your opinion) enough detail on their data manipulation and analysis steps (i.e. software codes and procedures), that you would be able to reproduce their analyses? NO
7. Did the authors clearly define any bias, misrepresentations (i.e. under or over representation of their intended study population), or limitations in their data and analysis procedures? NO
8. Did the authors provide summary tables and/or graphics providing good descriptive statistical summaries of the data they used in their analyses? YES

#### Data Provenance – using the information you reviewed to find the answers to my questions above, write a paragraph or two at most on the origins and data management and manipulation trail the authors followed that led to their analyses and conclusions (i.e. what all did they say they did to the data and what do you think they did to the data that was maybe not well described in their final paper)

The authors defined two data sources, RDP classifer and Vaginal 16S rDNA Reference Database, for which they used as reference databases to determine the genus and species of the bacteria within the microbiome. While they provided a reference [available from the Short Read Archive at NCBI (project ID phs000256; Fettweis et al., 2011)] for accessing their data, the process requires specific credentials and an application. Types of statistical analyses used were listed, but there was not enough information provided to recreate the results even if the data were provided. Codes for data manipulation were not provided.

#### Data Summary and Presentation – write a paragraph summarizing how well (or not) the authors did presenting their underlying data (via tables, figures or text descriptions), including any limitations or biases, and whether or not you think these were adequate to support their analyses and conclusions.

The authors provided clear, concise tables and colorful figures that meaningfully displayed the data descriptions and results. Supplemental tables were also provided for descriptive information about the data used for some of the analyses. Color code keys were also provided in the supplemental information, which helped identify the less dominant bacterial species found in the vaginal microbiome.

#### Limitations – write a short paragraph on what you think was not well described or presented in the paper. If you were a reviewer, what additional information would you ask the authors to provide – especially if you would like to either (a) reproduce their findings or (b) take their data, code, findings to apply in your own work to advance your research interests. [Keep in mind this is not us being judgmental per se – many journals have very limited word count limits which makes it extremely hard to get all of these details in – often these important data source, manipulation, handling and processing details are thought to be boring and not germane to the findings so these sections are often way too short, which is why some journals allow supplementary materials to be uploaded… thus, the point of this homework 02 exercise…]

The analysis platform, STIRRUP, used with the RDP classifer was developed was developed by the author's team and was not publicly available. Additionally, when some of the analyses were conducted, particular samples were included. All of the chosen samples purposefully did or did not possess a specific characteristic, but it is unclear why some that fit the criteria were not included. As a reviewer, I would request access to some data along with corresponding codes for analysis to spot check the results for reproducibility. It is highly likely that the group owns the rights to STIRRUP, so I would ask for, but not require detailed submission or access. However, because this project appears to be solely funded by the NIH, the group of authors would have to agree to adequately assist other researchers as they explore this developing field.

#### Upload a copy of your summary report

The report is also available in the github repository "N736-HW2"