Twitter Thread by Pat Schloss

Pat Schloss
@PatSchloss

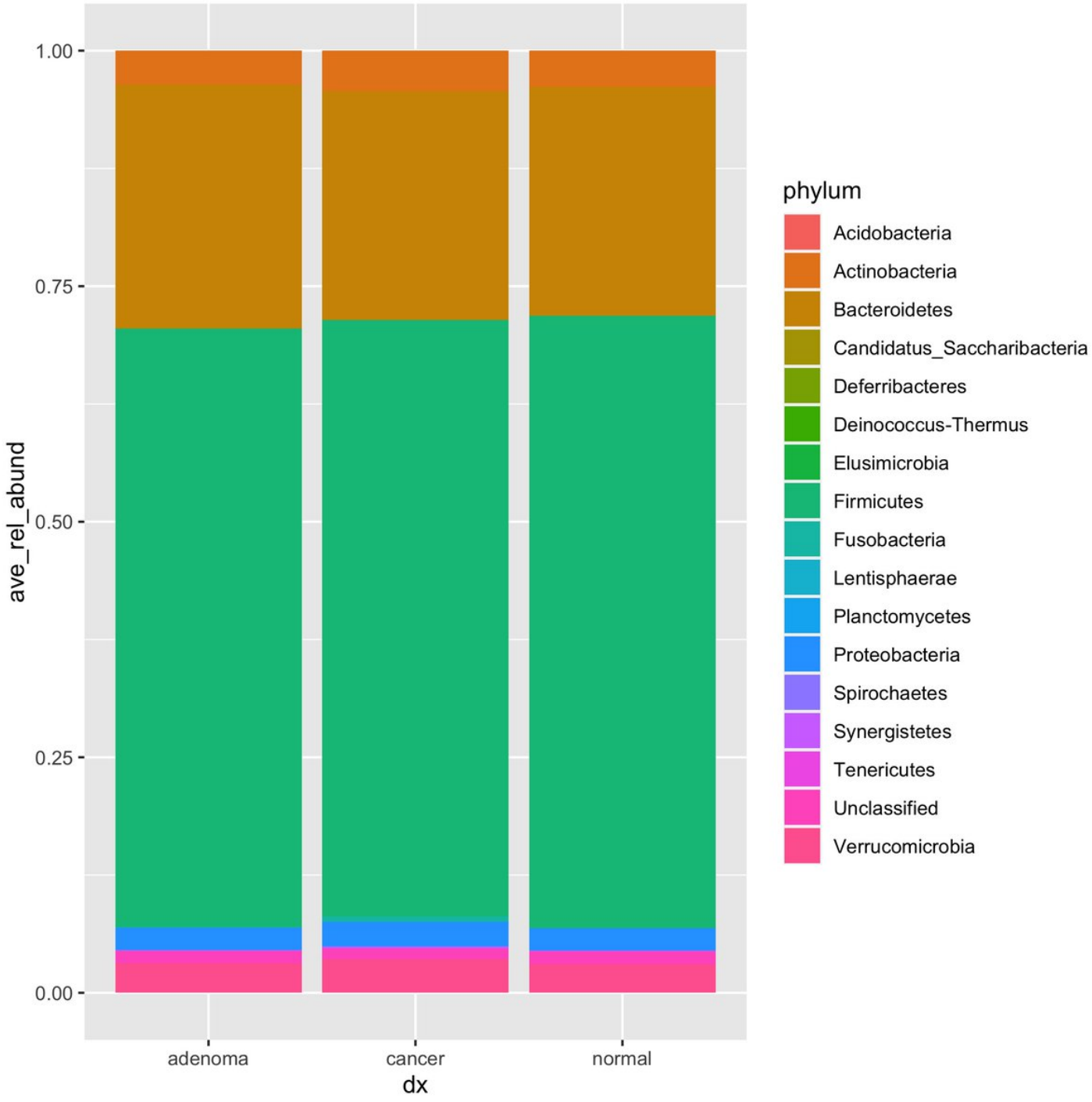
Wellll... A few weeks back I started working on a tutorial for our lab's Code Club on how to make shitty graphs. It was too dispiriting and I balked. A twitter workshop with figures and code:

> When are you doing pie charts?
>
> — #BlackLivesMatter (@surt_lab) October 13, 2020

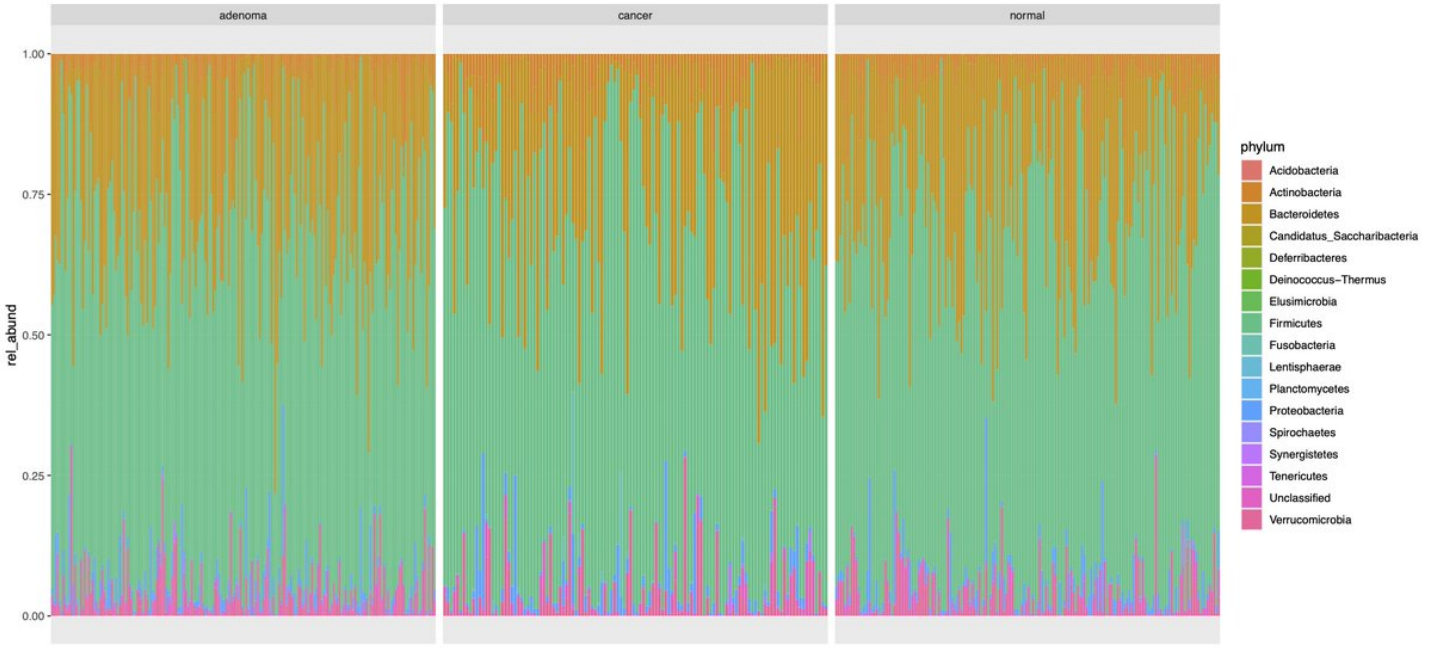
Here's the code to generate the data frame. You can get the "raw" data from https://t.co/jcTE5t0uBT

```
library(tidyverse)

taxonomy <- read_tsv("data/baxter.cons.taxonomy") %>%
  select(-Size) %>%
  mutate(Taxonomy = str_replace(Taxonomy, ";$", "")) %>%
  mutate(Taxonomy = str_replace_all(Taxonomy, "\\(\\d*\\)", "")) %>%
  separate(Taxonomy,
    into=c("kingdom", "phylum", "class", "order", "family", "genus"),
    sep=";") %>%
  mutate(phylum = str_replace(phylum, "^u", "U")) %>%
  select(OTU, phylum)

metadata <- read_tsv("data/baxter.metadata.tsv",
    col_types=cols(sample=col_character())) %>%
  select(sample, dx)

read_tsv("data/baxter.subsample.shared",
    col_type=cols(Group=col_character(), .default=col_double())) %>%
  select(-label, -numOtus) %>%
  rename(sample = Group) %>%
  pivot_longer(-sample, names_to="OTU", values_to="count") %>%
  mutate(rel_abund = count / 10530) %>%
  select(-count) %>%
  inner_join(., taxonomy, by="OTU") %>%
  group_by(sample, phylum) %>%
  summarize(rel_abund = sum(rel_abund), .groups="drop") %>%
  inner_join(metadata, ., by="sample") %>%
  write_tsv("dx_phylum_relabund.tsv")
```

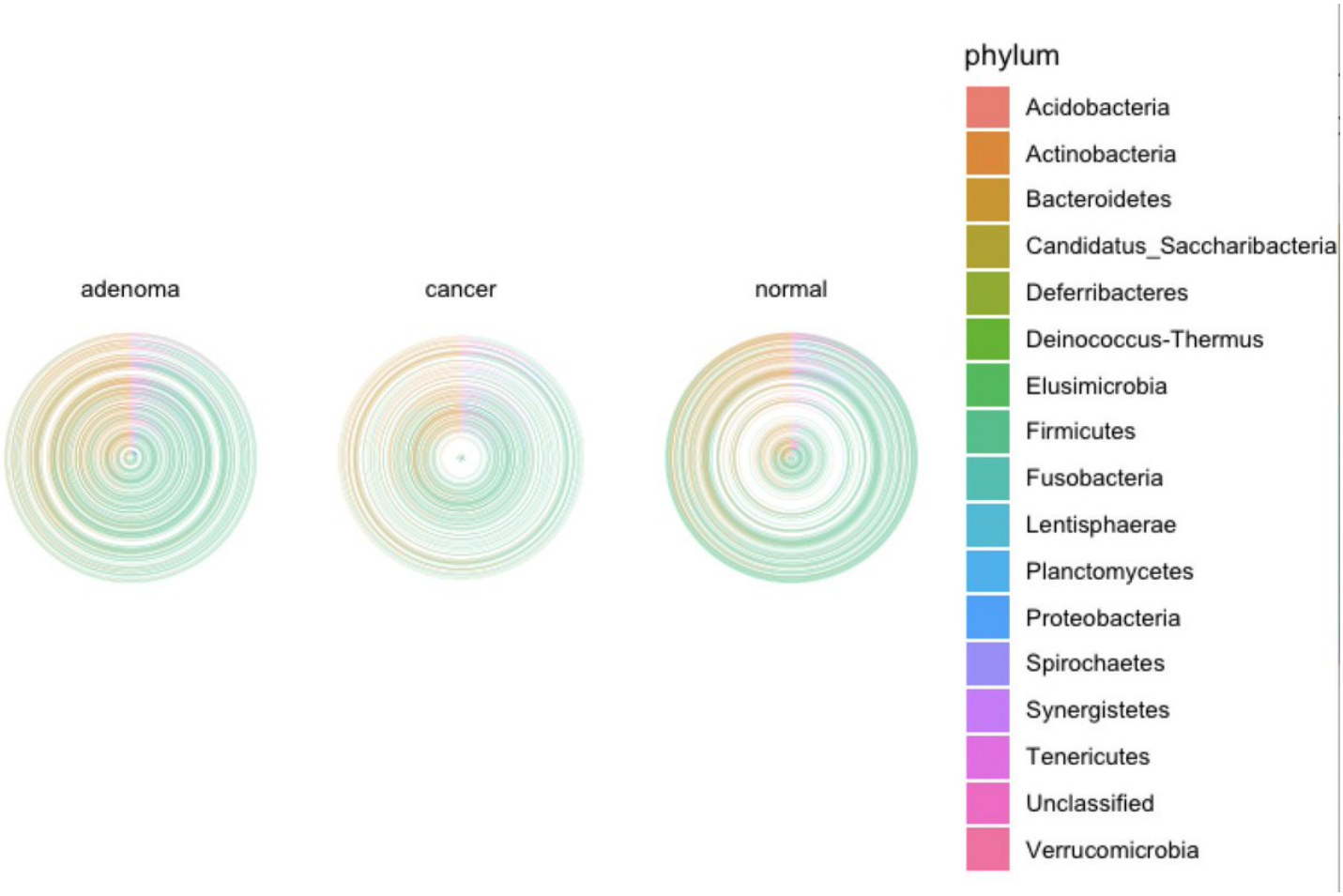
Obligatory stacked bar chart that hides any sense of variation in the data



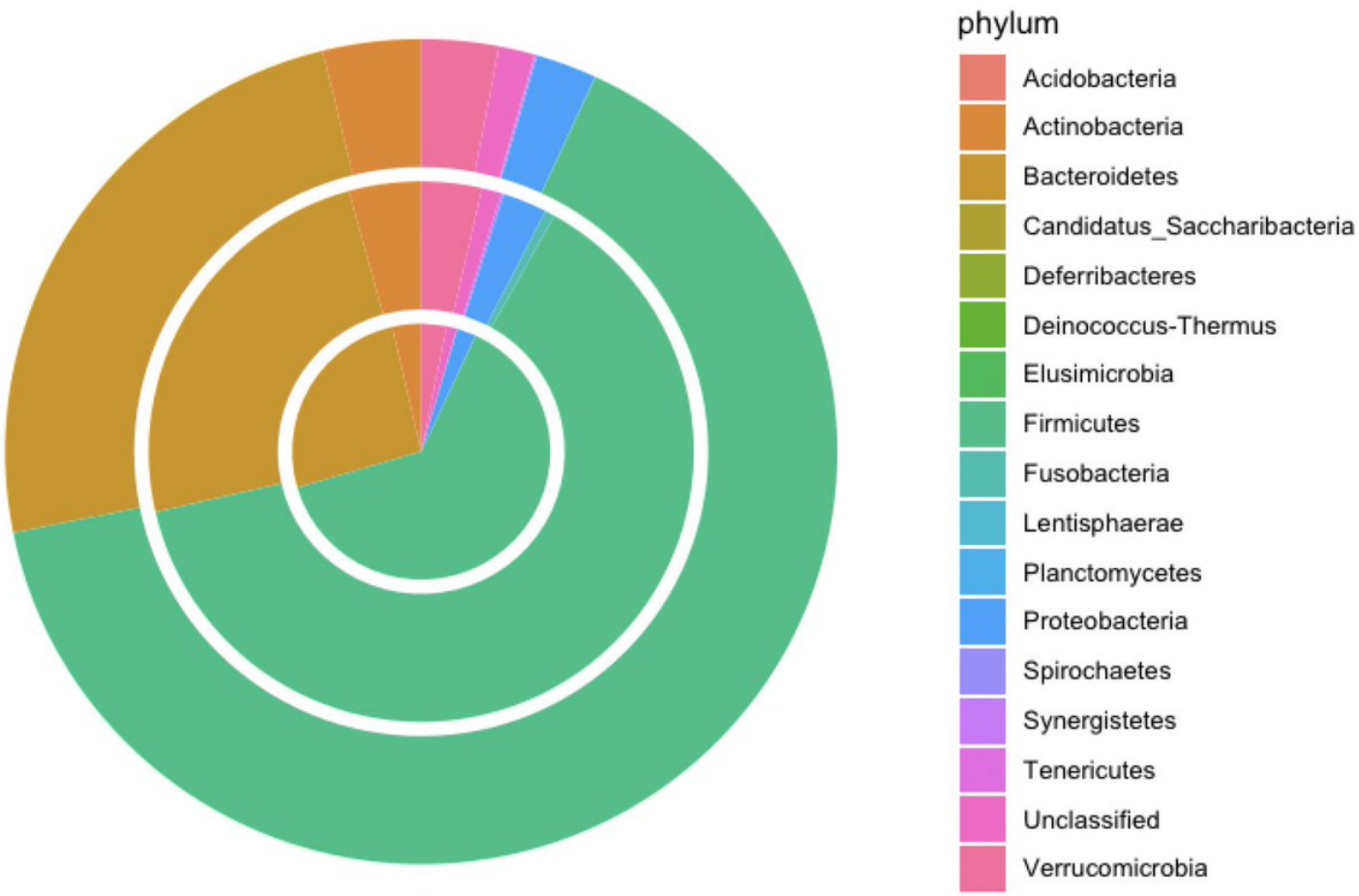
Obligatory stacked bar chart that shows all the things and yet shows absolutely nothing at the same time



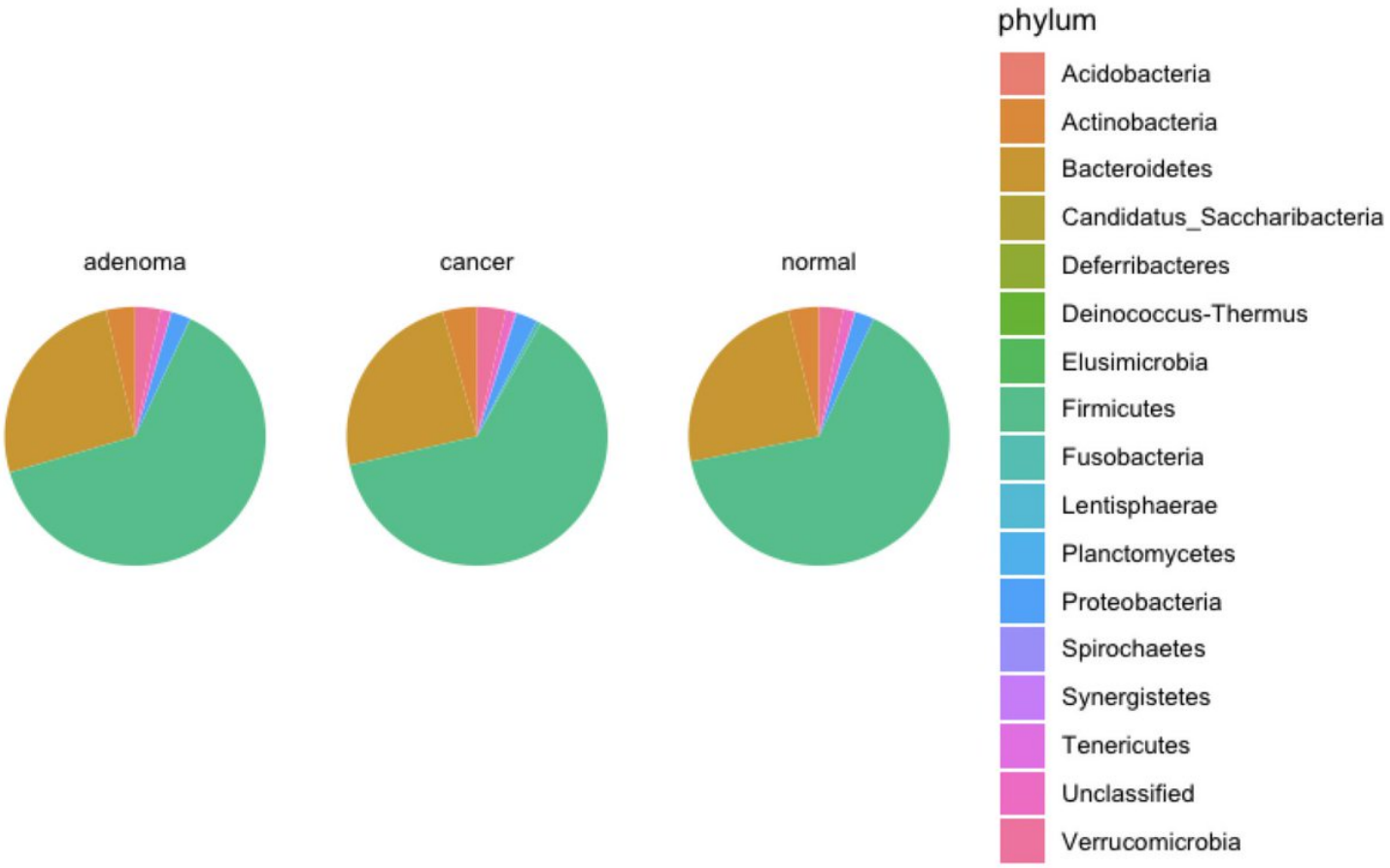
STACKED Donut plot. Who doesn't want a donut? Who wouldn't want a stack of them!?! This took forever to render and looked worse than it should because coord_polar doesn't do scales="free_x".



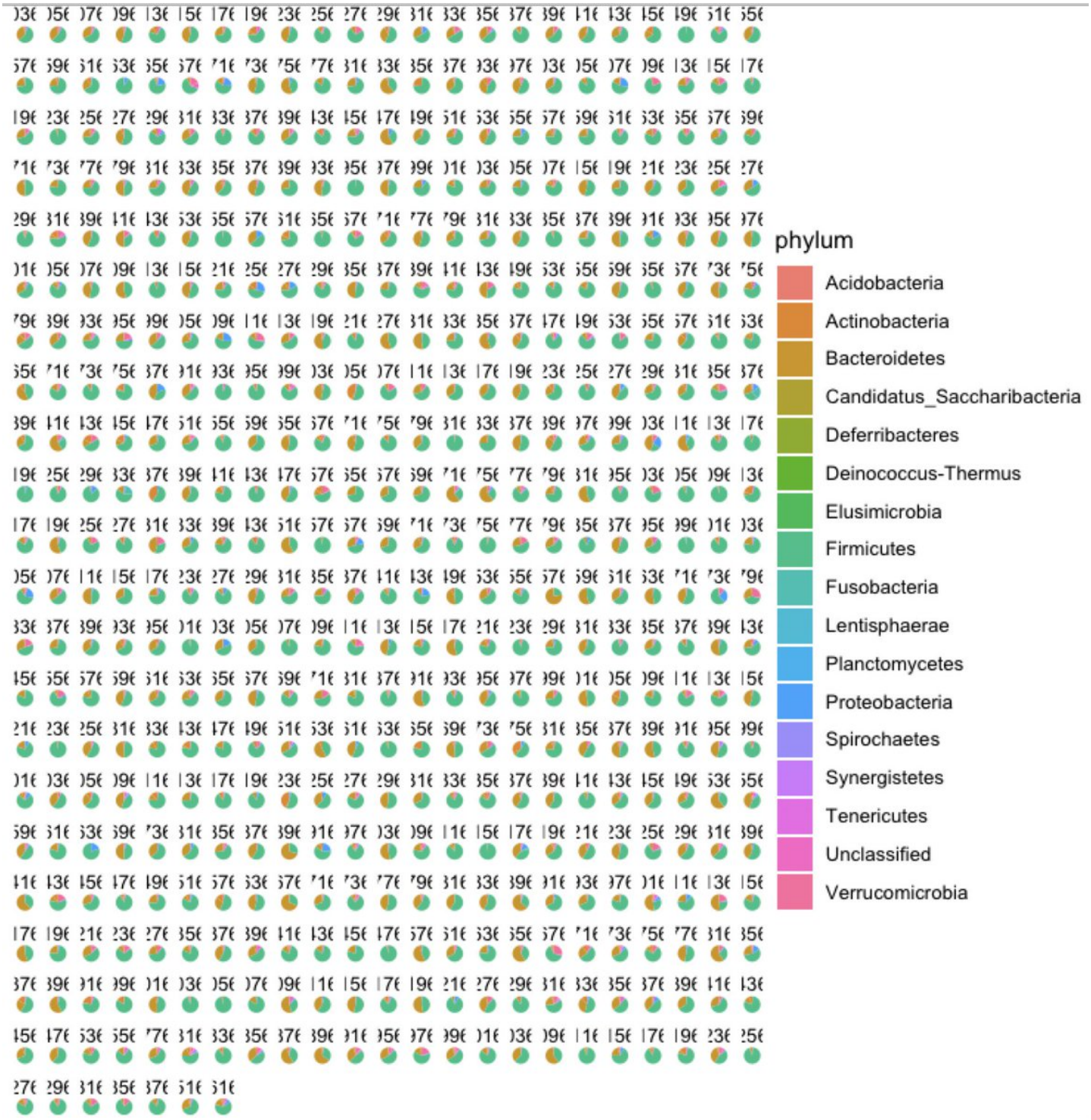
More donuts. Let's get rid of all that messy variation in the data



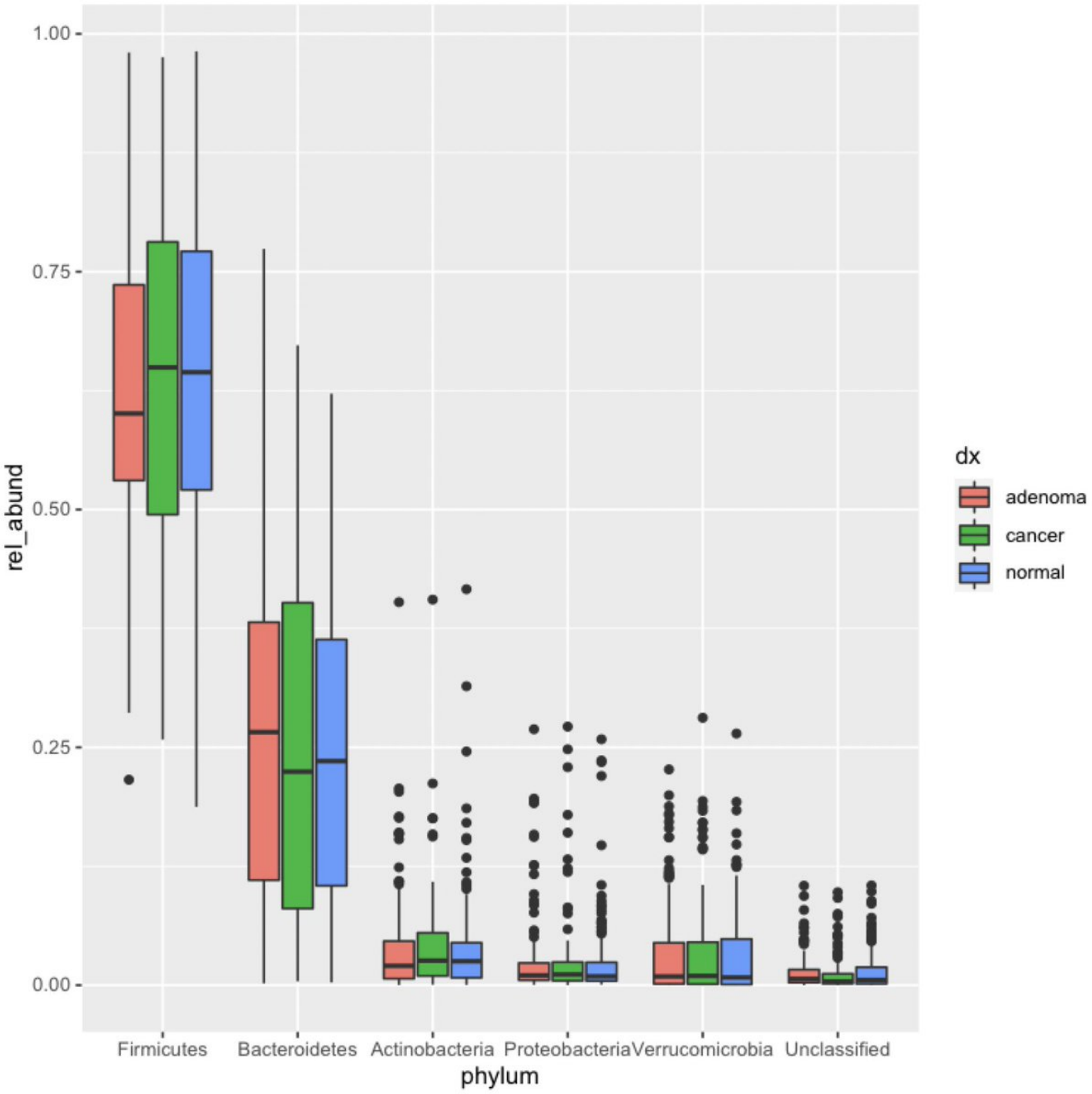
One pie for @surt_lab, one for @watermicrobe, and one waiting to explode for @a2binny



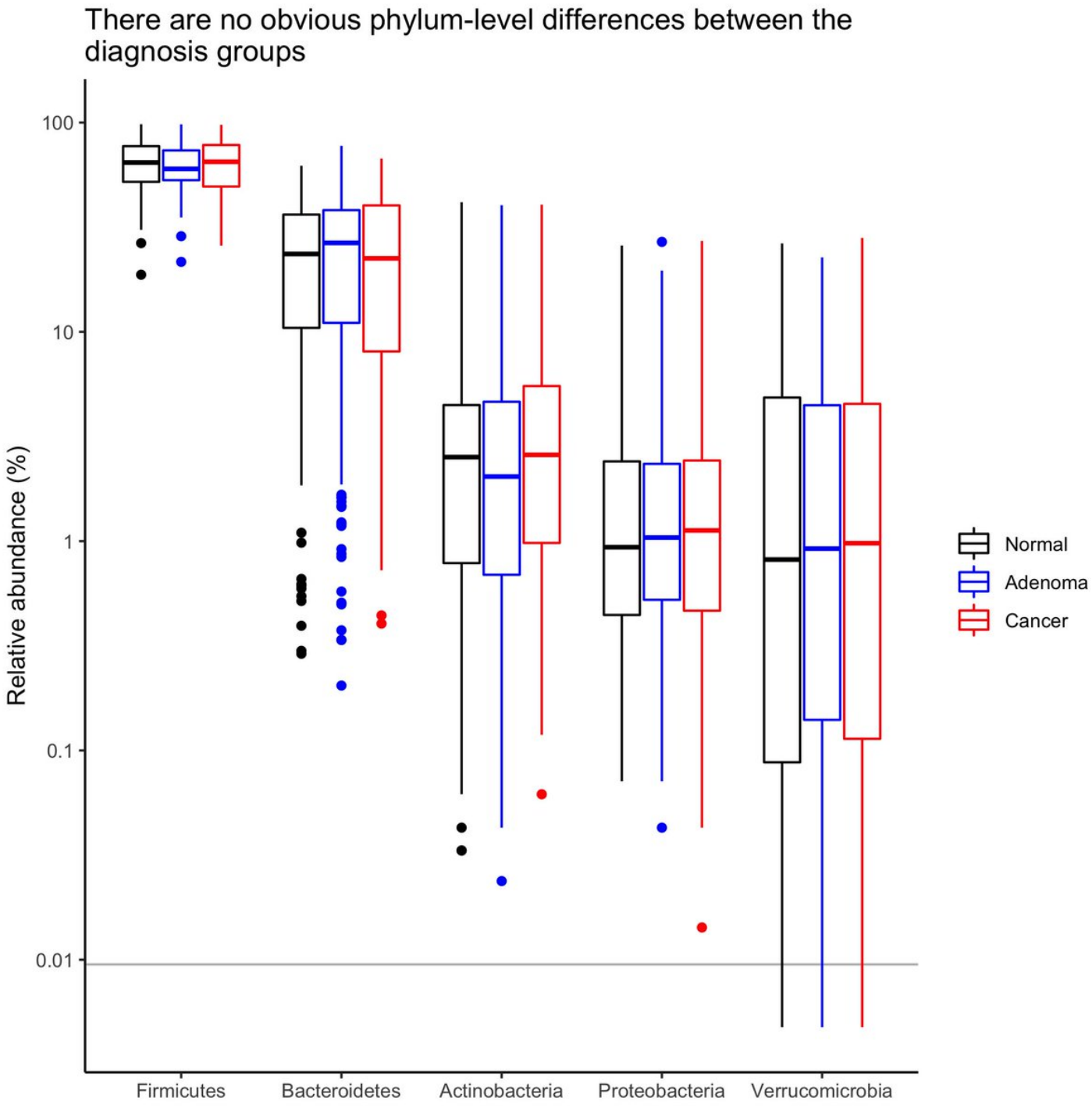
Fine. Here's a pie for those of you that are still watching... This also took forever to render. The numbers are subject IDs



In all seriousness, here's the type of plot that I encourage for showing relative abundance by taxonomic data. Not fully polished, but you get the idea. Here each diagnosis has about 160 samples. With fewer samples, I'd use geom_jitter rather than geom_histogram



I prefer the boxplot/jitter plot because it allows the viewer to directly compare what I think is important. It also shows the variation in the data. Here's more polished version.



You can see how to do this for other taxonomic levels, incorporate statistical analysis to pick levels to show, and how to add a log scale on y-axis at https://t.co/U30ehefQPE. Thanks for attending my twitter workshop.