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?

Status of Manually-Assessed Sequencing Manuscripts
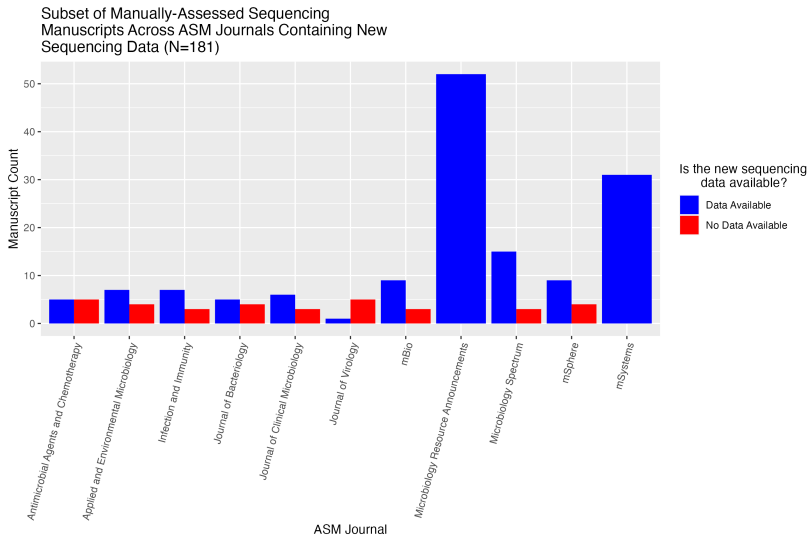Across ASM Journals Containing New Sequencing Data (N=446)
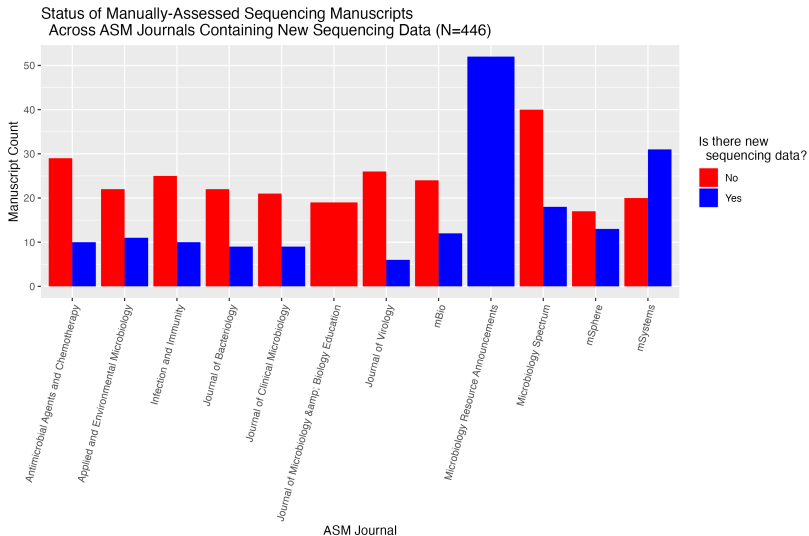
Distribution of papers with new sequencing data by year

Distribution by Year of Manually-Assessed Sequencing Manuscripts
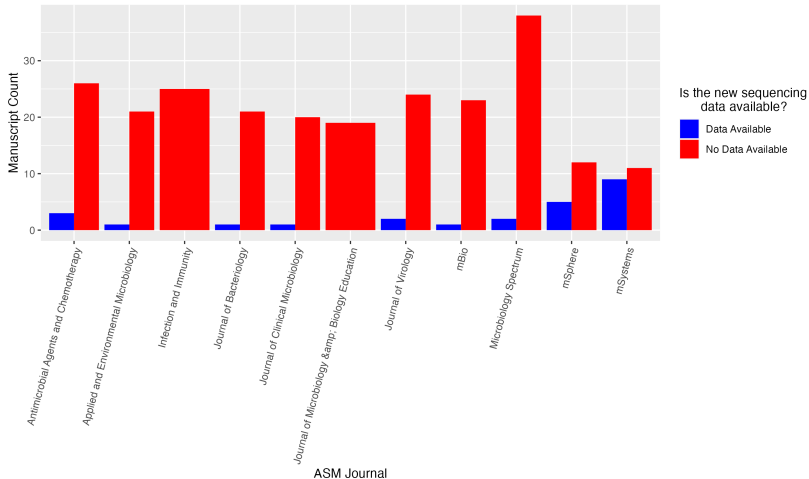Across ASM Journals Containing New Sequencing Data (N=446)
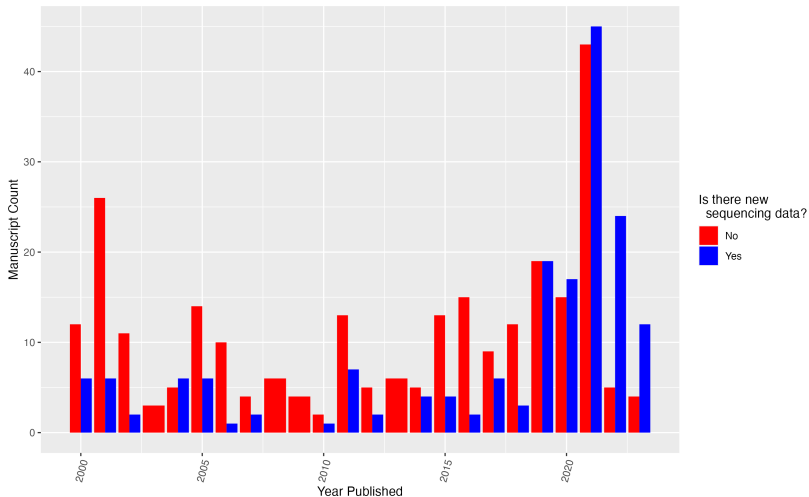
Of papers with new sequencing data, how many contain publicly available data?

Subset of Manually-Assessed Sequencing Manuscripts Across ASM Journals Containing New Sequencing Data (N=181)
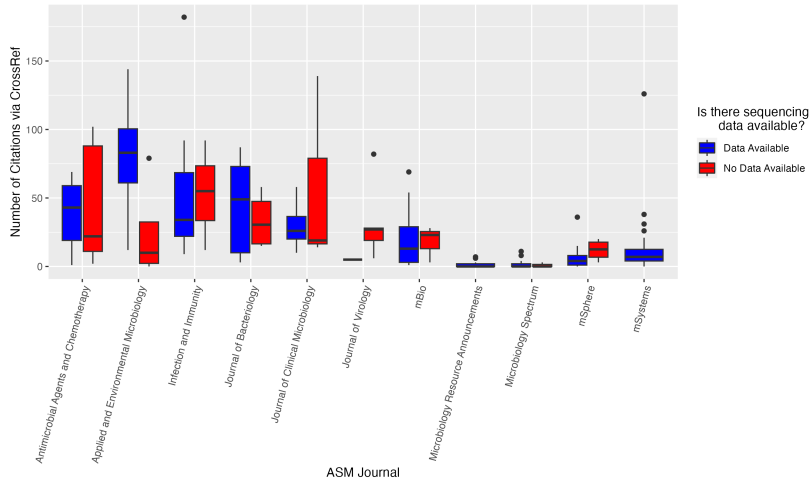
Distribution of papers WITHOUT new sequencing data

Subset of Manually-Assessed Sequencing Manuscripts Across ASM Journals Containing New Sequencing Data (N=265)
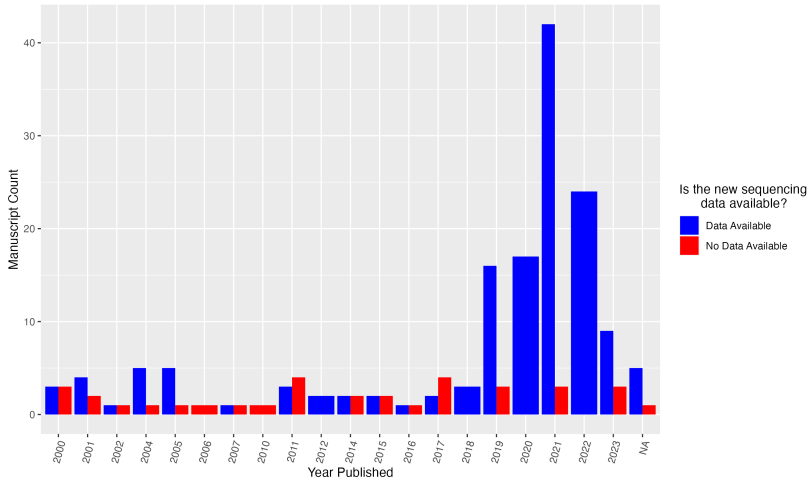
Do papers with sequencing data available have more citations?

Average Number of Citations for Subset of Manually-Assessed Sequencing Manuscripts Across ASM Journals Containing New Sequencing Data with Data Available (N=181)
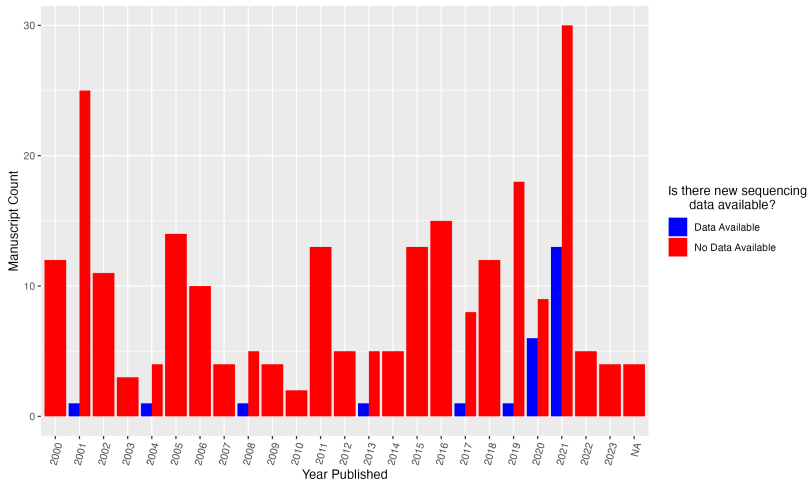
Does the number of sequencing papers with data available change based on date published?

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

Distribution of nonsequencing papers by date published
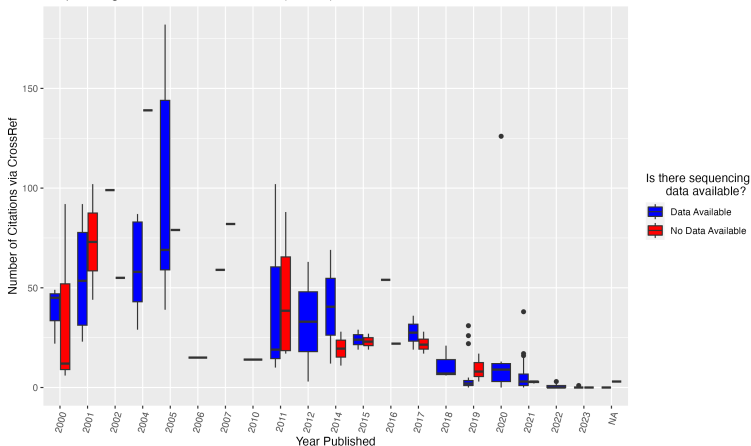

Subset of Manually-Assessed Sequencing
Manuscripts Across ASM Journals NOT Containing New
Sequencing Data (N=265)

Is there new sequencing
data available?

Data Available
No Data Available

Manuscript Count

Year Published

# Do new sequencing 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)

Do nonsequencing papers have more citations based on year published?

Average Number of Citations for Subset of Manually-Assessed Sequencing Manuscripts Across ASM Journals NOT Containing New Sequencing Data with Data Available (N=265)
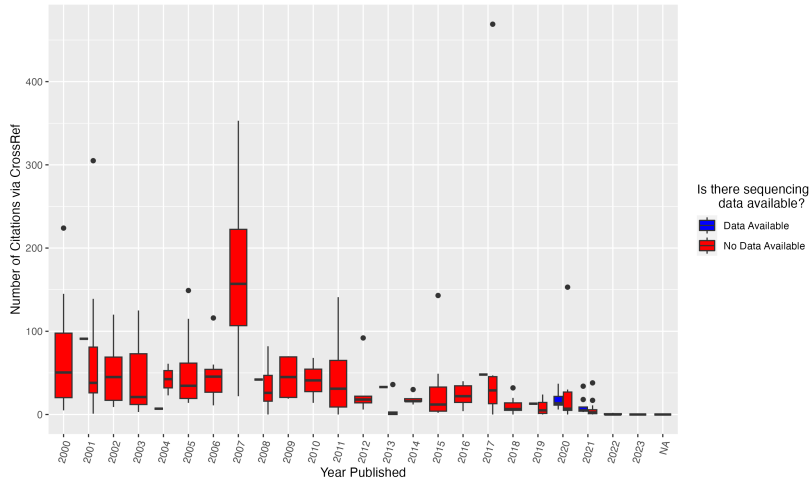
Webscraping Status

- 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
- Purrring 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

- ▶ Allison: When does ASM want data published? Is that consistent among journals?
  - ▶ Pat: no, not really consistent, we're trusting that if authors say that their data is on the SRA, it's actually there at that accession
  - ▶ authors in corporate settings often get away with "data available on request" which means data is really NOT available
  - ▶ might need to tighten up operational definition of data availability
    - ▶ data available on request is NOT available request, need to check AC/JVC coding of data to see if we coded incorrectly/kept track of this
    - ▶ variation may be driven by journal editors/reviewers
- ▶ Allison: do i have data on journal and year at once?
  - ▶ time since published and journal could be covariates
  - ▶ some journals have different goals (ie to be open access vs highly cited)
  - ▶ not all journals have been around the same length of time either

- ▶ does the training set need more papers?
  - ▶ why does it go back to the year 2000?
  - ▶ goal: need a training set to get a model to code for 20,000+ papers
    - ▶ want a representative training set of the actual data
    - ▶ does that focus on the last 5 years?
    - ▶ do we get equal papers for the last 20 years?
    - ▶ equal numbers by journal?
  - ▶ currently, roll with what we have for a training set with the addition to get back to 500 papers
    - ▶ pick papers randomly as best you can for underrepresented years and journals
    - ▶ look at AC 1300 papers set? does it have metadata already?
    - ▶ if we train the model on the training set and it's garbage, we can go back and work from there and start over or grab a ton more papers

- make pat figure of not new sequencing papers distribution (bar, boxplot)
- special character removal?
  - non-regex StringR cheatsheets (see back) to help you match character expressions
- can revisit link rot question if you want to poke your eyes out
  - will still need to get all of the paper text (with HTML tags tbh)
  - look for external links
    - this could actually help check for accession numbers that are live/viable