Summary Stats for DA Project

2025-08-20

```
#library statements
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
#import data
data <- read_csv("~/Documents/Schloss/Colovas_Data_Accessibility/Data/final/predictions_with_metadata.cs
```

```
## New names:
## * `NA...9` -> `NA...55`
## * `NA...10` -> `NA...56`
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 154720 Columns: 77
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (32): file, da, nsd, paper.x, doi, doi_no_underscore, journal_abrev, co...
## dbl  (12): issue, member, prefix, score, reference.count, references.count, ...
## lgl  (28): assertion, author, link, license, reference, update_to, subtitle,...
## date  (5): created, deposited, indexed, published.online, pub_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

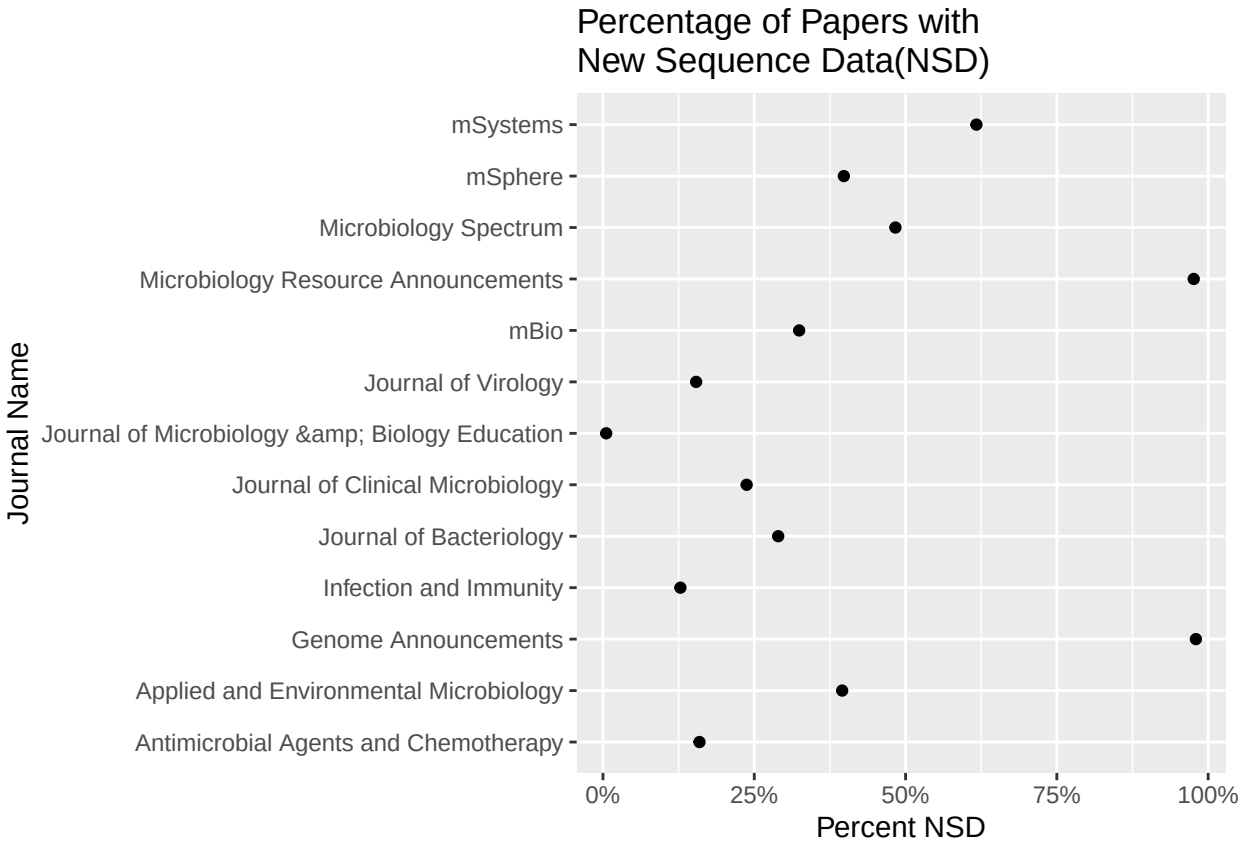
How many papers are nsd papers by journal?

```r
nsd_yes <- data %>%
  count(container.title,
        nsd) %>%
  filter(!is.na(nsd)) %>%
  group_by(container.title) %>%
  mutate(nsd,
            total = sum(`n`),
            nsd_fract = `n`/total) %>%
  filter(nsd == "Yes")

  knitr::kable(nsd_yes)
```

container.title	nsd	n	total	nsd_fract
Antimicrobial Agents and Chemotherapy	Yes	3237	20297	0.1594817
Applied and Environmental Microbiology	Yes	8638	21853	0.3952775
Genome Announcements	Yes	6578	6714	0.9797438
Infection and Immunity	Yes	1854	14490	0.1279503
Journal of Bacteriology	Yes	4867	16806	0.2895990
Journal of Clinical Microbiology	Yes	4374	18422	0.2374335
Journal of Microbiology & Biology Education	Yes	7	1305	0.0053640
Journal of Virology	Yes	4583	29761	0.1539935
Microbiology Resource Announcements	Yes	5738	5878	0.9761824
Microbiology Spectrum	Yes	2957	6119	0.4832489
mBio	Yes	2498	7705	0.3242051
mSphere	Yes	1041	2615	0.3980880
mSystems	Yes	1436	2327	0.6171036

```r
  ggplot(data = nsd_yes, aes(y = container.title, x = nsd_fract)) +
    geom_point() +
    scale_x_continuous(labels = scales::label_percent()) +
    labs(title = "Percentage of Papers with\nNew Sequence Data(NSD)",
         y = "Journal Name",
         x = "Percent NSD")
```



How many papers are nsd papers by journal?

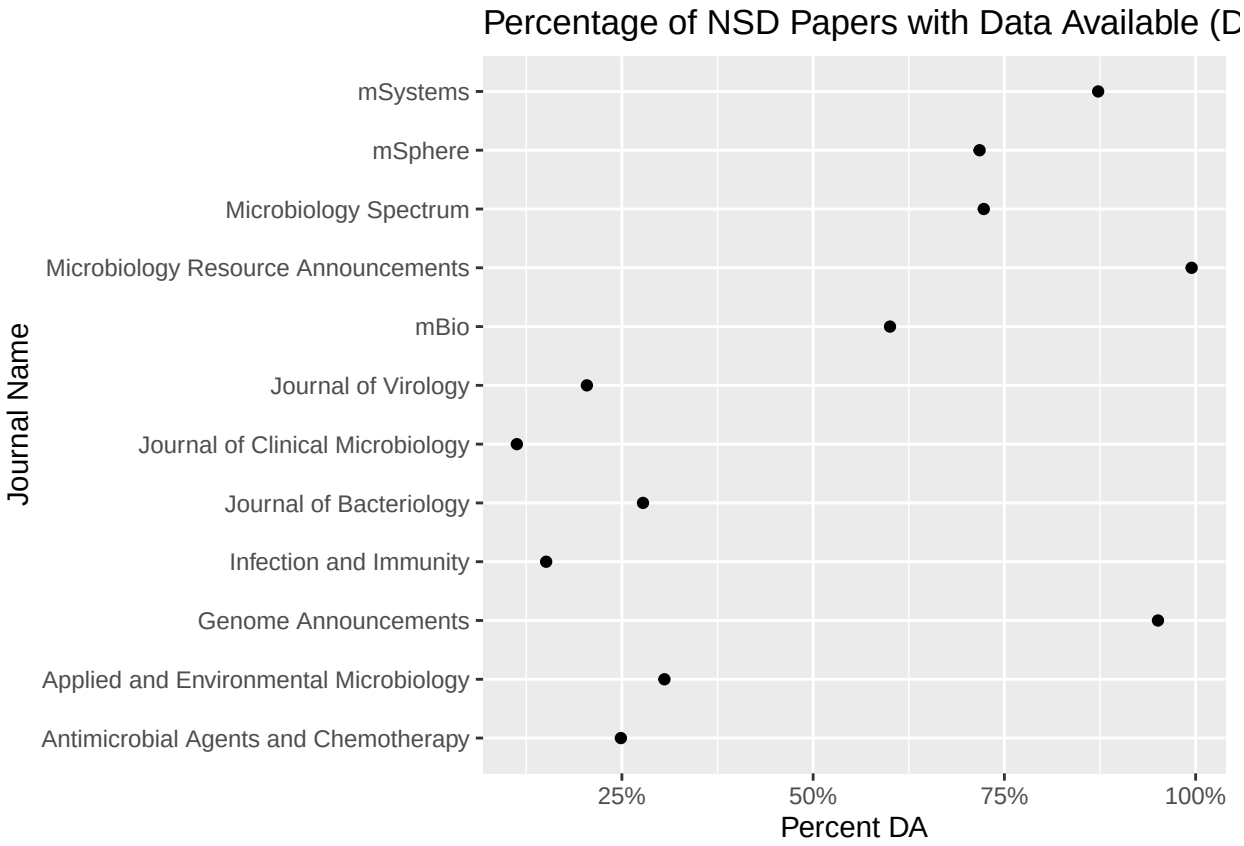
```
da_yes <- data %>%
  filter(nsd == "Yes") %>%
  count(container.title,
        da) %>%
  filter(!is.na(da)) %>%
  group_by(container.title) %>%
  mutate(da,
            total = sum(`n`),
            da_fract = `n`/total) %>%
  filter(da == "Yes")

  knitr::kable(da_yes)
```

container.title	da	n	total	da_fract
Antimicrobial Agents and Chemotherapy	Yes	805	3237	0.2486871
Applied and Environmental Microbiology	Yes	2640	8638	0.3056263
Genome Announcements	Yes	6254	6578	0.9507449
Infection and Immunity	Yes	280	1854	0.1510248
Journal of Bacteriology	Yes	1351	4867	0.2775837
Journal of Clinical Microbiology	Yes	493	4374	0.1127115

container.title	da	n	total	da_fract
Journal of Virology	Yes	936	4583	0.2042330
Microbiology Resource Announcements	Yes	5708	5738	0.9947717
Microbiology Spectrum	Yes	2138	2957	0.7230301
mBio	Yes	1500	2498	0.6004804
mSphere	Yes	747	1041	0.7175793
mSystems	Yes	1253	1436	0.8725627

```
ggplot(data = da_yes, aes(y = container.title, x = da_fract)) +
  geom_point() +
  scale_x_continuous(labels = scales::label_percent()) +
  labs(title = "Percentage of NSD Papers with Data Available (DA)",
       y = "Journal Name",
       x = "Percent DA")
```




```
#i think this needs to be of nsd yes papers
```