Looking at the number of bacterial human pathogens across taxa.

Daniel Padfield

04/07/2022

Dataset of bacterial human pathogens

We present a comprehensive list of known bacterial pathogens known to cause infectious symptoms in humans. A summary of all the datasets used and produced are summarised in this document.

The dataset is available on GitHub, but to allow the code to fit cleanly in these walkthroughs we created a shortened URL for the file (https://shorturl.at/hiwy7)

This work-through reads in the final list and then reproduces Figure 2, Figure 3, and Table 1 from the manuscript.

Load in packages and data

First we will load in the R packages used in the script.

```
# load packages
library(tidyverse)
library(lubridate)
library(patchwork)
library(gt)
library(rio)
```

Next we will load in full dataset.

```
# set url
url <- "https://shorturl.at/hiwy7"

d_pathogens <- rio::import(url, sheet = "Full List",
    range = "A1:J1514") %>%
    janitor::clean_names()

select(d_pathogens, -reference) %>%
    head()
```

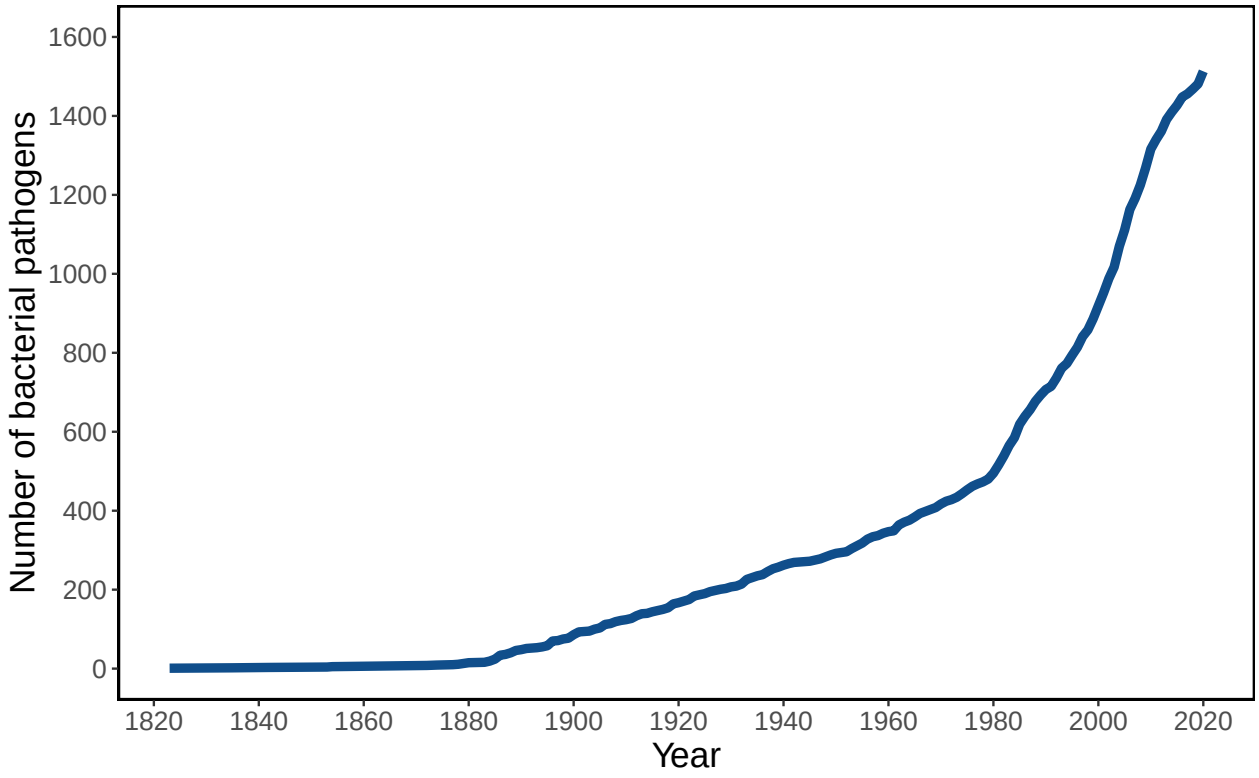
```
##   superkingdom         phylum               class            order
## 1     Bacteria Proteobacteria Gammaproteobacteria Enterobacterales
## 2     Bacteria     Firmicutes             Bacilli       Bacillales
## 3     Bacteria     Firmicutes          Clostridia    Eubacteriales
## 4     Bacteria   Fusobacteria       Fusobacteriia  Fusobacteriales
## 5     Bacteria Proteobacteria Gammaproteobacteria      Vibrionales
## 6     Bacteria     Firmicutes             Bacilli       Bacillales
##             family        genus    species year      status
## 1     Yersiniaceae     Serratia marcescens 1823 established
## 2      Bacillaceae     Bacillus   subtilis 1835 established
## 3   Clostridiaceae  Clostridium ventriculi 1842 established
## 4 Leptotrichiaceae Leptotrichia   buccalis 1853 established
```

```
## 5     Vibrionaceae       Vibrio   cholerae 1854 established
## 6      Bacillaceae     Bacillus  anthracis 1872 established
```

Plot discovery curve

We can see the rate at which new bacterial human pathogens were described by making a discovery curve. This can be done easily by calculated the number of pathogens discovered each year and taking the cumulative sum across years

```
# create cumulative counts for each year
species_counts <- group_by(d_pathogens, year) %>%
    tally() %>%
    arrange(year) %>%
    mutate(n_cum = cumsum(n))

# make the plot discovery curve
ggplot(species_counts, aes(x = year, y = n_cum)) +
    geom_line(size = 2, colour = "dodgerblue4") + theme_bw() +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.border = element_rect(colour = "black",
            fill = NA, size = 1), axis.title.y = element_text(size = 16),
        axis.title.x = element_text(size = 16), strip.background = element_blank(),
        strip.text = element_text(angle = 0, hjust = 0,
            size = 12), axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12), plot.title = element_text(size = 16),
        legend.position = "none") + labs(y = "Number of bacterial pathogens",
    x = "Year") + scale_x_continuous(breaks = round(seq(min(1820),
    max(2020), by = 20), 1)) + scale_y_continuous(breaks = round(seq(min(0),
    max(1600), by = 200), 1), limits = c(0, 1600))
```

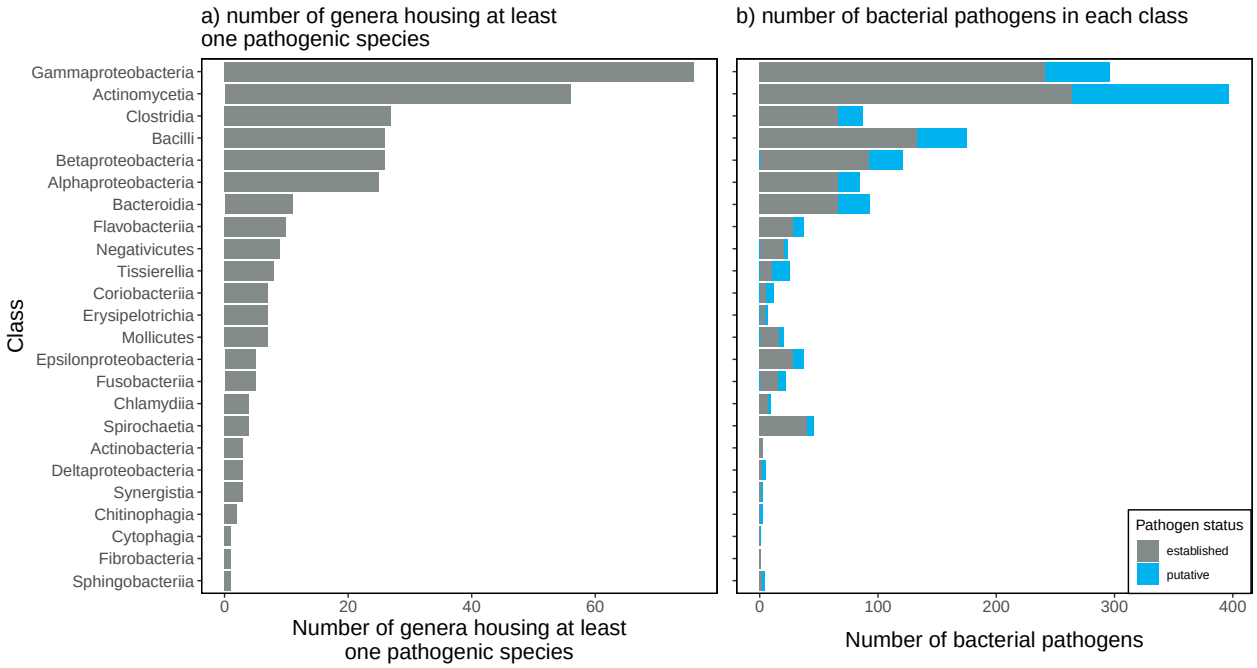


Visualise distribution of pathogens across different Classes

Bacterial human pathogens are unlikely to be distributed completely evenly across the bacterial taxonomy. We looked to see which bacterial Classes have (a) more unique genera that contain a human pathogen and (b) the total number of pathogenic species to see how pathogens are distributed across different taxa.

```
# just keep the columns we are interested in
d_pathogens2 <- select(d_pathogens, class, genus, status)

# first just find number of unique genera
# irrespective of status
genera <- select(d_pathogens2, -status) %>%
    distinct() %>%
    group_by(class) %>%
    tally()

# calculate number of species in each genera for
# each status
species <- d_pathogens2 %>%
    group_by(class, status) %>%
    tally() %>%
    spread(., status, n) %>%
    mutate(across(everything(), replace_na, 0), species_total = established +
        putative)

# create order for the bars in the plots (most
# speciose groups first)
to_order <- arrange(genera, desc(n)) %>%
    pull(class)

## number of genera per class
fig3a <- ggplot() + geom_bar(data = genera, aes(y = n,
    x = forcats::fct_relevel(class, to_order)), stat = "identity",
    fill = "azure4") + theme_bw() + theme(panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), panel.border = element_rect(colour = "black",
        fill = NA, size = 1), axis.title.y = element_text(size = 16),
    axis.title.x = element_text(size = 16), strip.background = element_blank(),
    strip.text = element_text(angle = 0, hjust = 0,
        size = 12), axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12), plot.title = element_text(size = 16),
    legend.position = c(0.875, 0.0825), legend.background = element_rect(fill = "white",
        color = "black")) + labs(y = "Number of genera housing at least\none pathogenic species",
    x = "Class", fill = "Pathogen status") + coord_flip() +
    scale_x_discrete(limits = rev) + ggtitle("a) number of genera housing at least\none pathogenic spec

# create Figure 3b
species2 <- select(species, -species_total) %>%
    pivot_longer(., names_to = "variable", values_to = "value",
        c(2:3))

fig3b <- ggplot() + geom_bar(data = species2, aes(fill = variable,
    y = (value/100), x = forcats::fct_relevel(class,
        to_order)), position = "stack", stat = "identity") +
    geom_bar(data = species, aes(y = species_total,
        x = forcats::fct_relevel(class, to_order)),
```

```
        position = "stack", stat = "identity", fill = "deepskyblue2") +
    geom_bar(data = species, aes(y = established, x = forcats::fct_relevel(class,
        to_order)), position = "stack", stat = "identity",
        fill = "azure4") + theme_bw() + theme(panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), panel.border = element_rect(colour = "black",
        fill = NA, size = 1), axis.title.y = element_blank(),
    axis.title.x = element_text(size = 16), strip.background = element_blank(),
    strip.text = element_text(angle = 0, hjust = 0,
        size = 12), axis.text.x = element_text(size = 12),
    axis.text.y = element_blank(), plot.title = element_text(size = 16),
    legend.position = c(0.878, 0.079), legend.background = element_rect(fill = "white",
        color = "black")) + labs(y = "Number of bacterial pathogens",
    x = "Class", fill = "Pathogen status") + scale_fill_manual(values = c("azure4",
    "deepskyblue2")) + coord_flip() + scale_x_discrete(limits = rev) +
    ggtitle("b) number of bacterial pathogens in each class\n")

fig3a + fig3b
```



We can then recreate Table 1 for the 10 genera which contain the most pathogenic species.

```
d_table <- d_pathogens2 %>%
    select(-class) %>%
    group_by(genus, status) %>%
    tally() %>%
    ungroup() %>%
    spread(., status, n) %>%
    mutate(across(2:3, replace_na, 0), species_total = established +
        putative) %>%
    slice_max(order_by = established, n = 10) %>%
    dplyr::rename(., Genus = genus, Established = established,
        Putative = putative, Total = species_total)

table1 <- gt(d_table)
```

```
table1 <- table1 %>%
    cols_align(align = "center") %>%
    tab_options(table.font.name = "Arial", table.border.top.style = "bold",
        table.border.bottom.style = "solid", table.border.bottom.width = px(3),
        column_labels.border.top.color = "white", column_labels.border.top.width = px(3),
        column_labels.border.bottom.width = px(3),
        data_row.padding = px(10), heading.align = "center") %>%
    opt_row_striping() %>%
    cols_label(Genus = md("**Genus**"), Established = md("**Established**"),
        Putative = md("**Putative**"), Total = md("**Total**")) %>%
    cols_width(columns = c(Genus) ~ px(150), columns = c(Established) ~
        px(120), columns = c(Putative) ~ px(100), columns = c(Total) ~
        px(100)) %>%
    tab_style(style = cell_text(style = "italic"),
        locations = cells_body(columns = Genus))

table1  # here is our table
```

Genus	Established	Putative	Total
Mycobacterium	91	21	112
Corynebacterium	36	20	56
Nocardia	35	18	53
Streptococcus	35	11	46
Staphylococcus	28	3	31
Prevotella	26	10	36
Clostridium	23	7	30
Acinetobacter	20	0	20
Bacteroides	20	7	27
Burkholderia	20	5	25
Rickettsia	20	2	22

```
# gtsave(table1,'Table_1.png', path = '..') #
# here we save it to our figures folder
```