BIOS0052: Human And Ecosystem Health In A Changing World

Week 10 practical session

Telling stories with data on ecology and environmental change

A critical aspect of data analysis and research is accurately and succinctly communicating your findings to other people, in particular via visualisation of data and model results. This is often especially important if we want our research to have a wider influence on the public and policy decisions. This means that the decisions we make around data visualisation are extremely important - we need to ensure we are representing the data and findings as precisely as possible, while finding creative ways to tell stories to a wider audience (usually including non-specialists).

The aim of today's short practical is to explore different ways of visually telling stories with the same dataset, using the R data visualisation package *"ggplot2"*. We will examine how different decisions we make around what outcome to visualise, and what visual approach to take, can significantly modify the stories that the data tell - and think about our responsibility as scientists and communicators to tell these stories accurately. The ggplot2 package (the "gg" stands for "grammar of graphics") is a hugely powerful, useful and flexible framework for data visualisation. Its approach focuses on defining a plot space to which we add visual elements - called "geoms" - which are almost infinitely customisable in terms of look, shape, colour etc. If you are interested in learning ggplot in more depth, this is a good tutorial resource to get started on the concepts.

This workbook is structured as a series of code snippets with short reflective questions, which are prompts to consider and discuss the results and visualisations. Hopefully this workshop will give you some ideas for how to explore and visualise the data for your assessment (and research projects!).

All the data you'll need for the workshop are in the GitHub, in the "Week10-Data-Visualisation" folder. Please download the whole folder using download-directory.github.io. Set this folder as your working directory, then all the materials you will need are contained within the "data" subfolder. The bat image in this file is by Fritz Geller-Grimm via Wikipedia.

```
# package dependencies
# use the "install.packages()" command if not already installed
library(ggplot2); library(dplyr); library(magrittr); library(sf); library(stringr)
library(rstudioapi); library(tidyr); library(Hmisc); library(pals); library(MetBrewer)

# automatically set working directory
# (or if this doesn't work,
# manually set your working directory to the folder "Week10-Data-Visualisation")
PATH = dirname(rstudioapi::getSourceEditorContext()$path)
setwd(PATH)
```

Today's research question: are bats "special" viral reservoirs?

In recent years - and especially since the 2013-16 Ebola epidemic and COVID-19 pandemic - a great deal of research and debate has focused on the question of whether bats (**Chiroptera**) are unusually effective reservoirs for high-consequence viruses. The eco-evolutionary hypotheses around this question are based in evidence that bats' physiological adaptations for powered flight, including modified immune responses to cope with the energetic demands of flight, might also enhance their ability to tolerate viral infections. Research into this question has focused variously on bat immunology, genomics and disease ecology, but also on bats' documented viral diversity. One hypothesis is that bats may host an unusually high diversity of viruses, including those that are able to infect humans ("zoonoses"). These include numerous high-consequence infections such as betacoronaviruses (e.g. SARS-CoV-1 and SARS-CoV-2), filoviruses (e.g. Ebola, Marburg) and lyssaviruses (e.g. rabies).

This topic became of wider public interest following COVID-19. The question of whether bats could be considered a "threat" to human health is obviously of high-concern for both public health and conservation, so communicating this evidence precisely is critically important. So today we will explore this question and communicating our findings using data on bats' documented viral diversity, in comparison to other mammals. These data were compiled from a huge dataset called **VIRION ("The Global Virome, In One Network")**, the largest and most comprehensive dataset of documented host-virus associations across the vertebrate web of life. These data are termed "host-virus association data" - this means that each record is a documented association between a host species (here, mammals) with a viral species, with evidence ranging from serology (antibody-based evidence of past infection), to genetic detection (e.g. PCR test), to viral isolation from biological samples. To read more about VIRION, the documentation is here and the original database paper is here. A great synthesis of the usefulness of this type of data is found in Albery et al. 2021, "The science of the host-virus network".

Telling stories with viral diversity data

A dataset of unique wild mammal host-virus associations derived from the original VIRION database is provided in the "data" subfolder. Each row contains information on a host and virus species with a documented association, the higher host and viral taxonomy (Order, Family, Genus), our current understanding of whether the virus is zoonotic, the number of times this virus has been reported in that species, the year of the first and last record, and the type of detection methods used. We'll read in the data and take a look.

```
# read VIRION data csv
vir = read.csv("data/virion_mammals.csv")

# look at this dataset - how many records are there?
head(vir)
nrow(vir)

# we can look at the number of records across different host and viral orders
# what do you notice about bats? (recall that bats' order is "Chiroptera")
table(vir$HostOrder)
table(vir$VirusOrder)
```

Let's start by looking at the dataset and examining how many hosts and viruses are represented in the data, both within bats and within mammals overall.

```
# how many host species are in the data?
dplyr::n_distinct(vir$Host)

# how many virus species?
dplyr::n_distinct(vir$Virus)

# how many species within each mammal order?
# which order has the most species represented, and which the least?
vir %>%
  dplyr::group_by(HostOrder) %>%
  dplyr::summarise(NumSpecies = n_distinct(Host))
```

We can ask lots of different questions with these data to get a picture of viral diversity in bats, and whether it is truly unusually high compared to other mammal groups. Let's start with the simplest - we'll examine the overall number of virus species, and viral genera, known in bats compared to all other mammals. There are lots of ways to visually represent this but we can start with a scatterplot.

```
# calculate, for each mammal order, the number of virus species and number of viral genera
vir_total = vir %>%
  dplyr::group_by(HostOrder) %>%
  dplyr::summarise(ViralRichness = n_distinct(Virus),
                   ViralGenera = n_distinct(VirusGenus)) %>%
```

```
    dplyr::mutate(HostOrder = Hmisc::capitalize(HostOrder))

# plot a scatterplot
vir_total %>%
  ggplot() +
  geom_point(aes(ViralRichness, HostOrder), size=4, color="coral2") +
  theme_bw() +
  xlab("Number of virus species") + ylab("Mammal Order")
```

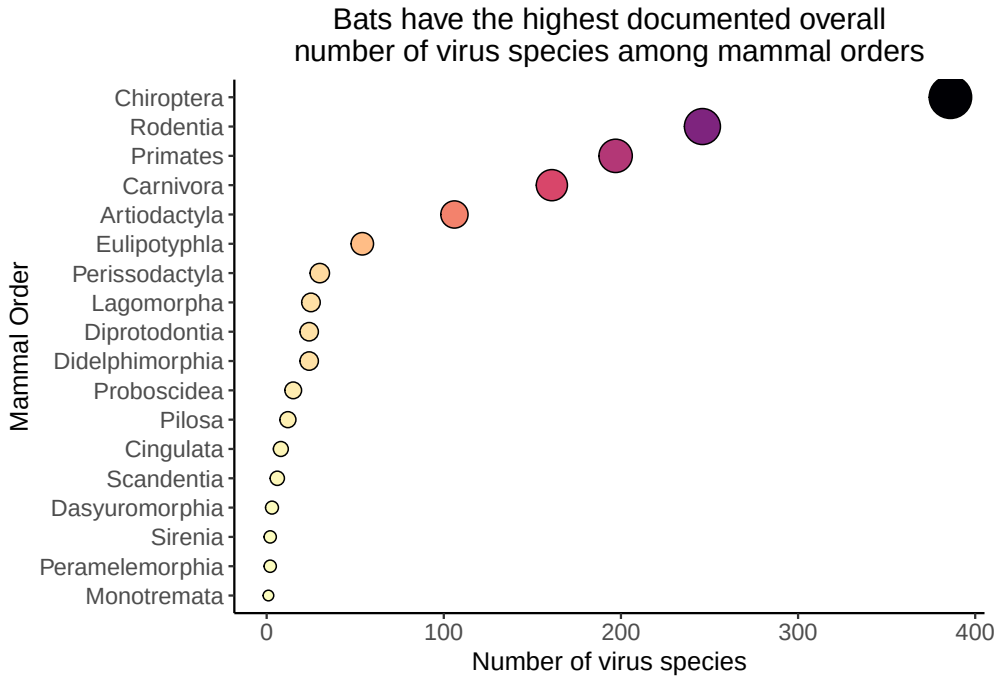
Is this visual representation useful, especially for a non-expert? How could we make this more intuitive to understand?

```
# one option to improve interpretability: we can order points by viral richness
# what does this suggest about bats' viral diversity?
vir_total %>%
  dplyr::arrange(ViralRichness) %>%
  dplyr::mutate(HostOrder = factor(HostOrder, levels=HostOrder, ordered=TRUE)) %>%
  ggplot() +
  geom_point(aes(ViralRichness, HostOrder), size=4, color="coral2") +
  theme_bw() +
  xlab("Number of virus species") + ylab("Mammal Order") +
  theme(axis.text = element_text(size=11),
    axis.title = element_text(size=12),
    plot.title = element_text(size=14, hjust=0.5))
```

Another way to add further visual emphasis is to colour and size scale the points. We do this by mapping ViralRichness to the "colour" and "size" dimensions of the geom_point() aesthetic. Here, the colour and size of the point scales with increasing viral richness. We can also add a declarative title to draw a viewer's eye to the lead result we want to emphasize.

Is this plot easier to interpret? What does it do well, and what does it do less well?

```
# colour and size scale points by viral richness
# add a title
vir_total %>%
  dplyr::arrange(ViralRichness) %>%
  dplyr::mutate(HostOrder = factor(HostOrder, levels=HostOrder, ordered=TRUE)) %>%
  ggplot() +
  geom_point(aes(ViralRichness, HostOrder, size=ViralRichness, fill=ViralRichness), pch=21) +
  theme_classic() +
  scale_size(range=c(2, 9)) +
  theme(legend.position="none") +
  scale_fill_viridis_c(option="magma", direction=-1) +
  xlab("Number of virus species") + ylab("Mammal Order") +
  ggtitle("Bats have the highest documented overall\nnumber of virus species among mammal orde
  theme(axis.text = element_text(size=11),
      axis.title = element_text(size=12),
      plot.title = element_text(size=14, hjust=0.5))
```



Now let's take the same approach for another measure of viral diversity - this time, the diversity of viral genera. Defining diversity at the genus level is probably more conservative - some viral general, such as betacoronaviruses, have been intensely studied so their species richness is likely to be inflated in the data. Calculating the richness of viral genera might help to alleviate these issues.

```
# colour and size scale points by viral richness
# add a title
vir_total %>%
  dplyr::arrange(ViralGenera) %>%
  dplyr::mutate(HostOrder = factor(HostOrder, levels=HostOrder, ordered=TRUE)) %>%
  ggplot() +
  geom_point(aes(ViralGenera, HostOrder, size=ViralGenera, fill=ViralGenera), pch=21) +
  theme_classic() +
  scale_size(range=c(2, 9)) +
  theme(legend.position="none") +
  scale_fill_viridis_c(option="magma", direction=-1) +
  xlab("Number of virus genera") + ylab("Mammal Order") +
  ggtitle("Bats have the fifth highest known number of\n viral genera among mammal orders") +
  theme(axis.text = element_text(size=11),
      axis.title = element_text(size=12),
      plot.title = element_text(size=14, hjust=0.5))
```

- **Q1.** What is different when we define viral diversity at the level of the genus? How does this change the story we are telling?

Different metrics and visual approaches tell different stories about bat viral diversity

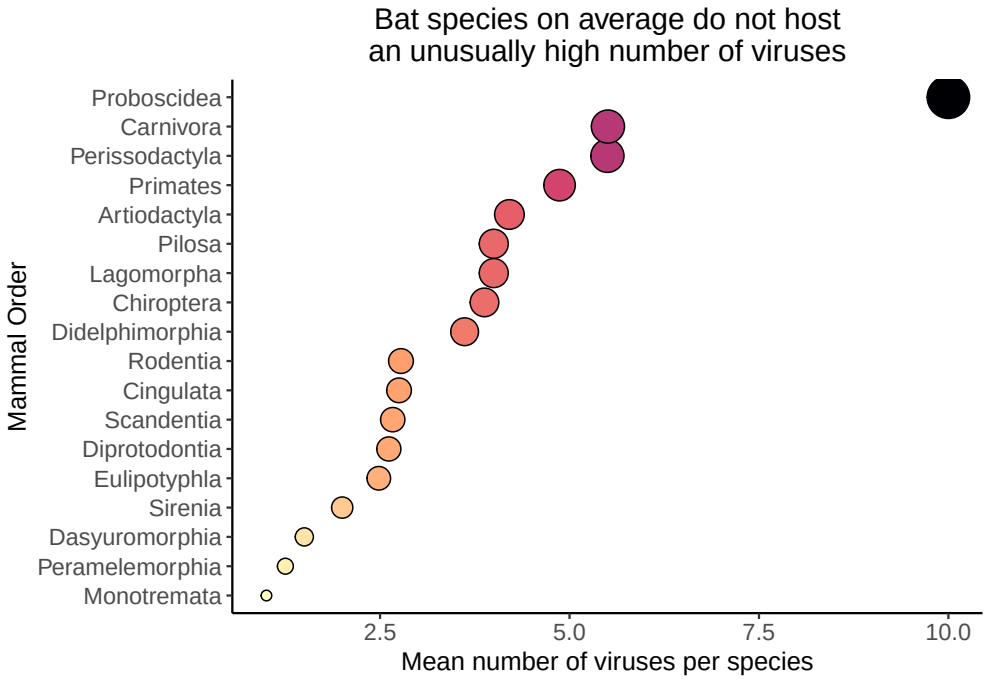
We defined viral diversity above, very simply, as the number of virus species or genera detected across all bats. However, this is only one very coarse way to define a viral diversity metric. It's also potentially misleading for the public - for a non-specialist, the above visualisations could easily be interpreted as "any specific bat individual or species carries lots of viruses", rather than the more nuanced reality that this the total summarised across *all* bats. This is potentially a risky message from a conservation perspective that could encourage the persecution of bats.

We could instead explore different ways to tell this story that do a better job of reflecting viral diversity at the *species* level. This is ultimately more relevant to how people will interpret the data. First, let's use the same visual approach, but instead plot the *average species-level viral richness* across all species in each mammalian Order. This tells us the average number of viruses known to be carried by each species in our dataset - but note that our data does not include counts of zero (i.e. species with no known viruses).

```
# first calculate the viral richness per host species
vr_sp = vir %>%
  dplyr::group_by(Host) %>%
  dplyr::summarise(ViralRichness = n_distinct(Virus),
                   HostOrder = head(HostOrder, 1) ) %>%
  dplyr::mutate(HostOrder = Hmisc::capitalize(HostOrder))


# then calculate mean host-level richness per order
vr_means = vr_sp %>%
  dplyr::group_by(HostOrder) %>%
  dplyr::summarise(ViralRichness = mean(ViralRichness))

vr_means %>%
  dplyr::arrange(ViralRichness) %>%
  dplyr::mutate(HostOrder = factor(HostOrder, levels=HostOrder, ordered=TRUE)) %>%
  ggplot() +
  geom_point(aes(ViralRichness, HostOrder, size=ViralRichness, fill=ViralRichness), pch=21) +
  theme_classic() +
  scale_size(range=c(2, 9)) +
  theme(legend.position="none") +
  scale_fill_viridis_c(option="magma", direction=-1) +
  xlab("Mean number of viruses per species") + ylab("Mammal Order") +
  ggtitle("Bat species on average do not host\nan unusually high number of viruses") +
  theme(axis.text = element_text(size=11),
      axis.title = element_text(size=12),
      plot.title = element_text(size=14, hjust=0.5))
```



This visualisation tells us something quite different about bat viral diversity - that, although bats in total host a very large number of known viruses, this does not translate to a particularly high individual species-level viral diversity. This is because there are a *lot* of bat species, and quite a lot of them have been sampled for viruses.

However, how useful is a mean? This tells us about the central tendency but very little about the spread of the data at species-level - it's possible a few bat species host an unusually large number of viruses, even if bat species overall do not. To do this, we can explore visualising both the central tendency and the individual species points, for example via boxplots. This is a more technical visualisation, but it once again helps to tell us a more nuanced story.

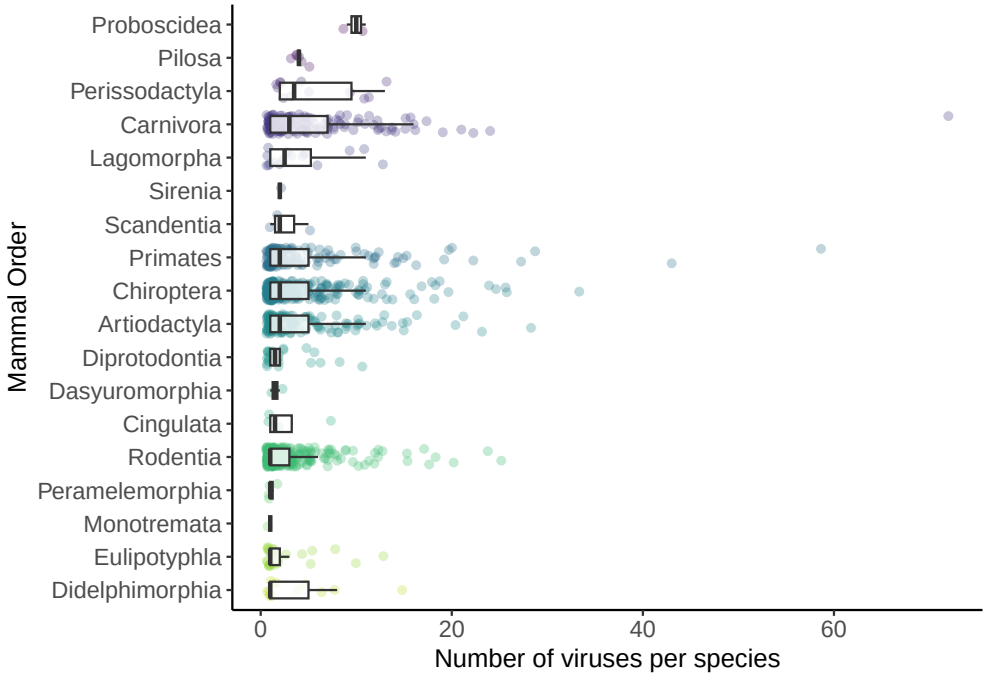
```
# we can see a much more holistic picture of the data, here
# but the way mammal Orders are arranged doesn't necessarily make sense
vr_sp %>%
  ggplot() +
  geom_jitter(aes(ViralRichness, HostOrder, color=HostOrder), alpha=0.3, height=0.3) +
  geom_boxplot(aes(ViralRichness, HostOrder), alpha=0.8, width=0.5, outliers = FALSE) +
  theme_classic() +
  guides(color="none") +
  xlab("Species-level viral richness") +
  ylab("Mammal Order") +
  scale_color_viridis_d(end=0.9, direction=-1) +
  theme(axis.text = element_text(size=11),
    axis.title = element_text(size=12),
    plot.title = element_text(size=14, hjust=0.5))

# for a more intuitive visual representation
# we could arrange mammal orders based on the median
```

```r
# calculate medians
vr_sp_medians = vr_sp %>%
  dplyr::group_by(HostOrder) %>%
  dplyr::summarise(ViralRichness_Median = median(ViralRichness)) %>%
  dplyr::arrange(ViralRichness_Median)

# arrange orders by median in vr_sp dataframe
vr_sp = vr_sp %>%
  dplyr::mutate(HostOrder = factor(HostOrder, levels=vr_sp_medians$HostOrder, ordered=TRUE))

# plot again
vr_sp %>%
  ggplot() +
  geom_jitter(aes(ViralRichness, HostOrder, color=HostOrder), alpha=0.3, height=0.3) +
  geom_boxplot(aes(ViralRichness, HostOrder), alpha=0.8, width=0.5, outliers = FALSE) +
  theme_classic() +
  guides(color="none") +
  xlab("Number of viruses per species") +
  ylab("Mammal Order") +
  scale_color_viridis_d(end=0.9, direction=-1) +
  theme(axis.text = element_text(size=11),
    axis.title = element_text(size=12),
    plot.title = element_text(size=14, hjust=0.5))
```



- **Q2.** What does this figure tell us about how (known) viral diversity is distributed across mammalian orders? This figure currently doesn't have a title - can you think of a suitable title that summarises this nuanced story?

There is a lot of unused white space on this figure, which isn't an ideal use of space. Is there a way to make better use of the space? Sometimes log-transforming data can help to use more space without losing nuance; let's give that a try.

```
# adding a log transformation to the X axis scale
vr_sp %>%
  ggplot() +
  geom_jitter(aes(ViralRichness, HostOrder, color=HostOrder), alpha=0.3, height=0.3, size=2) +
  geom_boxplot(aes(ViralRichness, HostOrder), alpha=0.8, width=0.5, outliers = FALSE) +
  theme_classic() +
  guides(color="none") +
  xlab("Number of viruses per species") +
  ylab("Mammal Order") +
  scale_color_viridis_d(end=0.9, direction=-1) +
  theme(axis.text = element_text(size=11),
    axis.title = element_text(size=12),
    plot.title = element_text(size=14, hjust=0.5))+
  scale_x_continuous(trans="log10") # log10 transform
```

- **Q3.** Does log-transforming improve this visualisation, and if so why? If not, why not?

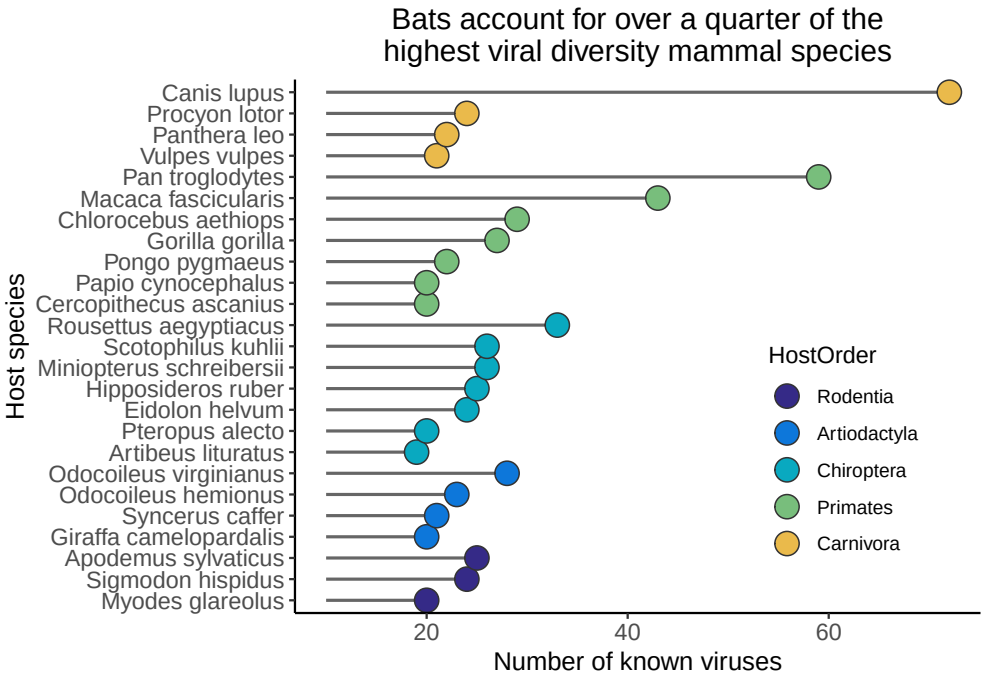
These plots have revealed that some species do have exceptionally high known diversity of viruses. One further option would be to visualise these species, to provide a visual exploration of potential species of high interest or risk. Let's examine this.

```
# take only the top 25 most virus rich species
# arrange them by Order and by richness
sp_25 = vr_sp %>%
  dplyr::arrange(desc(ViralRichness)) %>%
  dplyr::slice_head(n=25) %>%
  dplyr::arrange(HostOrder, ViralRichness) %>%
  dplyr::mutate(Host = Hmisc::capitalize(Host),
                Host = factor(Host, levels=Host, ordered=TRUE))

# how many of top 25 are bats? 7 (28%)
table(sp_25$HostOrder)

# visualise
sp_25 %>%
  ggplot() +
  geom_segment(aes(y=Host, x=10, xend=ViralRichness), size=0.7, color="grey40") +
  geom_point(aes(ViralRichness, Host, fill=HostOrder), color="grey20", pch=21, size=5) +
  theme_classic() +
  scale_fill_manual(values = pals::parula(6)) +
  xlab("Number of known viruses") +
  ylab("Host species") +
  ggtitle("Bats account for over a quarter of the\nhighest viral diversity mammal species") +
```

```
  theme(legend.position.inside = c(0.8, 0.3),
        legend.position = "inside") +
  theme(axis.text = element_text(size=11),
  axis.title = element_text(size=12),
  plot.title = element_text(size=14, hjust=0.5))
```



- **Q4.** What do you notice about the high viral diversity species? If you are unfamiliar, take the time to look up some of these species names. What do you think might be the reasons for such a high documented viral richness in some of these species?

Do bats host an unusual number of zoonotic viruses?

So far we have only been looking at the overall diversity of viruses documented in bats (and other mammalian orders). But we also have information on *zoonotic* viruses - those known to infect humans, and which therefore pose a potential public health risk. So another question we could ask and communicate with these data is, are bat viruses unusually likely to be zoonotic?

Let's start by summarising the total number of viruses and zoonotic viruses in each mammal order, and then visualising these like we did at the start of this workshop.

```
# calculate total and zoonotic virus richness
vr_zoo = vir %>%
  dplyr::group_by(HostOrder) %>%
  dplyr::summarise(ViralRichness = n_distinct(Virus),
                   ZoonoticRichness = n_distinct(Virus[VirusZoonotic==TRUE])) %>%
  dplyr::mutate(HostOrder = Hmisc::capitalize(HostOrder))
```

```
# visualise
vr_zoo %>%
  dplyr::arrange(ZoonoticRichness) %>%
  dplyr::mutate(HostOrder = factor(HostOrder, levels=HostOrder, ordered=TRUE)) %>%
  ggplot() +
  geom_point(aes(ZoonoticRichness, HostOrder, size=ZoonoticRichness, fill=ZoonoticRichness), p
  theme_classic() +
  scale_size(range=c(2, 9)) +
  theme(legend.position="none") +
  scale_fill_viridis_c(option="magma", direction=-1) +
  xlab("Number of virus species") + ylab("Mammal Order") +
  ggtitle("Primates and rodents host the highest\nnumber of known zoonotic viruses") +
  theme(axis.text = element_text(size=11),
      axis.title = element_text(size=12),
      plot.title = element_text(size=14, hjust=0.5))
```

- **Q5.** By far the highest number of known zoonoses are hosted by primates, followed by rodents. Does this make sense biologically or ecologically? Why?

The problem with this visualisation is that it doesn't directly tell us anything about whether bats host an unusually high number of zoonoses relative to their overall viral diversity; it doesn't account for how well-studied or speciose bats are, for example. Let's try a different visualisation approach. Here, we can explicitly estimate whether zoonotic virus richness is lower or higher than expected for each order, given its overall viral richness. To do this, we fit a regression model of the relationship between total and zoonotic viral richness, and extract its residuals.

```
# scatter plot of zoonotic vs total viral richness
# we can see these are highly correlated
vr_zoo %>%
  ggplot() +
  geom_point(aes(ViralRichness, ZoonoticRichness), size=3, color="coral2") +
  theme_classic() +
  theme(axis.text = element_text(size=11),
    axis.title = element_text(size=12),
    plot.title = element_text(size=14, hjust=0.5)) +
  xlab("Total viral richness") + ylab("Zoonotic viral richness")

# we can add a fitted line using Poisson regression to show expected
# zoonotic richness for each value of total viral richness
# points that fall below the fitted line have unusually low zoonotic richness
# and vice versa
vr_zoo %>%
  ggplot() +
  geom_point(aes(ViralRichness, ZoonoticRichness), size=3, color="coral2") +
  geom_smooth(aes(ViralRichness, ZoonoticRichness), method = "glm",
              se = F, method.args = list(family = "poisson")) +
```

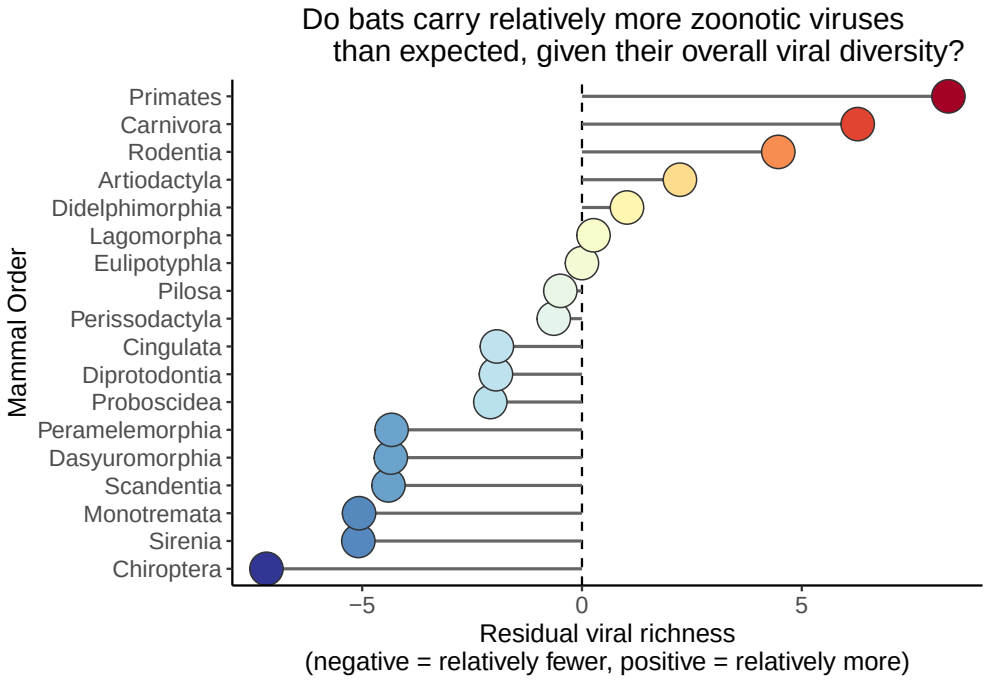
```
  theme_classic() +
  theme(axis.text = element_text(size=11),
    axis.title = element_text(size=12),
    plot.title = element_text(size=14, hjust=0.5)) +
  xlab("Total viral richness") + ylab("Zoonotic viral richness")

# we can see this isn't a great model!
# not very well fitted to the data
# if we were doing something more rigorous (rather than this workshop)
# we might explore other options
# but for now let's consider this to be acceptable

# fit the model outside the ggplot call and extract the residuals
mod = glm(ZoonoticRichness~ViralRichness, data=vr_zoo, family="poisson")
vr_zoo$ResidualRichness = resid(mod)

# arrange by residual richness
vr_zoo = vr_zoo %>%
  dplyr::arrange(ResidualRichness) %>%
  dplyr::mutate(HostOrder = factor(HostOrder, levels=HostOrder, ordered=TRUE))

# visualise the residuals with a diverging colour scale
vr_zoo %>%
  ggplot() +
  geom_vline(xintercept=0, lty=2) +
  geom_segment(aes(y=HostOrder, x=0, xend=ResidualRichness), size=0.7, color="grey40") +
  geom_point(aes(ResidualRichness, HostOrder, fill=ResidualRichness),
             color="grey20", pch=21, size=7) +
  theme_classic() +
  guides(fill="none") +
  scale_fill_gradientn(colors=rev(pals::brewer.rdylbu(100))) +
  xlab("") +
  ylab("Mammal Order") +
  xlab("Residual viral richness\n(negative = relatively fewer, positive = relatively more)") +
  ggtitle("Do bats carry relatively more zoonotic viruses
          than expected, given their overall viral diversity?") +
  theme(axis.text = element_text(size=11),
        axis.title = element_text(size=12),
        plot.title = element_text(size=14, hjust=0.5))
```



- **Q6.** If we set aside for now the issues with the model, what story is this visualisation telling us? Is it biologically or ecologically realistic? What message would this give to the wider public or decision-makers concerned about the role of bats in disease risk, and what might be missing from this story?

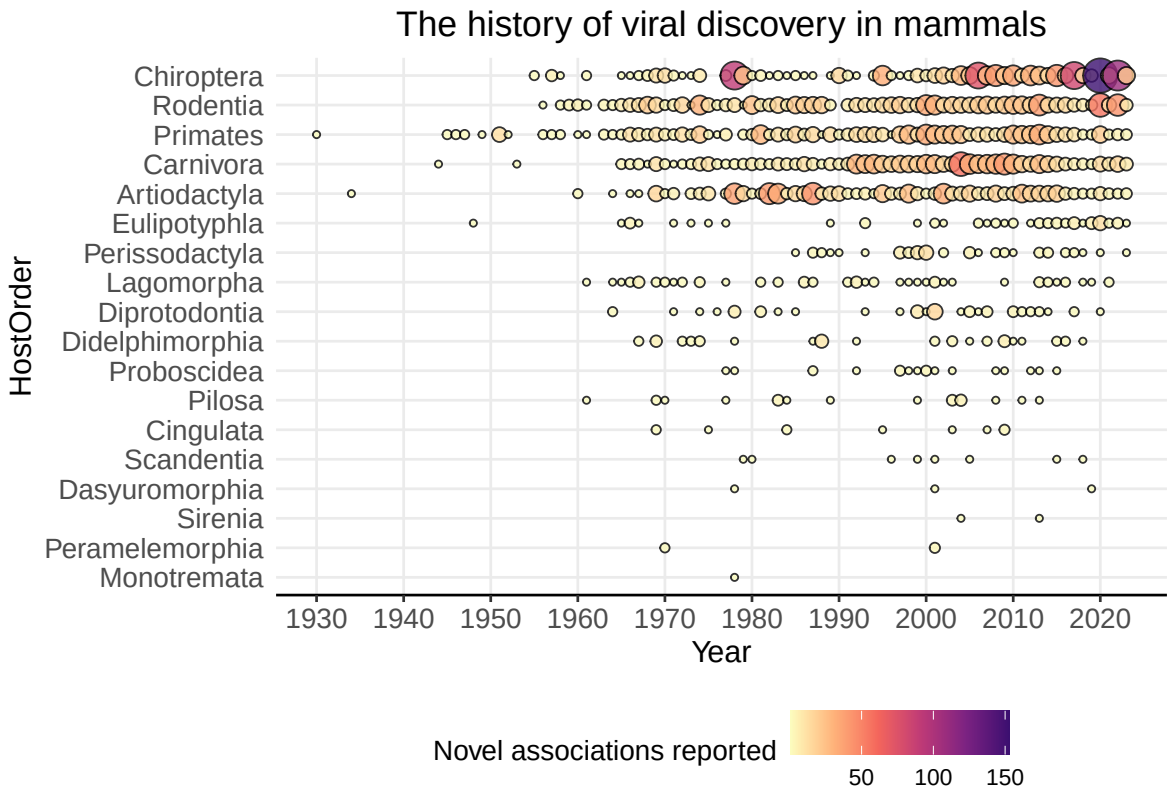
Bats and the discovery of high-consequence zoonoses

```
# calculate the number of new associations documented in each year
btc = vir %>%
  #dplyr::filter(VirusGenus == "betacoronavirus") %>%
  dplyr::group_by(HostOrder, YearFirstRecord) %>%
  dplyr::summarise(Discovered = length(YearFirstRecord)) %>%
  dplyr::mutate(HostOrder = Hmisc::capitalize(HostOrder))

# arrange by total
order_order = vir_total %>% dplyr::arrange(ViralRichness)

# visualise
btc %>%
  dplyr::mutate(HostOrder = factor(HostOrder, levels=order_order$HostOrder, ordered=TRUE)) %>%
  ggplot() +
  geom_point(aes(YearFirstRecord, HostOrder, fill=Discovered, size=Discovered), pch=21, alpha=
  theme_bw() +
  scale_fill_viridis_c(option="magma", direction=-1,
                       begin=0.2, name="Novel associations reported") +
  scale_size(range = c(1, 6)) +
```

```
  xlab("Year") + ylab("HostOrder") +
  theme(panel.border = element_blank()) +
  theme(axis.text = element_text(size=11),
    axis.title = element_text(size=12),
    plot.title = element_text(size=14, hjust=0.5)) +
  guides(size="none") +
  ggtitle("The history of viral discovery in mammals") +
  theme(axis.line.x = element_line(),
        axis.ticks.y = element_blank(),
        panel.grid.minor.x = element_blank(),
        legend.position = "bottom") +
  scale_x_continuous(labels=seq(1930, 2020, by=10), breaks=seq(1930, 2020, by=10))
```



- **Q7.** What does this visualisation tell you about our current estimates of viral diversity in mammals and their reliability? What do you think has driven the history of viral discovery efforts across mammals?

- **Q8.** Modify this code to only include the history of viruses in the genus "Betacoronavirus" - this is the genus that contains SARS-CoV-1 and SARS-CoV-2 (COVID-19). What does this tell you about the evolution of knowledge and discovery efforts for this genus of high-consequence viruses?