Virion: host-virus association data

Tad Dallas

Contents

VERENA

The Viral Emergence Research Initiative (VERENA) is a global consortium. Our goal is to curate the largest ecosystem of open data in viral ecology, and build tools to help predict which viruses could infect humans, which animals host them, and where they could someday emerge.

VIRION

With over 3 million records, the Global Virome in One Network (VIRION) database is a living encyclopedia of vertebrate viruses - including the ones that pose the greatest threats to human health. Data like these pave the way for a new era of predictive science, and form the backbone for a broader data ecosystem we're building for animal disease surveillance.

Exploring the Virion data

Download the Virion data from the link below:

> https://github.com/viralemergence/virion/tree/main/Virion

There are many different bits of information, including detection information (how was the host-virus association quantified?), taxonomic information on hosts and viruses, and the host-virus association data. Let's load the entire dataset and get a better idea of the scope and nature of the data

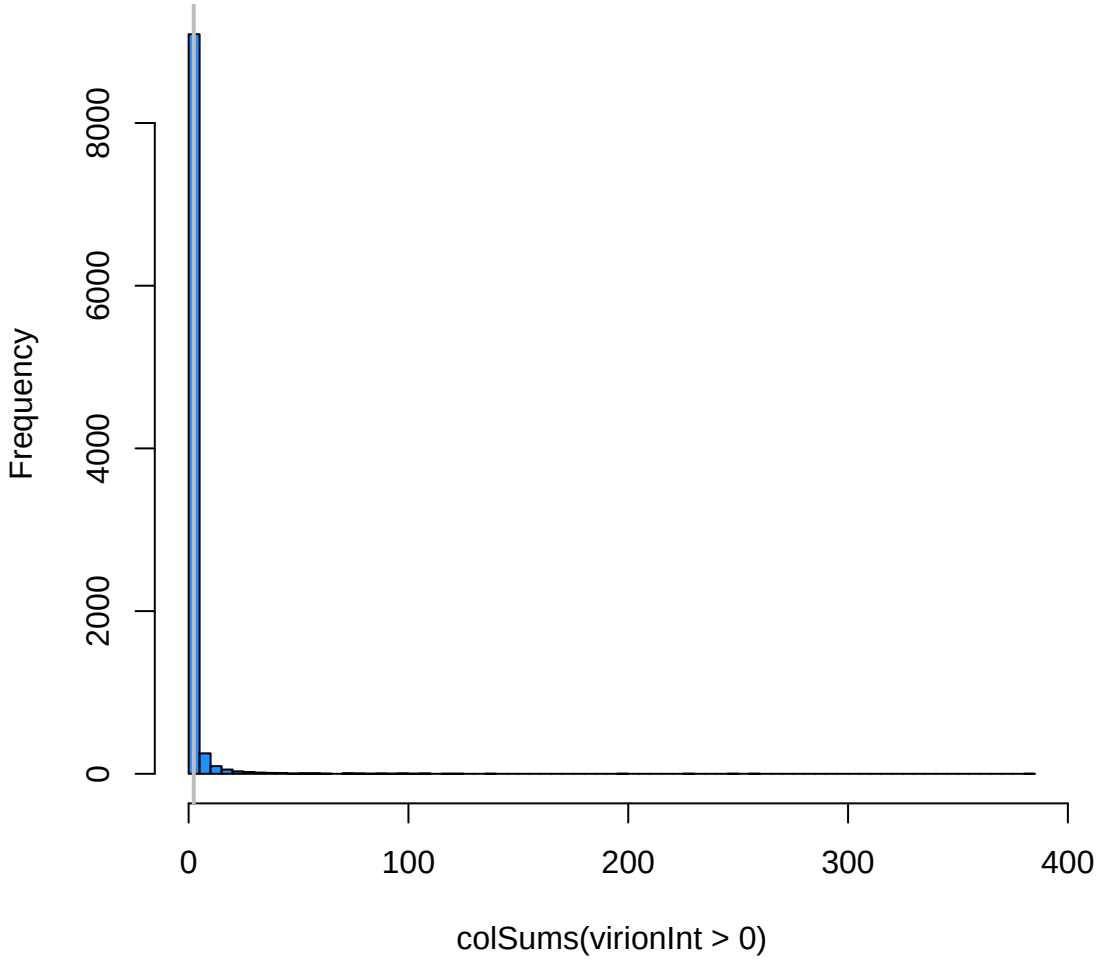
```
virion <- read.delim('data/Virion.csv')


# make an interaction matrix
virionInt <- table(virion$Host, virion$Virus)[,-1]

dim(virionInt)
```

```
## [1] 3643 9628
```

```r
# how host-specific are viruses?
hist(colSums(virionInt>0),
  col='dodgerblue', breaks=100,
  main='')
abline(v=mean(colSums(virionInt>0)), lwd=2, col='grey')
```




```r
# on average, viruses infect around 2 species
mean(colSums(virionInt>0))
```

```
## [1] 2.347736
```

```r
# but this one infects over 4500 species
which.max(colSums(virionInt>0))
```

```
## influenza a virus
##              4724
```

This is the sort of exploratory data analysis that researchers do to try to understand the data. It can also be used to generate research questions. Why is influenza a virus so common in the data relative to the majority of other pathogens? How do patterns of host-virus associations change across different host groups (i.e., are there some pathogens that only infect certain groups of host species?). These are the types of questions that could be explored here, but you are not limited to this. Some other fun ideas for exploring these data given below.

Project ideas

- Create a compelling visualization to share with friends to showcase some aspect of the data. e.g., take images of different host species from the internet and overlay information on how many viruses they have.
- More web-design savvy folks can think about how to serve these data in a web-friendly way or create interactive visualizations using different javascript libraries (e.g., d3.js)
- More analytical folks can consider what questions they could ask within the time window. e.g., what aspects of host species are related (or even predictive) of the number of viruses that infect them?
- What species share the most pathogens with humans? Create a networked image of the relationships between host species in terms of which viral pathogens they share (hint: the R package igraph might be helpful for starting to work with these types of structures)
- Select a particular host species (perhaps one of your favorite animals) and see the types of viruses that can infect it. Create a leaflet/webpage/artwork using this information (e.g., https://www.cdc.gov/brucellosis/index.html)
- Develop a tool to try and identify or model potential associations between hosts and viruses that are not currently known, or perhaps ones that shouldn't even exist given what the data.
- Do host taxa which are represented by more species in the data also tend to have more viruses per species on average? What might that suggest?
- Focus on a specific geographic area. For instance, what viruses might Mike the Tiger get? He is a tiger, but tigers might share viruses with some local Baton Rouge species, leading to Mike potentially becoming infected.
- Scenario analysis: you've just discovered you are infected by some virus (could be anything in the data). Given the name of the virus, identify the potential other host species that might have given you the virus (assuming direct transmission).
- Develop teaching materials aimed at communicating the ecology of infectious disease, using Virion as a starting point.
- ... plenty of great ideas, including yours. Best of luck, and feel free to bounce ideas off of the organizers and helpers.

Other relevant data sources to potentially bring in

- GBIF: Global Biodiversity Information Facility = a global database of species occurrence records. Could be used to start to understand the spatial distribution of these host species
- Host trait data. R package `EcologicalTraitData` has some relevant trait data built into it, such as the PanTHERIA data (https://rdrr.io/github/EcologicalTraitData/traitdataform/man/pantheria.html).
- GIDEON: a database of human infectious disease occurrences, broken down by pathogen and year. Accessible through R (https://bitbucket.org/gideononline/gideon-api-r/src/master/?mc_cid=97eba0ef6e&mc_eid=9a1a474bc0).
-