Visualising the phylogenetic distribution of bacterial human pathogens.

Daniel Padfield

05/07/2022

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)

Load in packages and data

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

```
# load packages
library(tidyverse)
library(ggtree)
library(ape)
library(RColorBrewer)
library(ggnewscale)
```

Load in data from GTDB

We will use the bacterial phylogeny produced by GTDB. The Genome Taxonomy Database (GTDB) is an initiative to establish a standardised microbial taxonomy based on genome phylogeny, with the bacterial taxonomy based on genome trees inferred using FastTree from an aligned concatenated set of 120 single copy marker proteins.

We can load in the phylogeny from the most recent release at the time this project was done (r207).

```
# read in tree
tree <- read.tree("https://data.gtdb.ecogenomic.org/releases/release207/207.0/bac120_r207.tree")
tree
```

```
##
## Phylogenetic tree with 62291 tips and 62290 internal nodes.
##
## Tip labels:
##   GB_GCA_001829155.1, GB_GCA_002450905.1, GB_GCA_003645695.1, GB_GCA_016934265.1, GB_GCA_011049595.1
## Node labels:
##   d__Bacteria, '96.0:p__Spirochaetota', 97.0, 55.0, '69.0:c__UBA6919', 39.0, ...
##
## Rooted; includes branch lengths.
```

Each tip of the tree is an accession number of the genome included in GTDB. We need a way to link these accession numbers to the species in our pathogen list. However, our taxonomy is assigned using NCBI

nomenclature and the taxonomy from GTDB is - well - GTDB taxonomy.

With each release GTDB have a file called `bac120_metadata_r207.tar.gz` where r207 represented the release of the database which contains both gtdb and ncbi taxonomic information of each genome. We will use this file to create a dataframe with both ncbi and gtdb taxonomy for each genome in GTDB.

```
# metadata file - its a big file
url <- "https://data.gtdb.ecogenomic.org/releases/release207/207.0/bac120_metadata_r207.tar.gz"
download.file(url, destfile = "tmp.tar.gz")

# unzip file
untar("tmp.tar.gz")

# read in gtdb taxonomy
taxonomy <- read_tsv("bac120_metadata_r207.tsv", col_names = TRUE) %>%
    # rename columns
select(accession, gtdb_taxonomy, ncbi_taxonomy)
```

```
## Warning: One or more parsing issues, see `problems()` for details
```

We can then remove all entries that are not present in the phylogenetic tree and split the taxonomies into each phylogenetic level.

```
taxonomy <- filter(taxonomy, accession %in% tree$tip.label) %>%
    # split taxonomy into columns for each level
separate(gtdb_taxonomy, c("kingdom", "phylum", "class",
    "order", "family", "genus", "species"), sep = ";") %>%
    # get rid of p__ c__ etc
mutate(across(kingdom:species, function(x) {
    gsub(".*__", "", x)
})) %>%
    separate(ncbi_taxonomy, c("ncbi_kingdom", "ncbi_phylum",
        "ncbi_class", "ncbi_order", "ncbi_family",
        "ncbi_genus", "ncbi_species"), sep = ";") %>%
    mutate(across(ncbi_kingdom:ncbi_species, function(x) {
        gsub(".*__", "", x)
    }))
```