Data Accessibility Update

Joanna Colovas

Lab Meeting 20240129

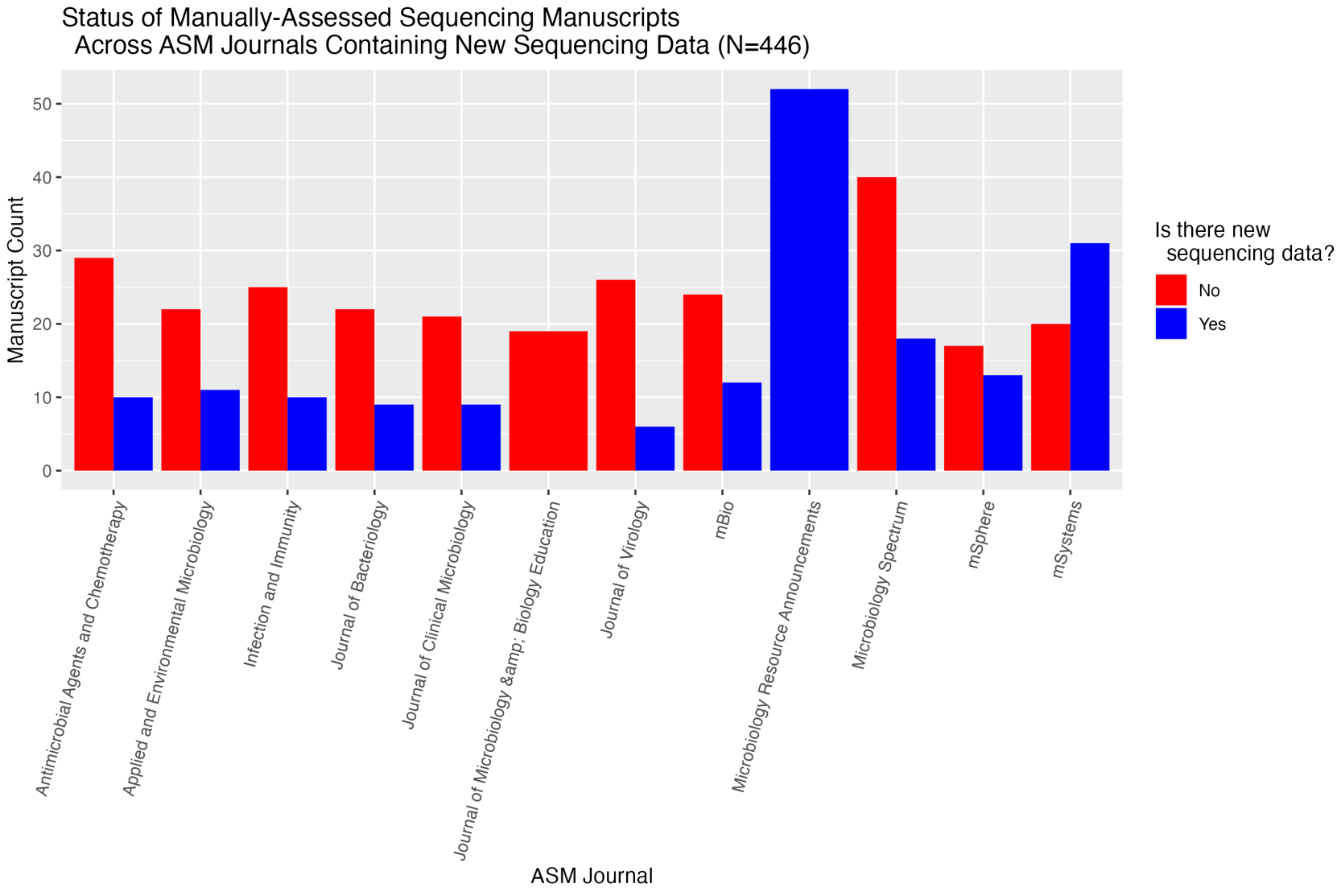
Project Goals

- 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

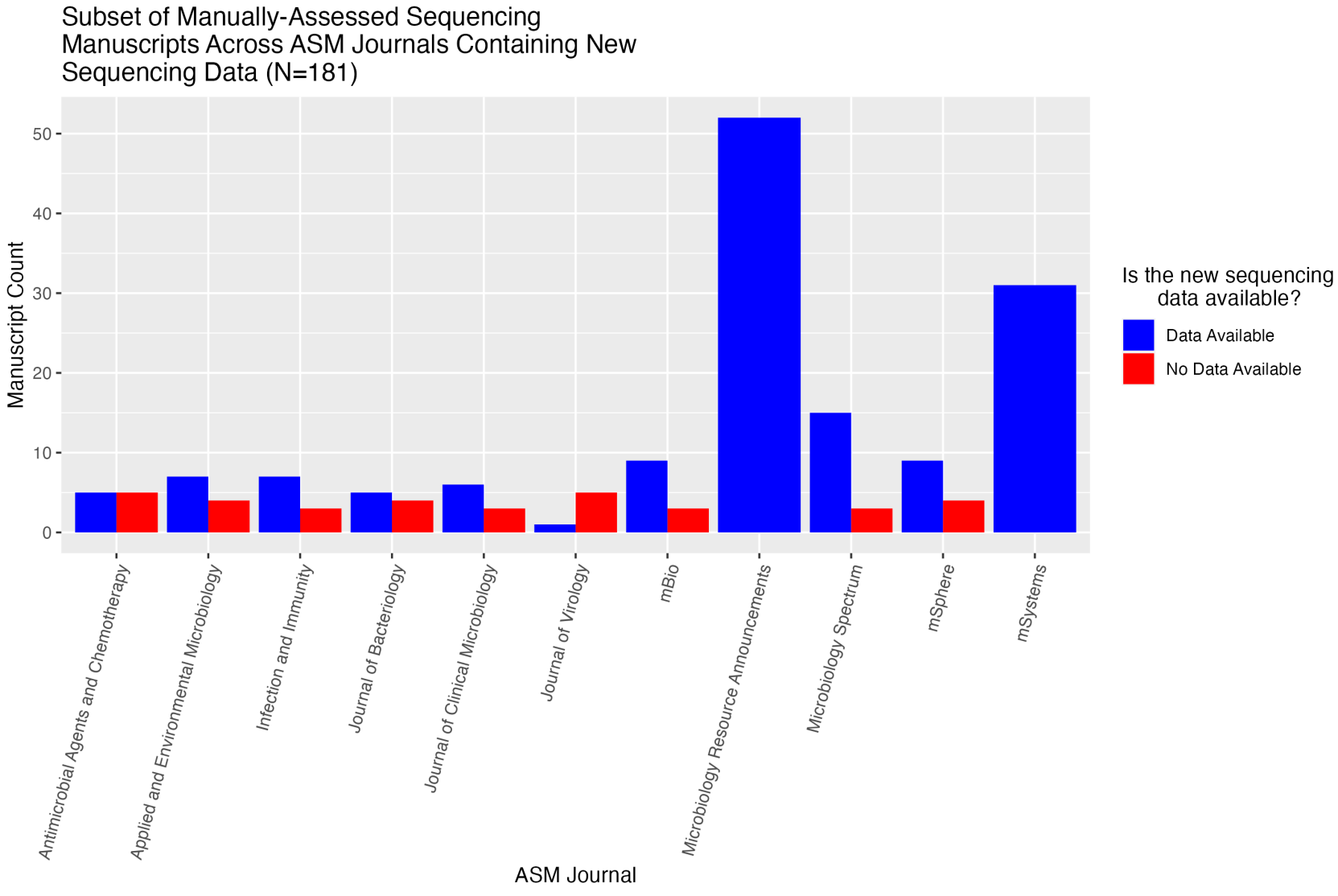
Completed Tasks

- 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

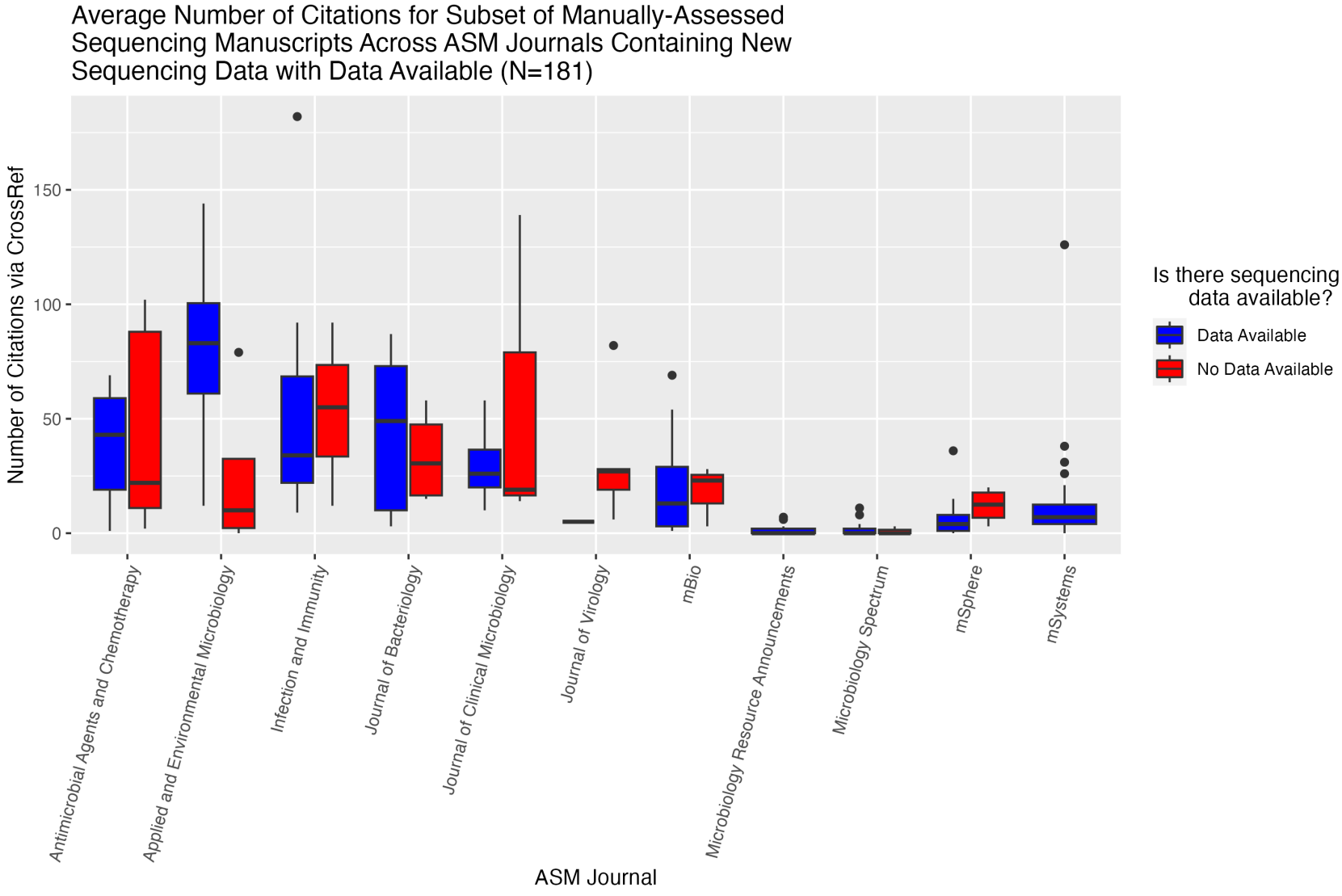
Which journals contain papers with new sequencing data?



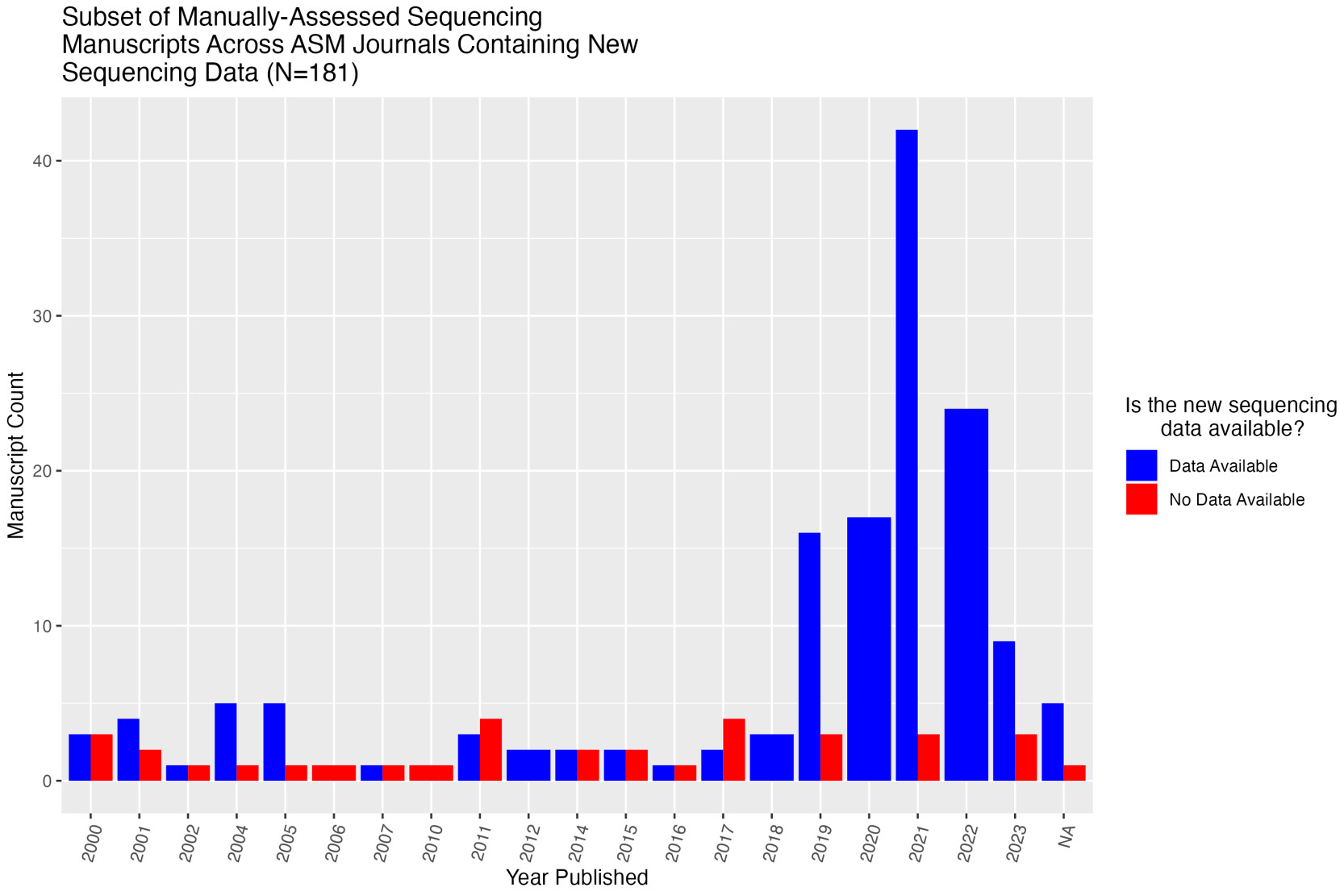
Of papers with new sequencing data, how many contain publicly available data?



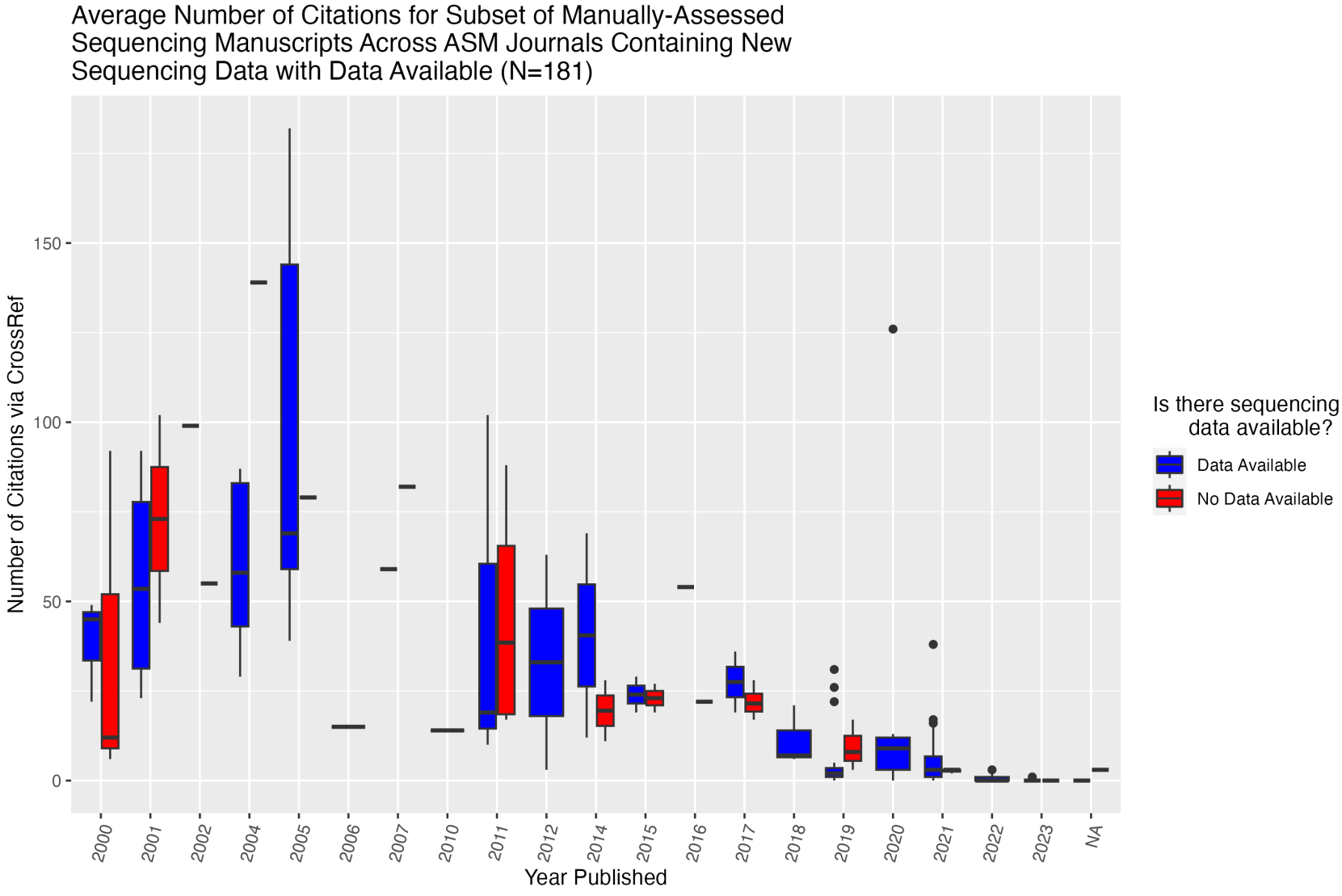
Do papers with sequencing data available have more citations?



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



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



Current and Next Steps

- Web scraping with the elimination of figure and table captions using package rvest
  - Issue: current selectors do not eliminate figure and table captions
  - Investigate rvest functions for correct "not" operator
  - Comb page HTML tags for more precise selectors of what we DO want from the web page
  - Several more test papers from ASM to determine if function works as intended
- Do we want the text scraped without HTML tags or keep them?
  - Does removing them later pose a problem?
- Do we need more papers for the training set based on the composition of the current training set?