How many insect viruses are there?

Colin J. Carlson

12/18/2020

The core philosophy

This is a back-of-the-envelope estimate of the global diversity of insect-infective viruses. It uses an approach developed by Carlson *et al.* (2019) *Nature Ecology and Evolution*, which was used at the time to estimate global viral diversity in mammals. That approach estimates total symbiont diversity by (1) generating a bipartite scaling curve from a host-symbiont association dataset, and extrapolating to a higher total number of hosts; then (2) using a more complete metagenomic dataset on one specific host species to correct for undersampling in the association data. For example, in that study, there are two species - a bat and a monkey - that have been fully inventoried. After using the HP3 dataset to estimate total mammal virus diversity, that number gets multiplied by 1 / (the proportion of monkey/bat viruses in HP3 out of all of their known viruses).

We're going to try the same thing here. On Twitter, Eddie Holmes said that *Drosophila melanogaster* probably has the best-sampled insect virome. We're going to go to GenBank and use that to grab every named *D. melanogaster* virus. That'll be our "complete" species. Then, we're going to use the `insectPathogen` package that Tad shared to generate our edgelist of insects and viruses, and use the `codependent` package I made in 2019 to estimate insect virus diversity. We'll use an estimate of ~6 million insects on Earth, based on:

Larsen, Brendan B., et al. "Inordinate fondness multiplied and redistributed: the number of species on earth and the new pie of life." *The Quarterly Review of Biology* 92.3 (2017): 229-265.

Some install

If you need these packages, you can install both from Github. However, because `insectDisease` is still private, you'll need to clone it and install locally.

```
devtools::install_github('viralemergence/insectDisease')
devtools::install_github('cjcarlson/codependent')
```

You will also need to **unzip the file called GenBank_as_Edgelist.zip** before you go any further (yes, the file is too big for GitHub, no, I don't feel like solving that problem a responsible way in the first version of this).

Some loading

```
library(codependent)
library(InsectDisease)
library(tidyverse)
```

How many viruses does *Drosophila melanogaster* have?

```
setwd("~/Github/drosophily")

gb <- read_csv(’GenBank_as_Edgelist.csv’)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Host = col_character(),
##   Species = col_character()
## )
```

```
gb %>%
  filter(Host == ’Drosophila melanogaster’) %>%
  select(Host, Species) %>%
  unique() -> dm

dm %>%
  nrow()
```

```
## [1] 30
```

```
dm %>% pull(Species)
```

```
##  [1] "Esparto virus"
##  [2] "Tomelloso virus"
##  [3] "Drosophila melanogaster sigmavirus"
##  [4] "Kallithea virus"
##  [5] "Drosophila melanogaster totivirus SW-2009a"
##  [6] "Drosophila A virus"
##  [7] "Nora virus"
##  [8] "Galbut virus"
##  [9] "Chaq virus"
## [10] "Vera virus"
## [11] "Chaq-like virus"
## [12] "Vesanto virus"
## [13] "Drosophila-associated nudivirus"
## [14] "Drosophila-associated filamentous virus"
## [15] "Yalta virus"
## [16] "Drosophila C virus"
## [17] "Mauternbach virus"
## [18] "La Jolla virus"
## [19] "Dansoman virus"
## [20] "Motts Mill virus"
## [21] "Craigies Hill virus"
## [22] "Newfield virus"
## [23] "Torrey Pines virus"
## [24] "Viltain virus"
## [25] "Drosophila virus JTM-2015"
## [26] "Drosophila reovirus"
## [27] "Thika virus"
## [28] "Flock House virus"
## [29] "Drosophila melanogaster birnavirus SW-2009a"
## [30] "Drosophila melanogaster tetravirus SW-2009a"
```

There are about thirty viruses in the NCBI taxonomy that infect *D.m.*

Bringing in the GenBank data

Let's grab the insect virus data, and check *D. melanogaster*:

```
id <- InsectDisease::viruses

id %>%
  filter(Host == 'Drosophila melanogaster') %>%
  pull(Virus)
```

```
## [1] Drosophila  sigma virus Drosophila  C virus     Retrovirus  (RTV)
## [4] reovirus-like particle
## 267 Levels:  ... Wiseana cervinata IV
```

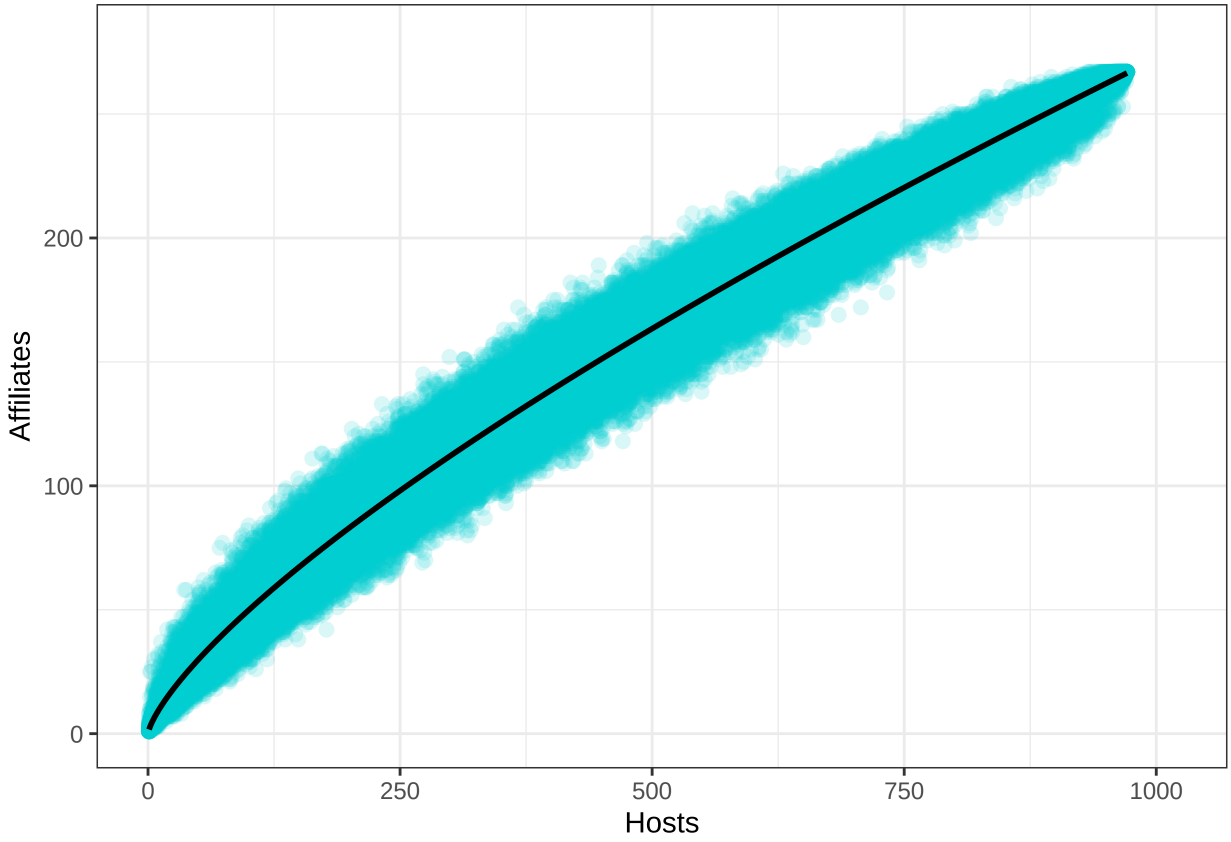
Drosophila C virus and sigma virus are named, both in GenBank. Retrovirus (RTV) and reovirus-like particle are not; reovirus is probably equivalent to Drosophila reovirus, though. So let's say there are 3 viruses here compared to 30 in GenBank - that means we'll use a multiplication factor of 10.

Building the edgelist and curve

Let's get a unique edgelist of hosts and vectors, and try building the scaling curve:

```
id %>%
  select(Host, Virus) %>%
  unique() -> assn

b <- binera(assn, iter = 100, plots = TRUE)
```



```
b
```

```
## Nonlinear regression model
##   model: n.par ~ b * n.host^z
##    data: cu
##      b      z
## 1.6739 0.7372
##  residual sum-of-squares: 8894122
##
## Number of iterations to convergence: 8
## Achieved convergence tolerance: 2.231e-09
```

Now let’s try extrapolating:

```
c <- copredict(assn, iter = 100, n.indep = 6000000)
c
```

```
## [[1]]
##   mean.b  lowerCI  upperCI
## 166785.5 163560.2 170068.8
##
## [[2]]
##         b         z
## 1.6682706 0.7376483
##
```

```
## [[3]]
##       2.5 %    97.5 %
## b 1.6587135 1.6778276
## z 0.7367653 0.7385314
```

And finally, let's multiply by that correction factor:

```
c[[1]][1]*10
```

```
##  mean.b
## 1667855
```

If there are about 6 million insect species on Earth, there should be proportionally at least ~1.66 million viruses in insects.

Frequently asked questions

Q: Carroll *et al.* (2018) *Science* famously estimated there are 1.67 million viruses in mammals and waterfowl, which is a number you corrected in the 2019 *Nature Ecology and Evolution* paper down to 40-60,000.

A: Yes, I was there. That's not a question, really.

Q: That 1.67 million viruses number is absurdly, weirdly close to 1.67 million insect viruses.

A: Yeah, it is.

Q: Does that mean anything?

A: Absolutely not, but it's going to make for a ridiculous sentence in the abstract.