

# Data Accessibility Paper

Joanna Colovas

Adena Collens

Patrick D. Schloss

Oct 2, 2025

- Abstract

- Importance

- Incentivize authors to publish/make available their original data
- Publishing data helps get more use out of research
- Helps eliminate file drawer effect as it shows negative data

- Keywords

- Data accessibility
- Data reproducibility

## Introduction

### Scientific Data as a Public Good

The United States Government spent over two hundred million dollars (USD) in 2024 on research expenditures ((1)). The result of all of these investments are data, paid for in part by taxpayers. Therefore, data is a public good. Public goods, for example public libraries, are able to be used by anyone without barrier to entry, and without diminishing the use of others. Data is available to be used by others, and best used as for the benefit of those who provided the funds for it. Once data has been generated, it can be used not only for initial analyses, but over and over again in future studies or meta analyses. A key tenet of the scientific method is the ability to replicate scientific

findings to ensure that they are not due to error. One way that scientific findings can be replicated is by re-completing the same analyses by another researcher. This is only possible if the data used to complete the original analyses is available for use. Additionally, data can be used to eliminate possible solutions to a problem by the publishing of negative or non-significant data. Thinking of data as a public good, if negative data is published, it can help researchers avoid sinking time and financial resources into the investigation of non-viable hypotheses. Unfortunately, researchers have few incentives to publish negative or non-significant results. This lack of publication of non-fruitful investigation is more commonly known as the “file drawer effect” ((2), (**rosenthal\_1979?**)).

Building on the file drawer effect, data availability (DA) is the practice of making raw experimental data, such as nucleic acid sequences, and analyses publicly accessible, often via upload in maintained databases. These databases are public “commons” and help widely benefit when they are well maintained and well used. With the latest and greatest methodologies available across fields, increasing amounts of data are being generated each day, especially in the biological sciences ( (3), (4)). Availability of large quantities of study data and metadata (data about data) is a necessary resource for appropriate use and re-use of data and protocols as well as the recreation of analyses. DA is a deeply important component of the scientific process in the digital age, and curation of digital records is a slowly emerging topic in data science( (**howe\_2008?**)). Available data and analyses are the gold standard for recreation of studies and replication of their results. Not only is replication a worthy goal, but large datasets are often underutilized, and can continue to provide benefit and resources to researchers via their re-use towards investigating and answering further questions. The National Institutes of Health (NIH) has called for grant proposals for the creation, enhancement, and maintenance, of new and existing data repositories ( (5), (6)).

### **Nucleic Acid Sequencing Efforts**

Beginning in 1996 with the International Strategy Meeting on Human Genome Sequencing in Bermuda, researchers have prioritized the release of all human genome sequencing information so that it may “maximize its benefit to society” ((7)). The meeting participants agreed that “primary

sequence data should be rapidly released”, with “sequence assemblies [to] be released as soon as possible, in some centres[sic], assemblies of greater than 1 kb would be released automatically on a daily basis”, and that “finished annotated sequence should be submitted immediately to public  
50 databases” ((7)). The “Bermuda Principles,” as they became known, have been embraced by the large scale Human Genome Project (HGP) since 1998. The HGP endeavored to publish the first complete human genome. In 2003, another meeting, held in Ft. Lauderdale, FL, re-affirmed the 1996 Bermuda Principles, expanded upon them to apply more broadly towards sequencing data, and called for further support of these practices ((7)). These foundational agreements set  
55 the stage for both the HGP and the Human Microbiome Project (HMP) to generate and share massive amounts of data over the course of their studies ((8), (9), (10)). The goal of the HMP was to sequence all body sites to determine the microbes found on and in the human body.

Starting with major projects such as the HGP and HMP, nucleic acid sequencing efforts have been commonly uploaded and released using public databases. This allows for researchers to use and  
60 re-use the data from the HGP and HMP. There are three major databases worldwide to support sequencing and sharing efforts. The National Library of Medicine’s (NLM) National Center for Biotechnology (NCBI) in the United States, the Research Organization of Information Systems’ (ROIS) National Institute of Genetics (NIG) in Japan, and the European Molecular Biology Lab’s (EMBL) European Bioinformatics Institute (EBI) in Europe. These three databases are part of the  
65 International Nucleotide Sequence Database Collaboration (INSDC) ((11)) . These large databases make comparative research possible. Genetic lineages of microbes are determined by creating phylogenetic trees which compare a new sequence to existing sequences. Phylogenetic trees show how closely related a new microbial genetic sequence is related to others studied before both in terms of evolution and mutation and in structure and function. Comparative genetics and genomics  
70 would not be possible without strong community commitment to data availability.

## Current DA Policies

Current data availability guidelines have been informed by a number of policies created by funding agencies, peer-review journals, conference and special task groups, as well as community interest groups. In 2011, after the Future of Research Communication (FoRC) conference in Germany, 75 scientists and others came together to establish FORCE11, a community interest group which seeks to encourage and promote data availability standards ((12)) . Also in 2011, the Genomic Standards Consortium (GSC) published a set of standards in *Nature Biotechnology* to promote the publication of the “minimum information about a marker gene sequence” (MIMARKS) or “minimum information about x sequence” (MIXS) ((13)). These standards are checklists usable by data generators and 80 uploaders towards inclusion of relevant data with sequence uploads in the International Nucleotide Sequence Database Collaboration (INSDC). Some checklist items include if the data were published to an INSDC database and metadata about the study systems, data collected, and authors. An important factor is the ability to link the data to the results and to the data generators.

In 2014 the FORCE11 group published the Joint Declaration of Data Citation Principles (JDDCP), a 85 document working towards the standardization of data citation and its future availability((14)). Some of the JDDCPs include crediting the authors of the data, providing data with unique identifiers, and the persistence of the available data. The Findable, Accessible, Interoperable, and Reuseable (FAIR) data science guiding principles that were put forth in 2016 in *Nature Scientific Data* urges readers to “improve the infrastructure supporting the reuse of scholarly data” ((15)). The FAIR 90 principles are often cited by NIH in funding calls for strong data science practices((5)). In 2021, a *Nature Medicine* publication put forth the “Strengthening the Organization and Reporting of Microbiome Studies” (STORMS) checklist to help authors identify report-worthy elements of their data and metadata ( (16)). Some items on the STORMS checklist include the sequencing method used in the study, the study design, and physical location of the study. Unfortunately, none of these 95 DA principles or checklists are enforceable by any agency. The National Institutes of Health (NIH) began enforcing the “Policy for Data Management and Sharing” (NOT-OD-21-013) in January of 2023, requiring all NIH funded studies to submit a data management and sharing plan (DMS) with

their funding applications, and comply with their DMS plan after generation and publication of the funded work((17)). A DMS plan includes detailed descriptions of data that will be generated in a study, related tools, standards, and data preservation plans. Non-compliance with NOT-OD-21-013 is identified by funding agencies during annual Research Performance Progress Reports (RPPRs), and may impact future funding decisions ((17)). We also believe that statements of data “available on request” are not sufficient to be considered available data. One example was published in *Microbiome*, that a reader in search of data may email the corresponding author, with varying results ((18)).

With the advent of next generation sequencing, microbiology research has generated large amounts of sequencing data, and it is common to upload sequence data to a public repository as well as to include data in research publications. Because of this, we set out to determine rates of published raw data in the biological sciences, specifically microbiology, by examining the American Society for Microbiology’s (ASM) library of published primary research journals. We were also interested in the relationship between DA and number of citations received by a published work.

### **American Society for Microbiology Journals (ASM)**

The American Society for Microbiology(ASM) is the major professional body recognized by microbiologists. They have eighteen journals, thirteen primary research journals, three review journals, and two archive journals. In addition, several journals have been folded into others or renamed over time. The ASM family of journals requires that authors “make data fully available, without restriction, except in rare circumstances” ((19)). They have adapted this policy from journals *Microbial Genomics* and *PLOS*. In the ASM open data policy they describe the use of a “Data Availability Statement” which includes “data description, name(s) of the repositories, and digital object identifiers (DOIs) or accession numbers” and encourages publishing data on relevant public repositories ((19)). Consequences of non-compliance to the ASM open data policy include contacting research article authors to inform of non-compliance, publication of an “Expression of Concern” for the author and their compliance issues, sanctions on publication in ASM journals, as

well as contacting the affiliated research institution and/or funding agencies of the authors ((20)).

125 We endeavor to evaluate how well the microbiology community is using reproducible data practices as we believe that this group of researchers will be early adopters of the technologies available as a result of both the ASM and NIH policies towards data availability.

## Data Availability Case Studies

The availability of datasets also allows new questions to be answered with existing data or the  
130 combination of multiple datasets, such as the use of the Human Microbiome Project's (HMP) sequencing data by researchers to create over 650 scientific publications ((21)), and the completion of metadata studies, including those efforts participated in by these authors ( (22), (23), (24)).

An important tool for creating phylogenies is the NCBI Basic Local Alignment Search Tool (BLAST)  
((altschul1990?)). The BLAST algorithm allows users to compare a nucleic acid or protein  
135 sequence to the NCBI database of over 1TB of data to find similar and related sequences. Without the upload of sequences to the NCBI database, the use and success of BLAST would not be possible, despite the effort required on part of the researcher to upload of sequences to one of the INSDC databases.

Availability of data contributed to the rapid sequencing of the SARS-CoV-2 virus during the 2020  
140 pandemic and subsequent expedition of vaccine development ((25)).

With microbiologists commonly uploading nucleic acid sequences to public databases, the aim of this study was to determine the current state of data availability in twelve primary research journals from the ASM family of journals. Primary research articles were classified with using two machine learning models to answer two questions; "Does this paper contain new sequence data?", and  
145 "Is the data available?" Once these questions were answered, we moved to statistical analyses to answer these and further questions, such as "How does making my data available impact my citation metrics over time?" We were interested in citation metrics as a concrete metric to examine how making DA benefits researchers and as a possible incentive towards making DA.

## Results

### General Description of the Experiment

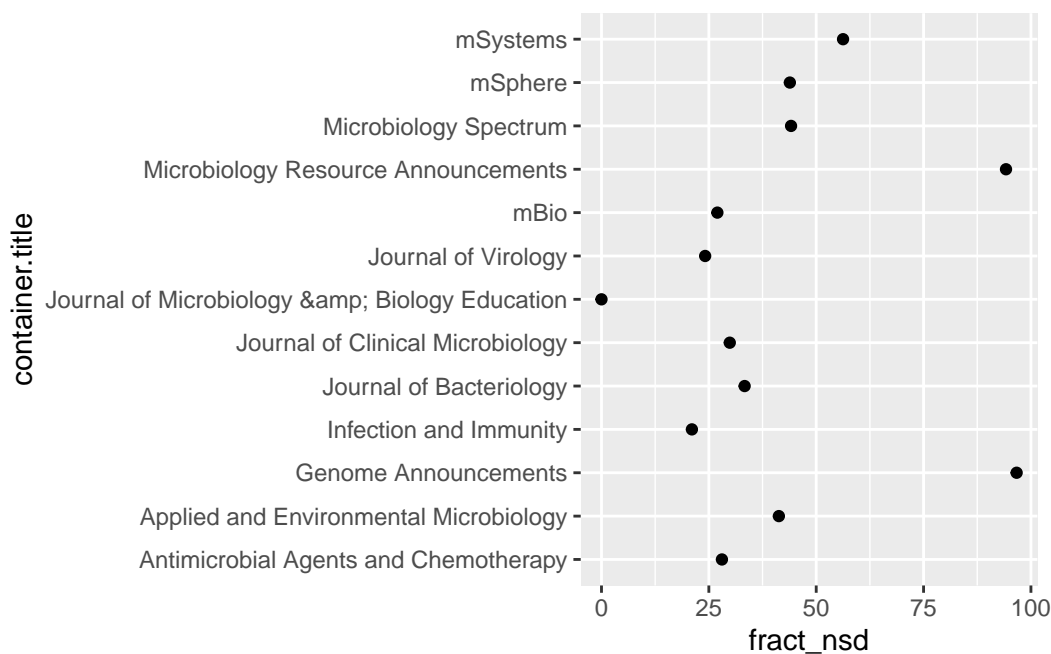
We set out to determine the current state of data availability in twelve primary research journals from the ASM family of journals. This objective was completed by first acquiring all papers published in the journals of interest between 2000 and 2024 using the Crossref database and command line tools. We then trained two random forest machine learning models to differentiate if each paper 1), contained “New Sequencing Data” (NSD) and 2) if the paper had “Data Available” (DA). To avoid overfitting the models, we trained each model multiple times, performing validations on a subset of data after each iteration. This allowed us to have a greater number of papers in the training dataset, as well as to have great accuracy and precision within our models. Using our trained models, we were able to classify over 150,000 papers from the whole dataset to determine if they were NSD or contained DA. After this, we could perform statistical modeling to describe the data, and generate summary statistics. We were especially interested in the ways in which NSD and DA impact citation metrics.

### ASM Journals

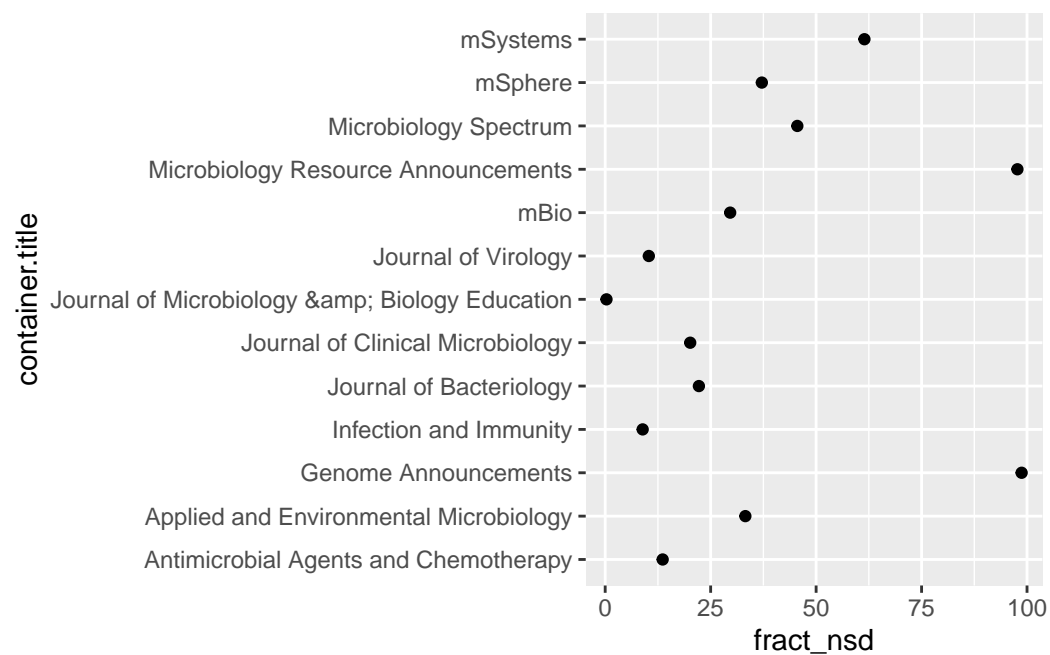
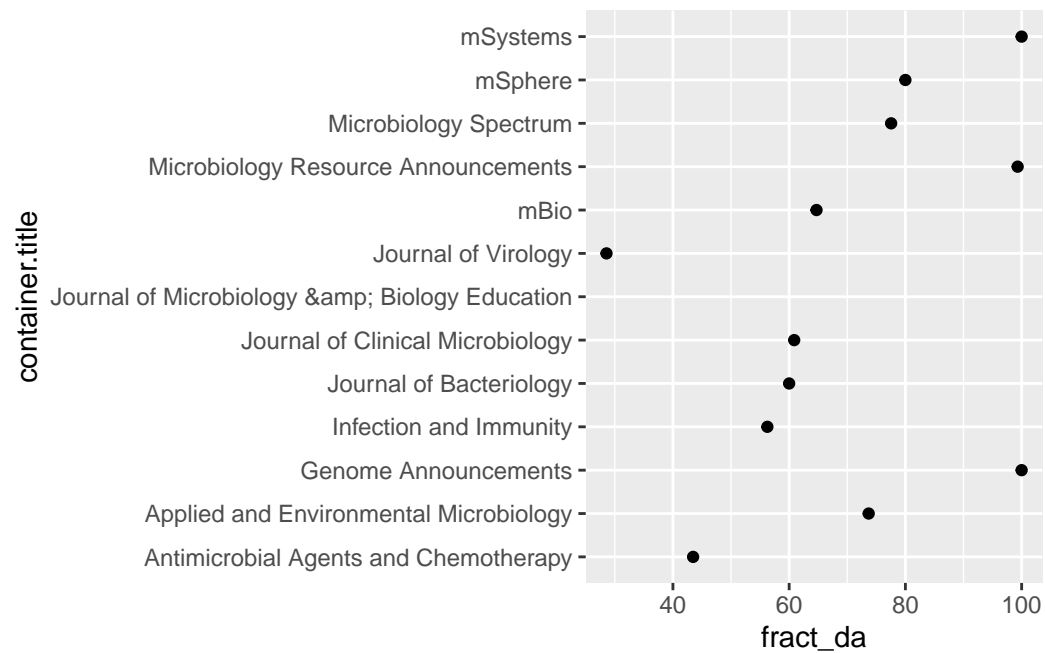
We used twelve of the ASM primary research journals in this study. Of note, several journals had changes to their publication goals during the 2000-2024 time period. The *Journal of Bacteriology* was the primary journal to publish new genome announcements until 2013 when ASM announced journal *Genome Announcements* as a more permanent home for this type of data. *Genome Announcements* was active from 2013 until 2018, when it was re-branded to *Microbiology Resource Announcements*, which has been active from 2018 until present. These two journals appear separately in our analyses due to the organization of the Crossref database. Not all journals are equally likely to contain NSD and have sequencing DA as a result of their field of interest. Journals publishing new genome announcements have a high percentage of NSD papers, and a high percentage of DA within those papers. As a result, since the introduction of *Genome*

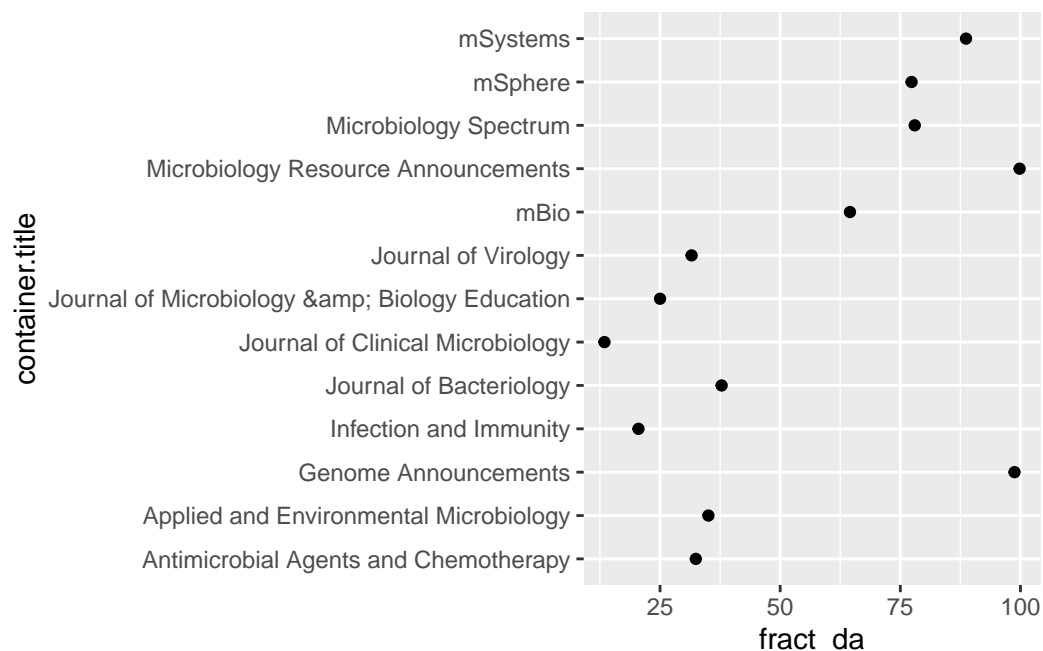
Announcements and Microbiology Resource Announcements, the *Journal of Bacteriology* has had fewer NSD papers, and fewer papers with DA since 2013. Another journal of note is *Microbiology Spectrum* and its re-brand. From 2013 until the fall of 2021, *Microbiology Spectrum* was a review journal. After this point, *Microbiology Spectrum* became a primary research journal ((26)). Review journals are less likely to publish articles with NSD, and to have DA. Several journals, including *Microbiology Spectrum*, do not span the entire time period for the study. Journals *mBio* (b.2010), *Microbiology Spectrum* (b. 2013, re-brand 2021), *mSphere* (b. 2016), *mSystems* (b. 2016), and *Genome Announcements* (2013-2018). Journals *Antimicrobial Agents and Chemotherapy*; *Infection and Immunity*; and the *Journal of Microbiology and Biology Education*; are all less likely to contain NSD and have DA than the other journals in the dataset. Therefore, aggregate measures of all journal activity together are thought to be less accurate than those depicting each journal separately.

## Descriptive Statistics









## Whole Dataset

Using the Crossref database of DOIs, with validation from the Web of Science, NCBI, and Scopus DOI databases, we downloaded N = 154720 unique records of papers published in ASM journals between January 1st, 2000 and December 31st, 2024. These papers came from journals *Applied and Environmental Microbiology*; *Antimicrobial Agents and Chemotherapy*; *Infection and Immunity*; *Journal of Clinical Biology*; *Journal of Virology*; *Journal of Bacteriology*; *Journal of Microbiology and Biology Education*; *Microbiology Resource Announcements* (formerly known as *Genome Announcements*); *mSystems*; *mSphere*; *mBio*; and *Microbiology Spectrum*. After downloading the HTML content of each paper, we cleaned the HTML content and readied it to apply our machine learning models to classify each paper. Overall, 26.9429% of papers had NSD, and 58.8615% of papers with NSD, had DA. See Figure 2XX for percentages of NSD and DA for each journal. The journal with the highest rate of NSD was *Genome Announcements* at 98.7191%, and the lowest was *Journal of Microbiology & Biology Education* at 0.3065%. The journal with the highest rate of DA was *\*Microbiology Resource Announcements\** at 99.8259%, and the lowest was *\*Journal of Clinical Microbiology\** at 13.4393%. This was expected as *\*Genome Announcements\** publishes

mainly new genomic sequence data and makes the data available. On average, papers in the dataset had a median of 25 citations/article. This number varies by journal, see table XXXX for data by journal. The journal with the highest median rate of citations/article was \*Applied and Environmental Microbiology\* at 38%, and the lowest was \*Microbiology Resource Announcements\* at 1%. The journals in the dataset span years 2000-2024. See table XXXX for the distribution of papers per year in the dataset.

### Training Dataset

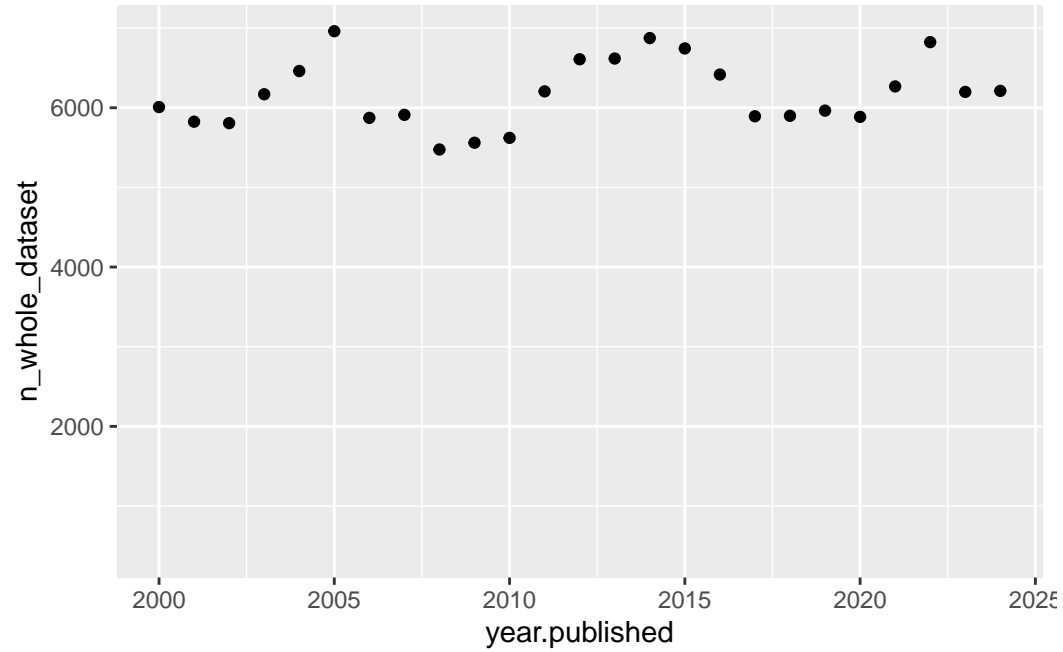
We created a subset of the whole dataset to train our machine learning models. The training dataset initially had N = 500 papers, but was increased over time due to gaps in the dataset, and after subsequent validation of the trained models (see below), a total of N = 1045. XXSee Figure 1 for the distribution of papers per journal in the training dataset.XX The journals in the dataset also span years 2000-2024. See Figure 5XX for the distribution of papers per year in the whole and training datasets. Overall, 43.7321% of papers had NSD, and 80.9628% of papers with NSD, had DA. See Figures 3 and 4 for percentages of NSD and DA for each journal. The journal with the highest rate of NSD was \*Genome Announcements\* at 96.6667%, and the lowest was \*Journal of Microbiology & Biology Education\* at 0%. The journals with the highest rate of DA in NSD papers were \*Genome Announcements and mSystems\* at 100%, and the lowest was \*Journal of Virology\* at 28.5714%. On average, papers in the dataset had median 10 citations/article. This number varies by journal, see table XXXX for data by journal. The journal with the highest rate of citations/article was \*Infection and Immunity\* at 40, and the lowest was \*Microbiology Resource Announcements\* at 0.

### Year Published Distribution Table

Table 1: Distribution of Year Published for Whole and Training Dataset

year.published	n_whole_dataset	n_training_dataset
2000	6009	30
2001	5825	43
2002	5807	23
2003	6170	22
2004	6461	21
2005	6961	36
2006	5873	29
2007	5911	18
2008	5476	12
2009	5561	13
2010	5622	21
2011	6206	41
2012	6609	24
2013	6618	29
2014	6875	31
2015	6745	34
2016	6417	38
2017	5893	40
2018	5899	40
2019	5964	63
2020	5886	65
2021	6268	119
2022	6824	70
2023	6199	57

year.published	n_whole_dataset	n_training_dataset
2024	6212	116
NA	429	10



## Descriptive Statisitcs about the Trained Models

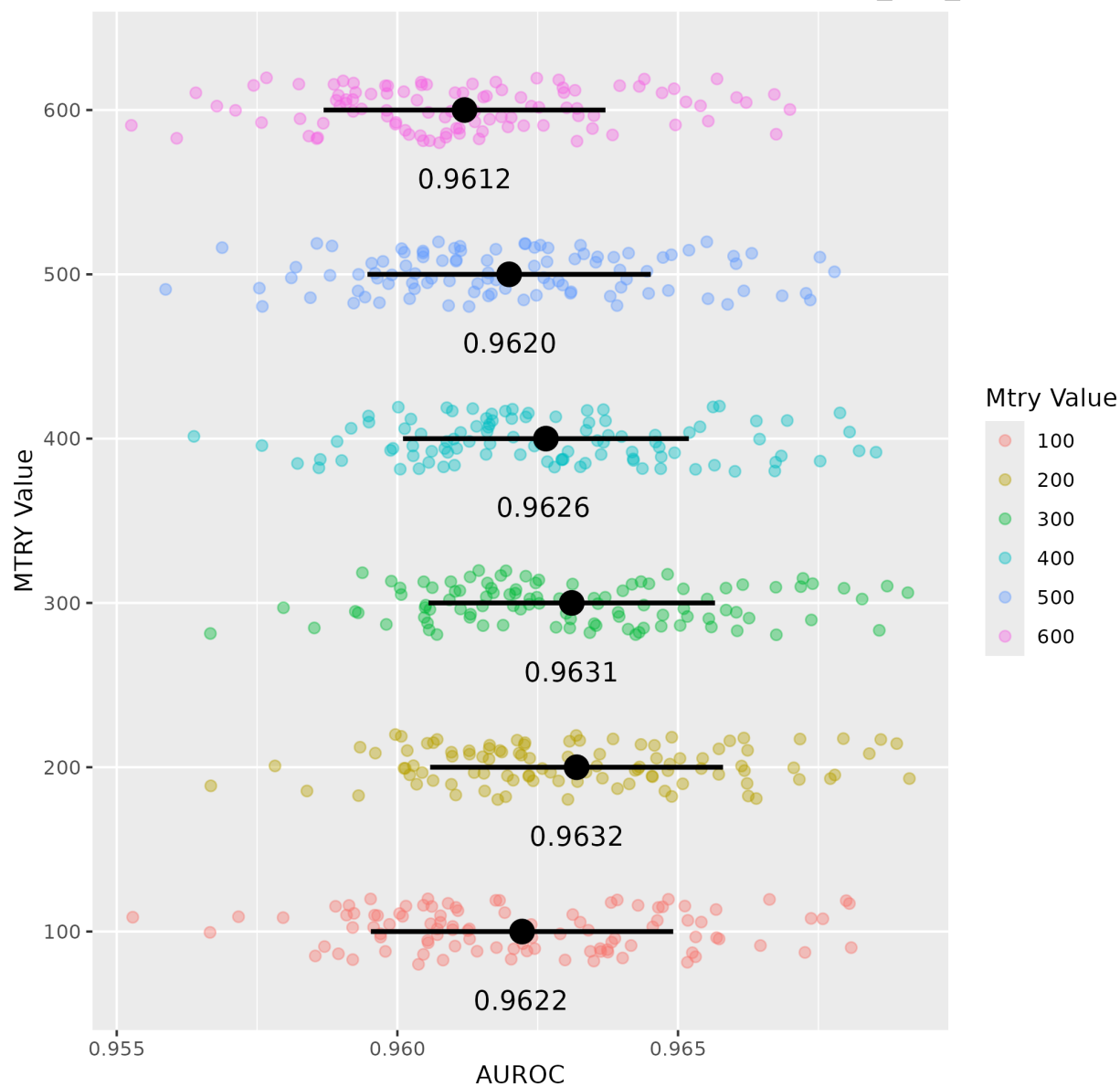
Table 2: Trained Model Summary Statistics

key	da_model	nsd_model
mtry	300.0000	200.0000
logLoss	0.1864	0.2848
AUC	0.9878	0.9645
prAUC	0.9468	0.9465
Accuracy	0.9464	0.9036
Kappa	0.8826	0.8029

key	da_model	nsd_model
F1	0.9586	0.9160
Sensitivity	0.9610	0.9350
Specificity	0.9199	0.8631
Pos_Pred_Value	0.9565	0.8984
Neg_Pred_Value	0.9290	0.9128
Precision	0.9565	0.8984
Recall	0.9610	0.9350
Detection_Rate	0.6204	0.5257
Balanced_Accuracy	0.9404	0.8991
logLossSD	0.0153	0.0206
AUCSD	0.0051	0.0107
prAUCSD	0.0131	0.0127
AccuracySD	0.0145	0.0189
KappaSD	0.0317	0.0388
F1SD	0.0112	0.0162
SensitivitySD	0.0163	0.0233
SpecificitySD	0.0294	0.0364
Pos_Pred_ValueSD	0.0152	0.0240
Neg_Pred_ValueSD	0.0277	0.0283
PrecisionSD	0.0152	0.0240
RecallSD	0.0163	0.0233
Detection_RateSD	0.0105	0.0133
Balanced_AccuracySD	0.0165	0.0199

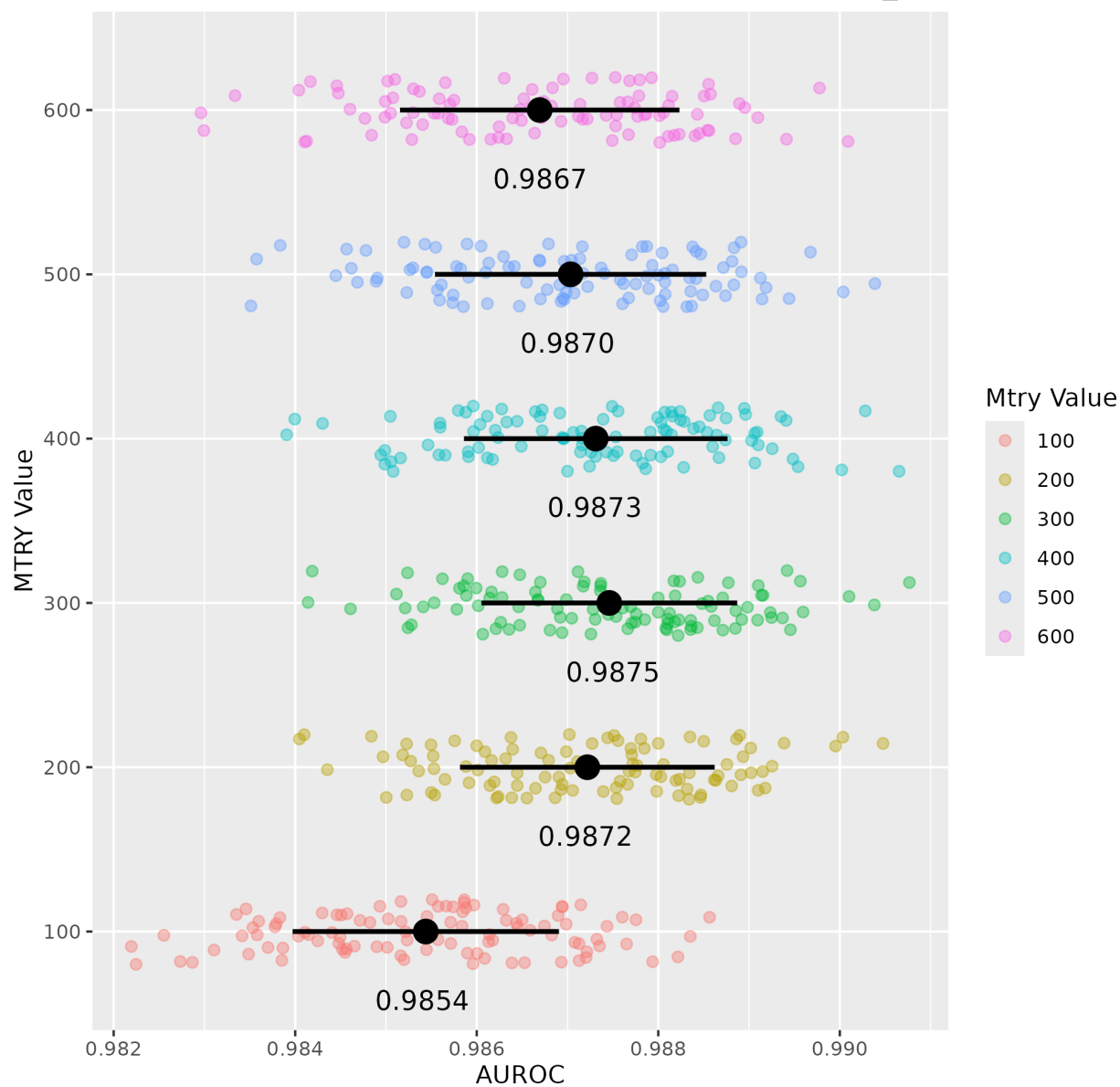
**Figures for trained models**

AUROC of Training Set by MTRY Value for model new\_seq\_data

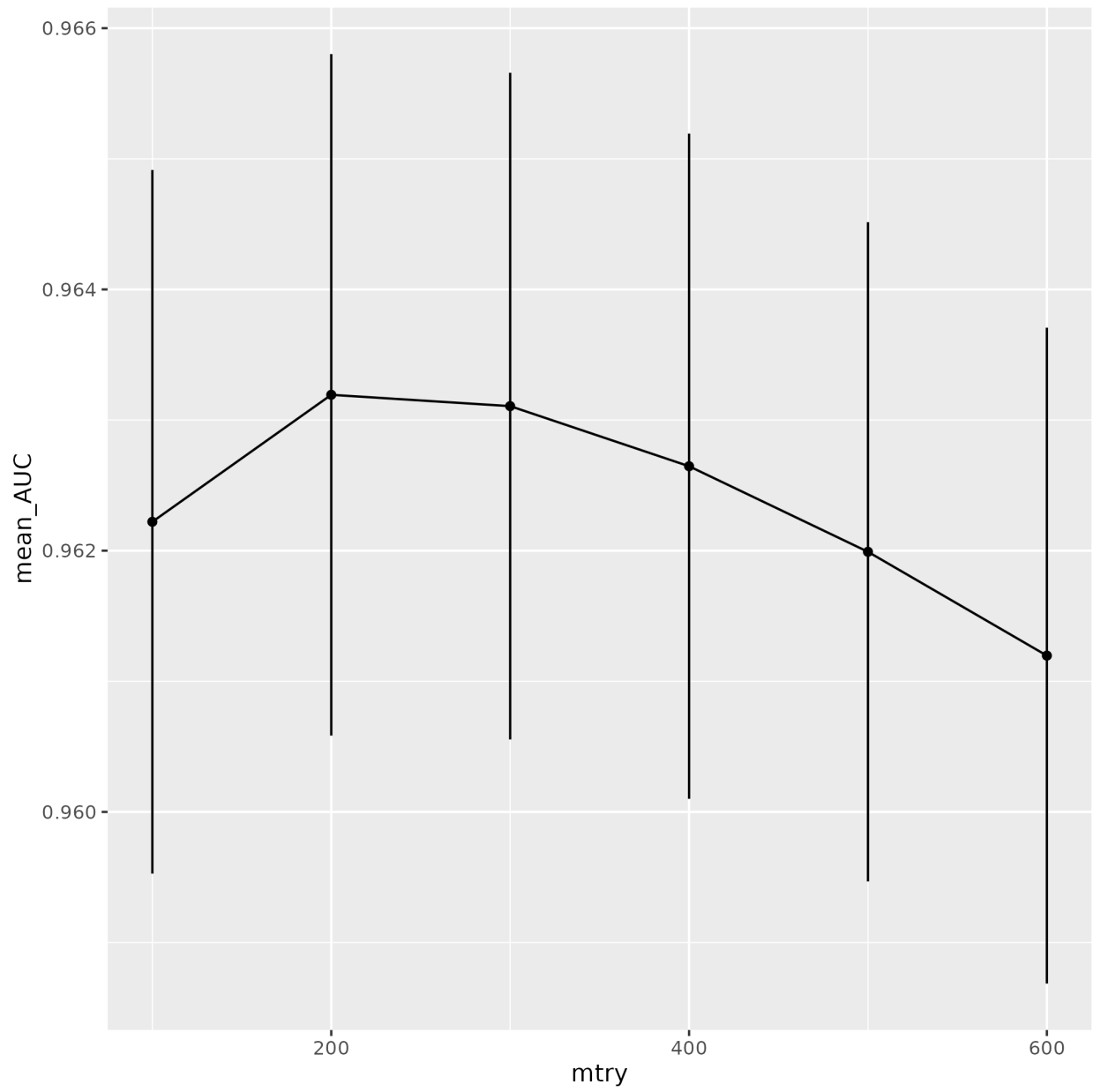


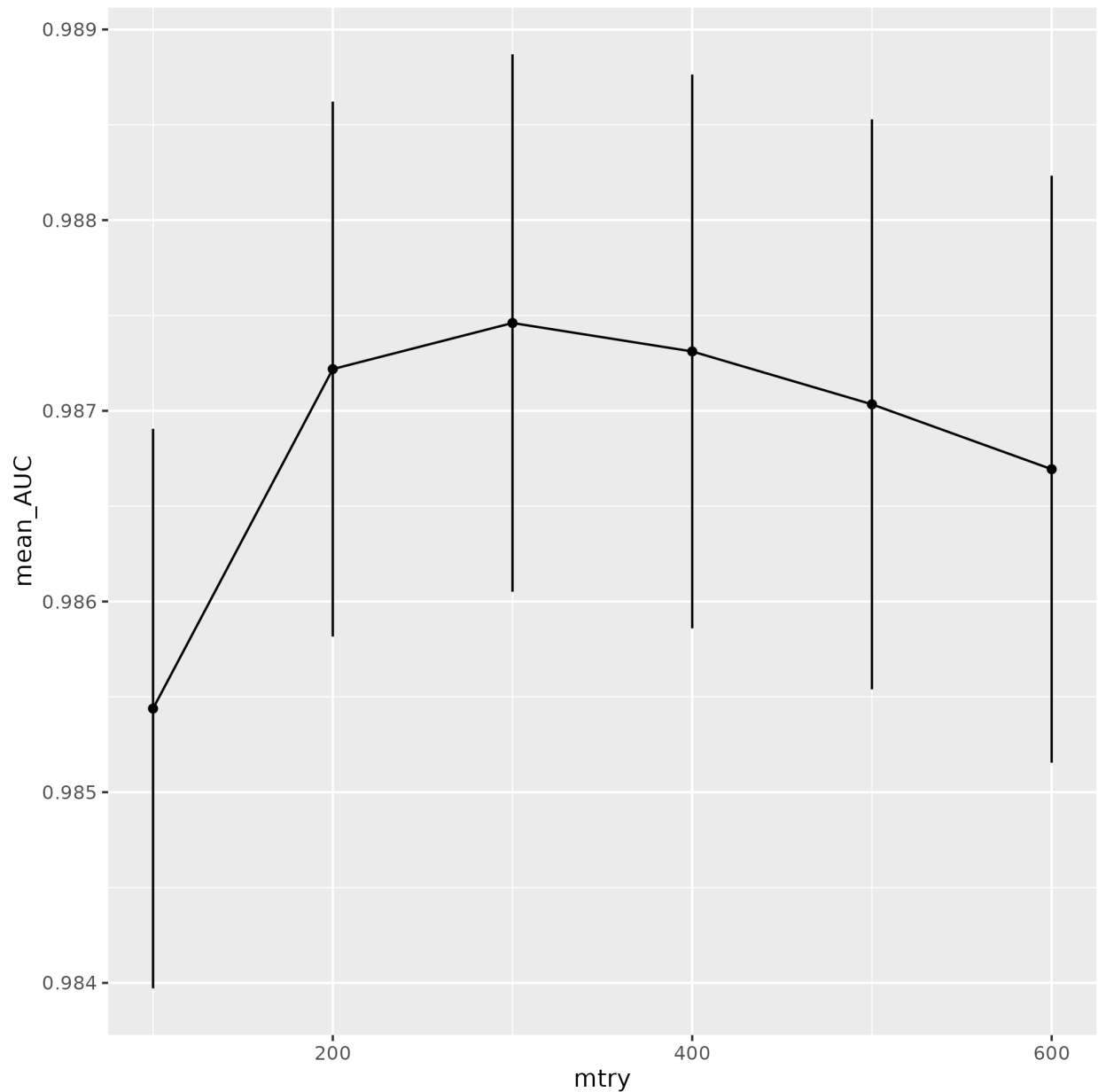
230

AUROC of Training Set by MTRY Value for model data\_availability









Two random forest (rf) models were trained to predict if published scientific papers “contained new  
sequence data” (NSD), and if the paper “had data available” (DA), one model for each variable.  
Other models such as generalized linear regression (GLM) and boosted trees (XGBoost) were  
explored, but were ultimately discarded in favor of the random forest model (data not shown).  
Random forest models were chosen to aid in this classification problem as the creation of many  
decision trees helps to improve accuracy and precision. This type of model has one hyperparameter,  
‘mtry’ or the number of predictors to be sampled at each decision. During iterative model training,

a subset of papers were validated after each completed training and deployment of each model. Papers from each journal, and extras from certain journals were hand-validated against model predictions to generate confusion matrices. Confusion matrices for the final version of each trained model are available in YYYY table(supplement?). To evaluate the fit of the model, we used the Area Under the Receiver Operator Curve (AUROC) which indicates how well the model classifies the data. An AUROC of 0.5 is a random 50/50 guess, and an AUROC of 1.0 indicates that the model always classifies a new item correctly. The NSD model used an mtry value of 200 had an Area Under the Curve(AUC) of 0.9645 and an accuracy of 0.9036. The sensitivity of the NSD model was 0.935, and the specificity of the model was 0.8631. The DA model used an mtry value of 300 had an AUC of 0.9878 and an accuracy of 0.9464. The sensitivity of the DA model was 0.961, and the specificity of the model was 0.9199 (See Table 4XX for more information on trained machine learning models). This shows that the models fit the data well, and can provide classifications on new data with an expected error rate of less than 10%. We deemed this as acceptable, accounting for variability in papers and data, as well as the large size of the dataset on which we deployed the models.

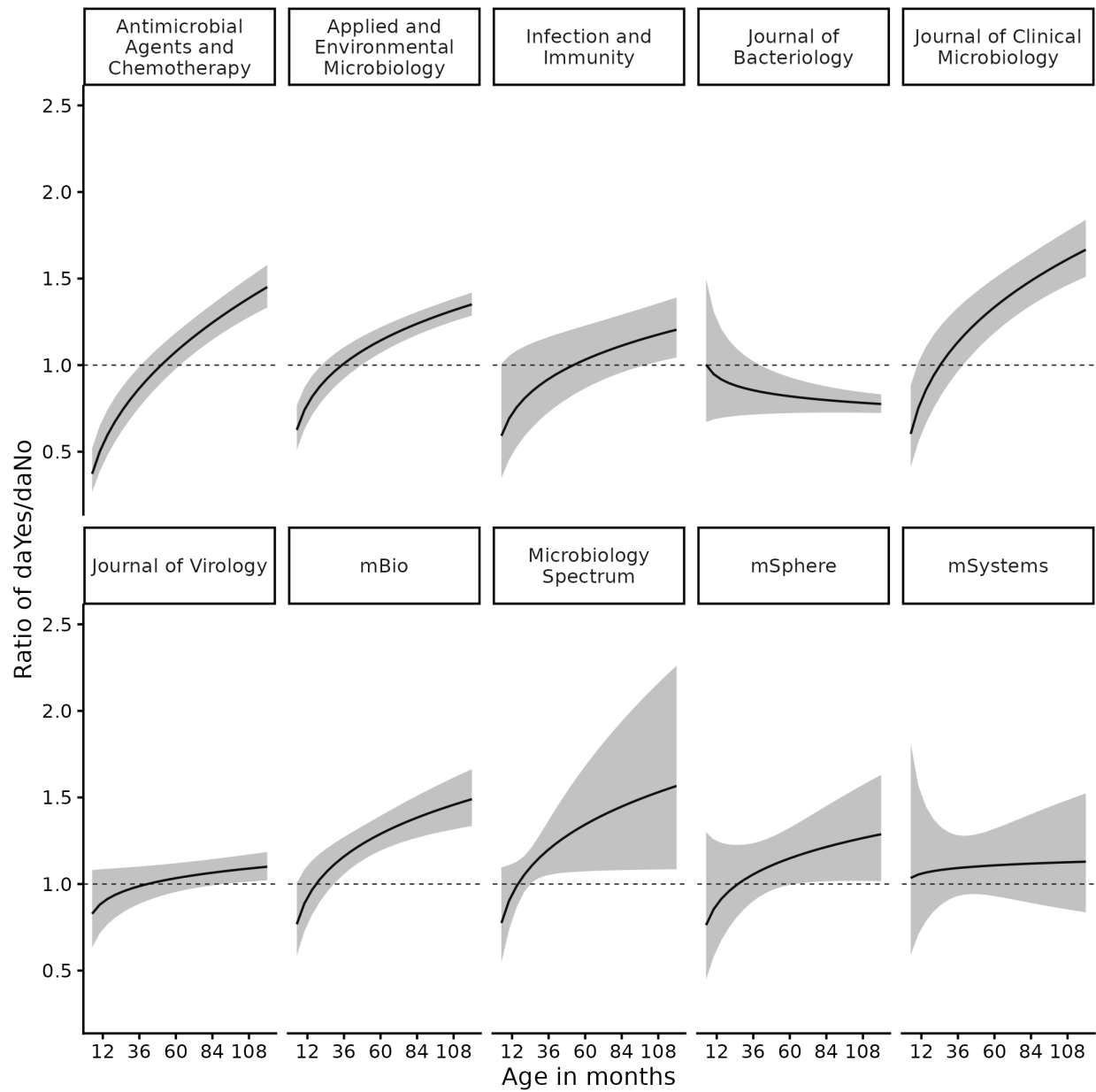
### **Regression Model using Negative Binomial Models**

In this study we sought to investigate the effect of NSD and DA on the number of citations received by a given paper. We focused on NSD papers to determine the effect of having DA. This led us to the use of a negative binomial regression model to best describe our data. All regression data had NSD (NSD == "Yes"). We focused on the continuous outcome of "number of citations" with predictor variables journal (categorical), age in months (continuous), and DA status (dichotomous). Due to the number of citations being bell shaped with a long right tail (very few papers at advanced age with many citations), the model that best described our data was a negative binomial regression model. A negative binomial model is appropriate for data that begins at zero and has a long 'tail' of data, as well as has differing means per group. This model also includes a dispersion parameter, to describe the spread of the data. We applied a log transformation to our age variable (age.in.months) to

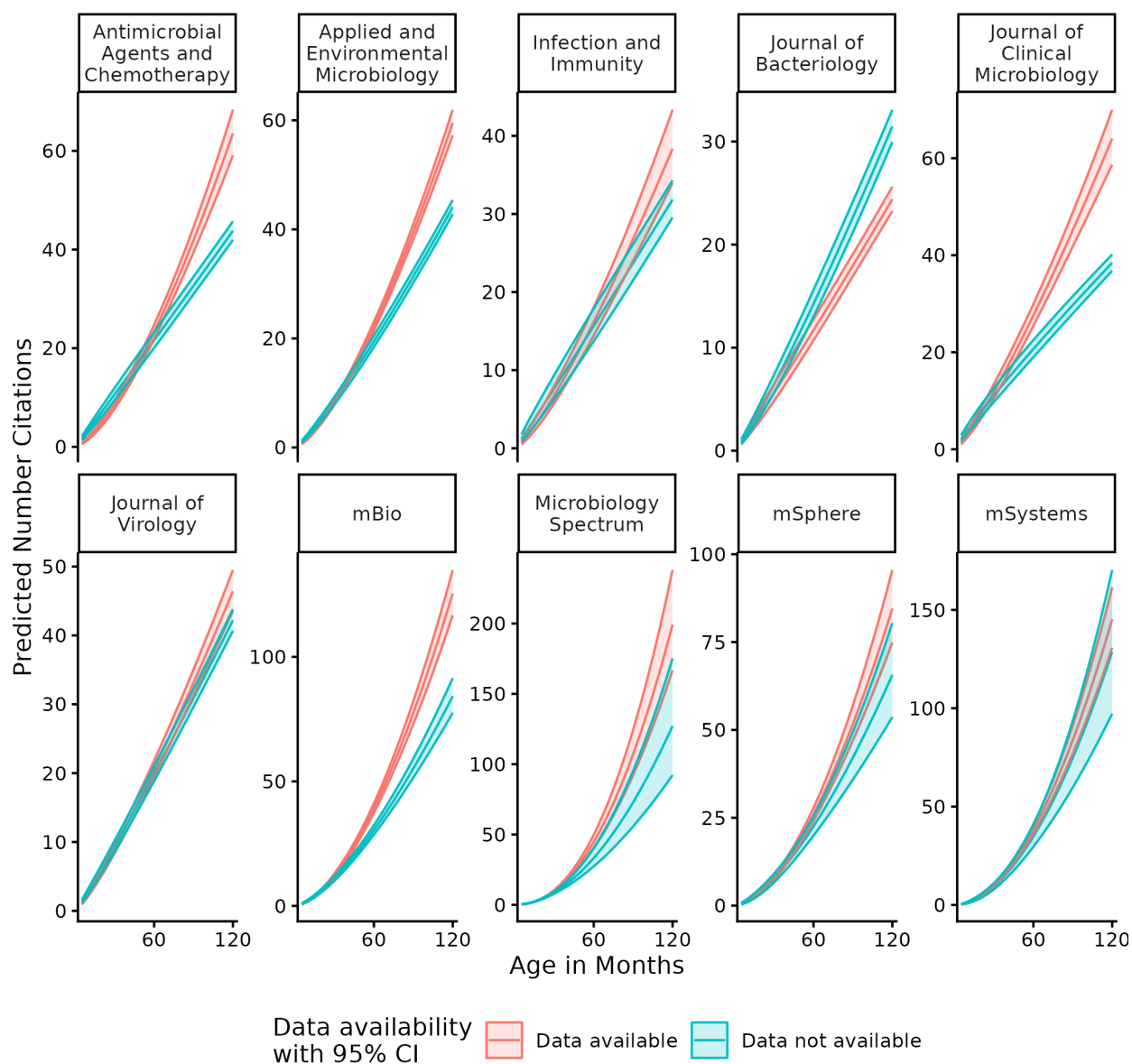
better describe the relationship between time and number of citations received. See supplemental table XXX6 for model coefficients. In general, we found that NSD papers that made DA received more citations over time than those that did not. See figure XXXX for trends in each major journal.

270 ??XX In Figure 7XX, we have calculated the ratio of number of citations for papers of similar age containing data vs those that do not ( $Citations[DA = Yes]/(Citations[DA = No])$ ). Over time, papers with DA receive more citations than those without up to well over 1.5x in some journals (*Journal of Clinical Microbiology*). In all journals excepting the *Journal of Bacteriology*, papers with DA have a greater number of citations at time point 60 months after publication, if not sooner via  
275 ratio plot. Over time this ratio increases further, demonstrating increased citations for papers with DA available over time. Figure 8XX shows that the gap between predicted number of citations for papers with DA vs those without widens over time, even beyond the width of the 95% confidence interval.

## Figures for Negative Binomial



## Predicted Number of Citations from GLM.NB



## Discussion

We investigated the impact of data availability on citation metrics in new sequencing papers by deploying machine learning models on over 150,000 papers from the ASM family of journals.

Overall, making data available increases citation metrics over time, an added benefit to authors for the effort of making their data available.

On average, more than half of papers with NSD had DA (58.8615%), showing that authors of NSD papers are more often than not, making their data publicly available. This is in line with recent NIH policy requiring that data be made available using the Data Management and Sharing Plan (DMS plan) outlined in the NIH's NOT-OD-21-013 ((17)). This NIH policy went into effect in 2023. XXDo we want percent for 2024?XX Expectedly, journals *Genome Announcements* and *Microbiology Resource Announcements* had the highest rates of DA in NSD papers. These journals publish primarily new genomic sequences and are required by ASM to make data available. Journals such as the *Journal of Microbiology and Biology Education* and *Infection and Immunity* which publish fewer NSD papers due to their specific subject matters, have lower rates of NSD and therefore DA in their journals.

Next, we looked further into the impact of DA on citation metrics using a negative binomial regression model. Using this model we found that over time, papers with DA receive more citations than those without up to well over 1.5x the amount of citations in some journals (Fig XXXX. *Journal of Clinical Microbiology*). This effect intensifies over time, with the greatest differences in citations occurring at the 108 months since publication time point. This is great news to manuscript authors, that simply making their DA can provide as much as 50% increase in citations over time. We believe that this more than justifies the work of making data available. We hope that these data will help to incentivize authors to make data available.

We acknowledge the limitations of our study data, that by focusing only on papers published in the ASM family of journals, our results will not be as generalizable. We understand that this relationship between DA and citation metrics may not be as strong in other families of journals, but we hope with the newest NIH funding policies, these trends will continue to improve DA. Another limitation is the availability of paper metadata from various databases, with Crossref having the most complete metadata available for each paper. We were also not able to track citation metrics for individual papers over time. Our database sources only had available citation metrics at the time of dataset download (February 2025), with no intermediate time points for any given paper. This leaves us unable to understand trends over time due to popularity or obsolescence of technique or results.

While making DA in publications has a citation advantage, we hope that is not the only reason  
315 authors choose to make their data available. The need for reproducibility in science and demand  
for large datasets are additional reasons, as well as the NIH funding requirement. We hope that  
authors recognize the need to contribute to the data “commons”, to further work done by others,  
and that even their negative results have value and power to stop others from continuing down  
dead ends.

## 320 **Materials and Methods**

### **Preparation of the Larger Experimental Dataset**

To fully answer our research questions, we created a larger dataset with  $N = 155779$  papers  
curated from reference databases Crossref, NCBI, Scopus, and the Web of Science ((**crossref?**),  
(**ncbi?**), (**scopus?**), (27)). These papers span all twelve ASM journals of interest from January  
325 1st, 2000 to December 31st, 2024. The ASM Journals of interest were *Applied and Environmental*  
*Microbiology*; *Antimicrobial Agents and Chemotherapy*; *Infection and Immunity*; *Journal of Clinical*  
*Biology*; *Journal of Virology*; *Journal of Bacteriology*; *Journal of Microbiology and Biology Education*;  
*Microbiology Resource Announcements* (formerly known as *Genome Announcements*); *mSystems*;  
*mSphere*; *mBio*; and *Microbiology Spectrum*. The data was updated as of February 10th, 2025  
330 with all citation counts frozen at that date.

### **Creation of the Training set**

To train our random forest machine learning model, we first created an appropriate training data  
set. For our initial training set, we chose an initial set of papers from across each journal and the  
time period of interest, adding special emphasis to include papers that were part of our desired  
335 set of interest (i.e. contained published data) to ensure that our two models could adequately  
characterize each paper as a new sequencing paper and if it published raw sequencing data or not.  
After creating our initial dataset, it was necessary to identify the status of both variables by hand



and determine if each paper contained “new sequencing data” (NSD), and if each one had “data available” (DA). This was completed by opening each paper in an internet browser window, and searching for a “data availability” or similar statement. See Table 1XXX for specific cases and how each of these cases were identified for the purpose of this study.

Table 3: Possible Data Scenarios

Scenario	NSD Status	DA Status
Paper is not about generating new sequencing data	No	No
Paper is about generating new sequencing data but has no data available	Yes	No
Paper is about generating new sequencing data and has data available	Yes	Yes
Paper uses sequencing as a confirmation of experimental technique (i.e. confirmation of plasmid insertion)	No	No
Paper discusses new computational or experimental tools	No	No
Paper has microarray data	No	No
Papers using MLST ONLY	No	No
Papers using qPCR ONLY	No	No
Papers about protein sequencing that have nucleotide sequencing	Yes	Yes/No depending on DA
Papers using iRNA techniques	No	No
Papers using pyrosequencing/454 techniques	Yes	Yes/No depending on DA

## Adding Additional Training Set Papers

After initial training of our random forest models, a random sampling of papers was collected for each journal to audit the efficacy of the models. To audit the efficacy of the models, we hand  
345 identified the status of both variables of interest, NSD and DA. We looked for weaknesses in the models, and updated methodology to reflect important areas of interest. For example, in 2023 the ASM journals changed their formatting to include the data availability statement of a paper in a sidebar of the webpage. We identified this by noticing that all papers from journal *Microbiology Resource Announcements* from 2023-2024 were incorrectly characterized by the model as DA =  
350 No. The sidebar of the webpage was not included in the text the model was considering, and code had to be updated to include all sidebar data for all papers. These improvements to the model created a larger and more comprehensive training set of N = 1045. These validations allowed us to create confusion matrices for each model. Confusion matrices for the final version of each trained model are available in XXX table(XXsupplement?).

## 355 Creation of the Training Data from Training dataset

To perform the computational steps required for these experiments, we used the python tool Snakemake ( (28) ), and the University of Michigan's high performance computing cluster (see acknowledgements). Using our selected papers from the training dataset, we downloaded the entirety of each paper's source HTML using the command line tool wget. This allowed us to use the  
360 source HTML multiple times for updated analyses without the need to re-query the ASM web servers numerous times. Next, we performed cleaning of the HTML using R packages rvest ((29)), textstem( (30)), and xml2 ((31)) to get the desired portions of the paper from the HTML including the abstract, the body of paper, all tables and figures with captions, as well as the side panels for all papers, but especially those containing the data availability statements in papers published after the 2023  
365 change in webpage format (see above). Then we removed unnecessary text using R packages tm(text manipulation)((32), (33)) and textstem ((30)), as well as converting all text to lowercase,

and the removal of digits and non-alphabetic characters such as whitespace. To have the fewest number of unique words, we lemmatized (sort words by grouping inflected or variant forms of the same word) words to trace them back to their root words and eliminate any possible issues with word tense. After this, we created and counted our ‘tokens’, phrases of up to 1-3 consecutive words from the text of the paper using R package `tokenizers` ((34)). Towards the goal of the fewest meaningful number of words, we used the ‘Snowball’ ((35)) dictionary of ‘stop words’ to remove non-meaningful words such as articles ‘a’, ‘an’, and ‘the’. We removed the ‘space’ character with an underscore in multi-word tokens for ease of processing, and created a count table for the tokens in each paper.

Once the tokens in each paper were counted, we transformed the data into a sparse matrix format useable by the R package `mikropml` ((36)), using R packages `caret` and `dplyr` ((37), (38), (39)). Tokens were filtered to those which appear in greater than one paper. This allows comparison between papers by the model. We removed near zero variants (tokens with frequency very close to zero) as well as collapsing perfectly correlated tokens (tokens that always appear together) using R packages `caret` and `mikropml` to reduce model complexity. The data was then simplified to keep only the following variables; tokens, frequency, journal information, and hand identified NSD and DA variables. This simplified sparse matrix data had the mean and standard deviation calculated and saved for the frequency of each token to later apply a z-scoring method to future data to be predicted by the model.

## Training of the DA and NSD Models

We trained two random forest machine learning models using `mikropml`’s “`run_ml`” function, one to determine if a paper contained new sequence data (NSD), and another to determine if the paper had data available (DA). The `mikropml` “`run_ml`” function uses methodology described by \*Topcuoglu et al.\*((40)) to split data for model training. Random forest models have one hyperparameter to tune, the `mtry` value. We began with `mtry` values of 100, 200, 300, 400, 500, and 600, to find peak hyperparameter performance given *N tokens*. We trained the models multiple times in accordance

with existing methodologies, first to find the optimal Area Under the Receiver-Operator Curve (AUROC) value for each model with N=100 seeds. Then to find the best mtry performance for each model, with N=1 seed. Finally, with N=1 seed to train each final model for use on experimental data.

## Deploying the RF Models

Once our RF models were ready we applied the same steps to ready papers for application of machine learning models as the model training dataset. See above for descriptions of web-scraping HTML, cleaning HTML, removing unnecessary text, and creation of token count table for application in each of the machine learning models to determine the NSD and DA statuses for each paper. Once the frequency count tables were prepared for each paper, a z-score was applied using the saved data from each model appropriately, using the formula  $XX \frac{((Observedtokenfrequency - Modeltokenfrequency))}{(modeltokenfrequencysd)}$ . This z-scoring formula was applied to standardize the frequency of each token. Only tokens included in the machine learning models were retained in experimental datasets. Finally, each model was deployed on each paper to determine its NSD and DA status.

## Statistical Methodology

After each random forest model was deployed on our experimental dataset, we used a negative binomial regression model using the R package MASS ((MASSref?)). A negative binomial regression model allows us to investigate data where group means are different than the overall dataset mean. A log transformation of the time variable(age.in.months) was applied prior to estimating the statistical regression model to correct for the nature of time as compared to other variables in the dataset. After applying the negative binomial model, we calculated ratios of the estimated number of citations per paper over time by DA status using R package emmeans ((41)). We also used R package sjPlot to estimate the 95% CI for estimated citations over time for each journal by DA status ((sjPlotref?)).

- Supplemental Material file list (where applicable)
- Acknowledgments

– *“This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor.”*

ARC’s RRID is: [SCR\\_027337](#)

- References
- Figures/Tables/stats to make/get

– table of conditions to add to the methods of classification?

1. [Federal Research and Development \(R&D\) Funding: FY2024](#). legislation.
2. **Moniz P, Druckman JN, Freese J.** 2025. The file drawer problem in social science survey experiments. *Proceedings of the National Academy of Sciences* **122**:e2426937122. doi:[10.1073/pnas.2426937122](#).
3. **Li Y, Chen L.** 2014. Big Biological Data: Challenges and Opportunities. *Genomics, Proteomics & Bioinformatics* **12**:187–189. doi:[10.1016/j.gpb.2014.10.001](#).
4. **Pal S, Mondal S, Das G, Khatua S, Ghosh Z.** 2020. Big data in biology: The hope and present-day challenges in it. *Gene Reports* **21**:100869. doi:[10.1016/j.genrep.2020.100869](#).
5. [Expired NOT-OD-24-096: Notice of special interest \(NOSI\): Promoting data reuse for health research](#).
6. [PAR-23-236: Early-stage biomedical data repositories and knowledgebases \(R24 clinical trial not allowed\)](#).

7. **Maxson Jones K, Ankeny RA, Cook-Deegan R.** 2018. The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. *Journal of the History of Biology* **51**:693–805. doi:[10.1007/s10739-018-9538-7](https://doi.org/10.1007/s10739-018-9538-7).
8. **Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, Zhou W, Buck GA, Snyder MP, Strauss JF, Weinstock GM, White O, Huttenhower C, The Integrative HMP (iHMP) Research Network Consortium.** 2019. The Integrative Human Microbiome Project. *Nature* **569**:641–648. doi:[10.1038/s41586-019-1238-8](https://doi.org/10.1038/s41586-019-1238-8).
9. **Gevers D, Pop M, Schloss PD, Huttenhower C.** 2012. Bioinformatics for the Human Microbiome Project. *PLOS Computational Biology* **8**:e1002779. doi:[10.1371/journal.pcbi.1002779](https://doi.org/10.1371/journal.pcbi.1002779).
- 435 10. **Group JCHMPDGW.** 2012. Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. *PLOS ONE* **7**:e39315. doi:[10.1371/journal.pone.0039315](https://doi.org/10.1371/journal.pone.0039315).
11. [International nucleotide sequence database collaboration.](#)
12. [About FORCE11 – FORCE11.](#)

13. **Yilmaz P, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PSG, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, Gil IS, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methé BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spor A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, Yatsunenko T, Glöckner FO.** 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MlxS) specifications. *Nature Biotechnology* **29**:415–420. doi:[10.1038/nbt.1823](https://doi.org/10.1038/nbt.1823).
14. **Altman Director of Research and Head/Scientist, Micah, Borgman Professor and Presidential Chair, Christine, Crosas Director of Data Science M, Matone Co-Director M.** 2015. An introduction to the joint principles for data citation. *Bulletin of the Association for Information Science and Technology* **41**:43–45. doi:[10.1002/bult.2015.1720410313](https://doi.org/10.1002/bult.2015.1720410313).

15. **Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, Silva Santos LB da, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen PAC 't, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, Schaik R van, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, Lei J van der, Mulligen E van, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B.** 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**:160018. doi:[10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
  
16. **Mirzayi C, Renson A, Zohra F, Elsafoury S, Geistlinger L, Kasselmann LJ, Eckenrode K, Wijgert J van de, Loughman A, Marques FZ, MacIntyre DA, Arumugam M, Azhar R, Beghini F, Bergstrom K, Bhatt A, Bisanz JE, Braun J, Bravo HC, Buck GA, Bushman F, Casero D, Clarke G, Collado MC, Cotter PD, Cryan JF, Demmer RT, Devkota S, Elinav E, Escobar JS, Fettweis J, Finn RD, Fodor AA, Forslund S, Franke A, Furlanello C, Gilbert J, Grice E, Haibe-Kains B, Handley S, Herd P, Holmes S, Jacobs JP, Karstens L, Knight R, Knights D, Koren O, Kwon DS, Langille M, Lindsay B, McGovern D, McHardy AC, McWeeney S, Mueller NT, Nezi L, Olm M, Palm N, Pasolli E, Raes J, Redinbo MR, Rühlemann M, Balfour Sartor R, Schloss PD, Schriml L, Segal E, Shardell M, Sharpton T, Smirnova E, Sokol H, Sonnenburg JL, Srinivasan S, Thingholm LB, Turnbaugh PJ, Upadhyay V, Walls RL, Wilmes P, Yamada T, Zeller G, Zhang M, Zhao N, Zhang M, Zhao L, Zhao L, Bao W, Culhane A, Devanarayan V, Dopazo J, Fan X, Fischer M, Jones W, Kusko R, Mason CE, Mercer TR, Sansone S-A, Scherer A, Shi L, Thakkar S, Tong W, Wolfinger R, Hunter C, Segata N, Huttenhower C, Dowd JB, Jones HE, Waldron L.** 2021. Reporting guidelines for human microbiome research: The STORMS checklist. *Nature Medicine* **27**:1885–1892. doi:[10.1038/s41591-021-01552-x](https://doi.org/10.1038/s41591-021-01552-x).



17. [NOT-OD-21-013: Final NIH Policy for Data Management and Sharing](#).
18. **Langille MGI, Ravel J, Fricke WF**. 2018. “Available upon request”: Not good enough for microbiome data! *Microbiome* **6**:8. doi:[10.1186/s40168-017-0394-z](#).
19. [Open Data Policy](#). ASM Journals.
- 445 20. [Publishing Ethics Policies and Procedures](#). ASM Journals.
21. [Human Microbiome Project \(HMP\) | NIH Common Fund](#).
22. **Ding T, Schloss PD**. 2014. Dynamics and associations of microbial community types across the human body. *Nature* **509**:357–360. doi:[10.1038/nature13178](#).
23. **Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methé B, Schloss PD, Gevers D, Mitreva M, Huttenhower C**. 2012. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. *PLOS Computational Biology* **8**:e1002358. doi:[10.1371/journal.pcbi.1002358](#).

24. **Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Paul Brooks J, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PSG, Chen I-MA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Michael Dunne W, Scott Durkin A, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K, Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney LJ, Foster L, Di Francesco V, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez A, Griggs A, Gujja S, Kinder Haake S, Haas BJ, Hamilton HA, Harris EL, Hepburn TA, Herter B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM, Izard J, Jansson JK, Jiang H, Jordan C, Joshi V, Katancik JA, Keitel WA, Kelley ST, Kells C, King NB, Knights D, Kong HH, Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL, Lemon KP, Lennon N, Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu Y, Lo C-C, Lozupone CA, Dwayne Lunsford R, Madden T, Mahurkar AA, Mannon PJ, Mardis ER, Markowitz VM, Mavromatis K, McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ, Newsham I, Nusbaum C, O’Laughlin M, Orvis J, Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Pop M, Priest ME, Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera MC, Rodriguez-Mueller B, Rogers Y-H, Ross MC, Russ C, Sanka RK, Sankar P, Fah Sathirapongsasuti J, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, Sheth NU, Simone GA, Singh I, Smillie<sup>24</sup>CS, Sobel JD, Sommer DD, Spicer P, Sutton GG, Sykes SM, Tabbaa DG, Thiagarajan M, Tomlinson CM, Torralba M, Treangen TJ,**

- 450 25. **Maxmen A.** 2021. One million coronavirus sequences: popular genome site hits mega milestone. *Nature* **593**:21–21. doi:[10.1038/d41586-021-01069-w](https://doi.org/10.1038/d41586-021-01069-w).
26. **Cuomo CA.** 2021. The Relaunch of Microbiology Spectrum. *Microbiology Spectrum* **9**:10.1128/spectrum.00396–21. doi:[10.1128/spectrum.00396-21](https://doi.org/10.1128/spectrum.00396-21).
27. Accessed 2025. [Web of science](#). Internet Database, Clarivate.
28. **Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J.** Sustainable data analysis with Snakemake. doi:[10.12688/f1000research.29032.2](https://doi.org/10.12688/f1000research.29032.2).
29. **Wickham H.** 2025. [Rvest: Easily harvest \(scrape\) web pages](#).
- 455 30. **Rinker TW.** 2018. [textstem: Tools for stemming and lemmatizing text](#). Buffalo, New York.
31. **Wickham H, Hester J, Ooms J.** 2025. [xml2: Parse XML](#).
32. **Feinerer I, Hornik K, Meyer D.** 2008. Text mining infrastructure in r. *Journal of Statistical Software* **25**:1–54. doi:[10.18637/jss.v025.i05](https://doi.org/10.18637/jss.v025.i05).
33. **Feinerer I, Hornik K.** 2025. [Tm: Text mining package](#).
34. **Mullen LA, Benoit K, Keyes O, Selivanov D, Arnold J.** 2018. Fast, consistent tokenization of natural language text. *Journal of Open Source Software* **3**:655. doi:[10.21105/joss.00655](https://doi.org/10.21105/joss.00655).
- 460 35. **Benoit K, Muhr D, Watanabe K.** 2021. [Stopwords: Multilingual stopword lists](#).

36. **Topçuoğlu BD, Lapp Z, Sovacool KL, Snitkin E, Wiens J, Schloss PD.** 2021. mikropml: User-friendly r package for supervised machine learning pipelines. Journal of Open Source Software **6**:3073. doi:[10.21105/joss.03073](https://doi.org/10.21105/joss.03073).
37. **Kuhn, Max.** 2008. Building predictive models in r using the caret package. Journal of Statistical Software **28**:1–26. doi:[10.18637/jss.v028.i05](https://doi.org/10.18637/jss.v028.i05).
38. **Wickham H, François R, Henry L, Müller K, Vaughan D.** 2025. [Dplyr: A grammar of data manipulation](#).
39. **Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Golemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H.** 2019. Welcome to the tidyverse. Journal of Open Source Software **4**:1686. doi:[10.21105/joss.01686](https://doi.org/10.21105/joss.01686).
- 465 40. **Topçuoğlu BD, Lapp Z, Sovacool KL, Snitkin E, Wiens J, Schloss PD.** 2021. Mikropml: User-Friendly R Package for Supervised Machine Learning Pipelines. Journal of Open Source Software **6**:3073. doi:[10.21105/joss.03073](https://doi.org/10.21105/joss.03073).
41. **Lenth RV.** 2025. [Emmeans: Estimated marginal means, aka least-squares means](#).