|  |
| --- |
| # load packages  library(rbiorxiv)  library(tidyverse)  library(gridExtra)    # make directory for output if it doesn't exist  if (dir.exists("./output")==FALSE) dir.create("./output")  if (dir.exists("./output/plots")==FALSE) dir.create("./output/plots") |

Use the R client to get a data frame of preprints uploaded in 2020.

|  |  |
| --- | --- |
| 1 | data <- biorxiv\_content(from = "2020-01-01", to = "2020-03-29", limit = "\*", format = "df") |

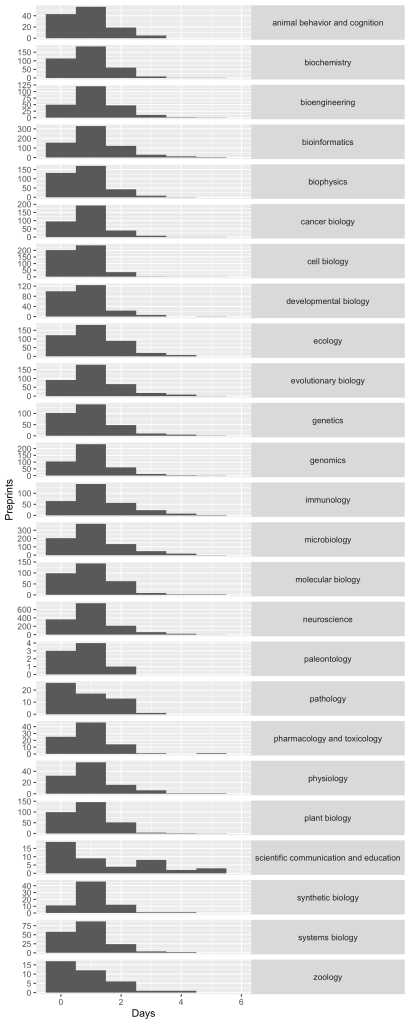
We only want to look at new preprints (Version 1) and not revisions, so let’s filter for that. Then, we’ll take advantage of bioRxiv’s new style DOIs to find the “submission date”.

|  |  |
| --- | --- |
| 1  2  3  4  5 | data <- filter(data, version == 1)  data$doi\_date <- substr(data$doi, 9, 18)  data$doi\_date <- gsub("\\.", "-", data$doi\_date)  data$days <- as.Date(data$date) - as.Date(data$doi\_date)  data$category <- as.factor(data$category) |

We now have a column called ‘days’ that shows the time in days from “submission” to “publication”. We will use this as a measure of screening time. Note: this is imperfect because the submission date is when an author begins uploading their preprint (they could take several days to do this) and not when it actually gets submitted to bioRxiv.

Let’s look at the screening time per subject area.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8 | p1 <- ggplot(data, aes(x = as.numeric(days))) +    geom\_histogram(binwidth = 1) +    xlim(NA, 6) +    facet\_grid(category ~ ., scales = "free") +    labs(x = "Days",         y= "Preprints") +    theme(strip.text.y = element\_text(angle = 0))  ggsave("./output/plots/screenLag.png", p1, height = 15, width = 6, dpi = 300,) |

Histogram of screening times per subject

I was surprised to see that, with the exception of “Scientific Communication and Education”, **the screening times were pretty constant across categories.**

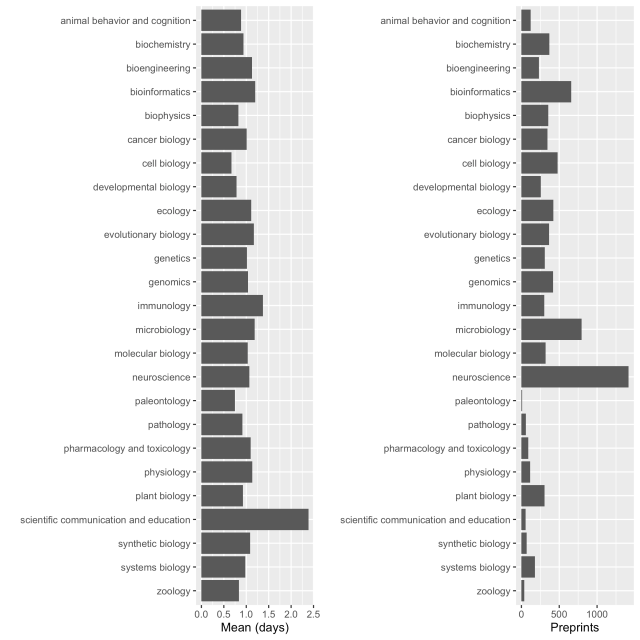
The subject areas on bioRxiv are not equal in size. Look at the numbers on the axes for Zoology and for Neuroscience to get a feel for the difference. The histogram view conceals these differences.

Next, we can calculate the average screening time and see if the busiest categories suffer delayed screening.

|  |  |
| --- | --- |
| 1  2  3  4  5 | df1 <- aggregate(as.numeric(data$days), list(data$category), mean)  colnames(df1) <- c("category","mean\_days")  df2 <- count(data$category)  colnames(df2) <- c("category","count")  summary\_df <- merge(df1,df2) |

And then make some bar charts to look at the data.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12 | p3 <- ggplot(summary\_df, aes(x = category, y = mean\_days)) +    geom\_bar(stat = "identity") +    scale\_x\_discrete(limits = rev(levels(summary\_df$category))) +    labs(x = "", y = "Mean (days)") +    coord\_flip()  p4 <- ggplot(summary\_df, aes(x = category, y = count)) +    geom\_bar(stat = "identity") +    scale\_x\_discrete(limits = rev(levels(summary\_df$category))) +    labs(x = "", y = "Preprints") +    coord\_flip()  p5 <- grid.arrange(p3,p4, nrow = 1, ncol = 2)  ggsave("./output/plots/summary.png", p5, height = 8, width = 8, dpi = 300,) |



The **average screening time is 1 day or less**. Neuroscience, microbiology and bioinformatics (the biggest categories) have similar screening delays to less busy categories. So, assuming that Affiliates screen on the basis of expertise, the pool is either enriched for these popular areas, or those affiliates are more busy!

The longest lag is for “Scientific Communication and Education”, which is a very small category. Assuming the authors take a similar time to upload these manuscripts, I guess the Affiliates tend to screen these preprints as a lower priority. These papers do tend to be a bit different from other research papers and have separate screening criteria. Anyway, they still get screened in just over 2 days, which is still impressive.

I was pleased to see “Cell Biology” had the shortest screening time (around half a day)!

**Conclusion**

Even though my impression was that Bioinformatics and Neuroscience papers linger in the queue, this is not actually the case. There’s likely more of them in the queue because there are more of them, period.

The bioRxiv team have done a great job in maintaining a pool of Affiliates that can screen the huge number of preprints that are uploaded.