

Data Accessibility Update

Joanna Colovas

Lab Meeting 20240129

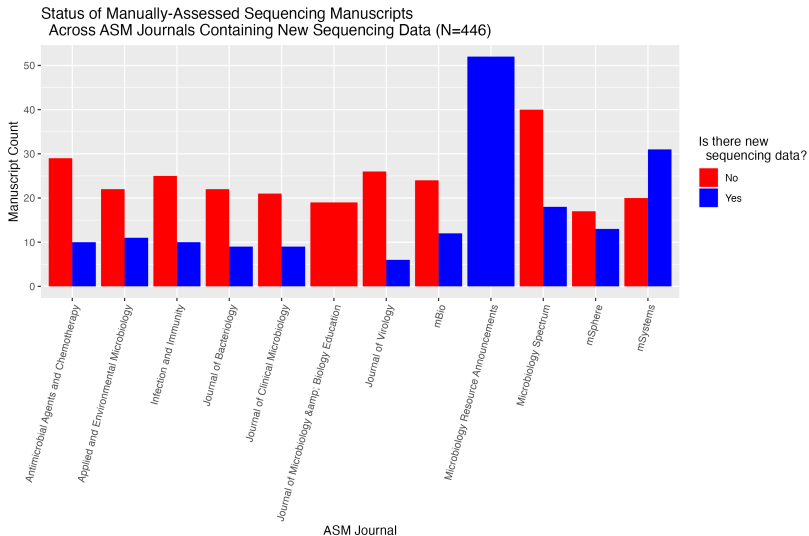
Project Goals

- ▶ General- Report statistics on the number of citations per paper as a function of data availability from the 12 ASM journals to answer question “Does making publication data available increase citation index of publications?”
- ▶ Proposal - Quantify the benefits of adhering to data accessibility policies for sequencing data at microbiology journals

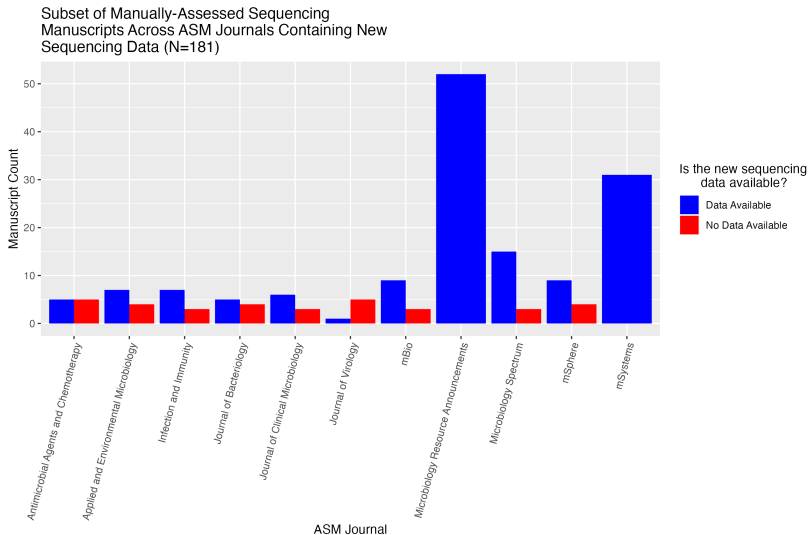
Completed Tasks

- ▶ Create “groundtruth.csv” file of N=446 papers with complete metadata from 12 ASM journals
 - ▶ Manually assessed each paper to determine if it was a “New Sequencing Paper” or not, and if “Data Available.”
- ▶ Creation of summary figures for the groundtruth dataset on its composition
 - ▶ Separation of papers by journal and by year based on data availability (N=181 with data available)
- ▶ API Key obtained for Clarivate Web of Science-starter API
 - ▶ Clarivate alternatives investigated:
 - ▶ CrossRef doesn't appear to return citation metric information
 - ▶ Scopus API from Elsevier should get citation metrics, has institutional API key available

Which journals contain papers with new sequencing data?

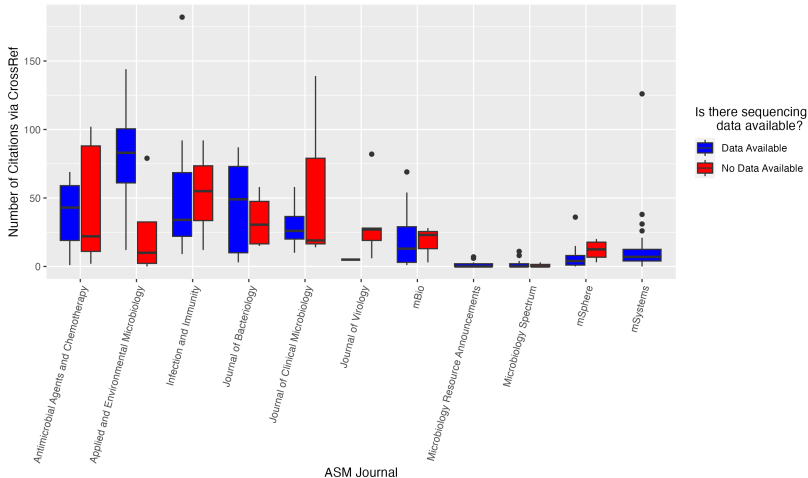


Of papers with new sequencing data, how many contain publicly available data?



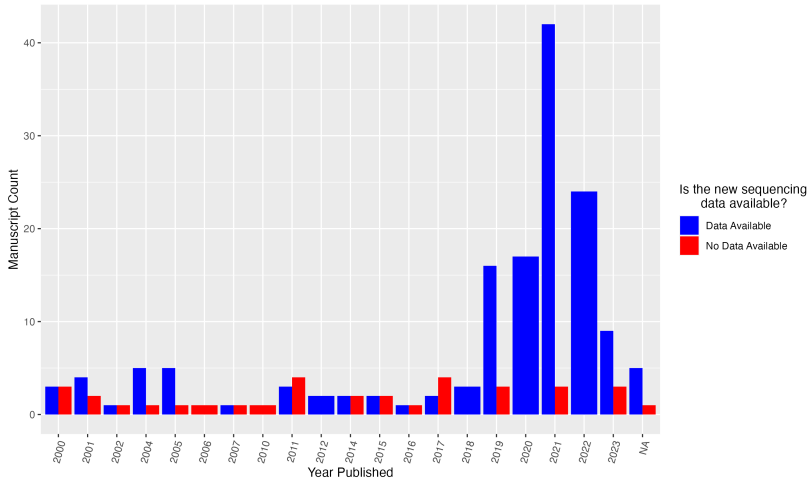
Do papers with sequencing data available have more citations?

Average Number of Citations for Subset of Manually-Assessed Sequencing Manuscripts Across ASM Journals Containing New Sequencing Data with Data Available (N=181)



Does the number of papers with data available change based on date published?

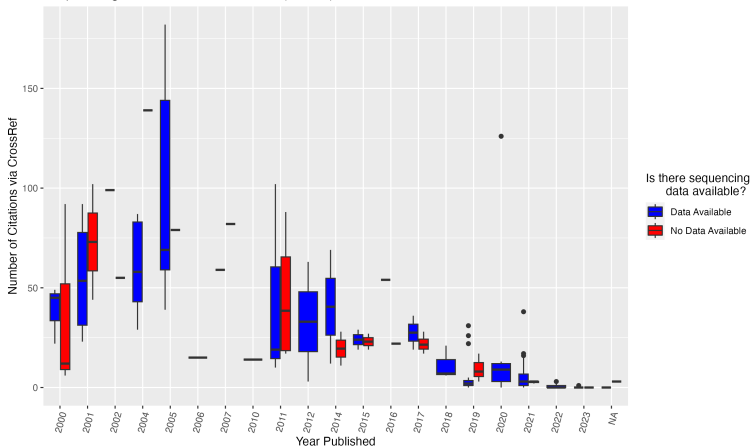
Subset of Manually-Assessed Sequencing Manuscripts Across ASM Journals Containing New Sequencing Data (N=181)



Do papers with sequencing data available have more citations based on year published?

- ▶ Do we need more papers for the training set based on the composition of the current training set?

Average Number of Citations for Subset of Manually-Assessed Sequencing Manuscripts Across ASM Journals Containing New Sequencing Data with Data Available (N=181)



Webscraping Status

- ▶ Test script works to eliminate table and figure captions using rvest and XML2! (Thanks Greg!)
- ▶ Some remaining issues for web scraping:
 - ▶ Headings still in scrape (h2 and h3)
 - ▶ Hyphen removal using [:punct:] smashes some words together
 - ▶ Some special characters still appear

Next Steps - Tidy Text

- ▶ Tackle converting example scraped text into tidy text format (one token/line)
 - ▶ `readLines()` %>% `unnest_tokens()`
 - ▶ allows for greater use of tidytext package, helps compute metrics and work with text
- ▶ Anti join with stop words df to remove common words
 - ▶ could also remove words < 3 letters/characters (AC did, would probably get rid of special characters)
 - ▶ can also make custom df of stop words to remove
- ▶ Does word frequency matter for each paper or just collectively for each larger grouping (ie contains new data or has data available)
 - ▶ if word frequency doesn't matter for each paper, we can remove duplicate words and make dfs smaller

Next Steps - Storage and GL

- ▶ Storage of tidy text data frames for each paper, dependent somewhat on size
- ▶ Purring for scraping and cleaning of all paper texts from the groundtruth dataset
 - ▶ GL scripting and polite package
 - ▶ ensure correct conda env is active
 - ▶ polite must be installed in R, not able to install using conda env