GenBank Data Analysis for International Equity in Scientific Research

Will Devitt

**Code Process for Top Virus Discovery & Health Event Discovery:**

For this analysis, I took two sets of virus research activity – submissions to the network, and publications. Each excel file for each of these datasets represents one year of activity – with each row in each file representing a virus and the corresponding number of submissions or publications relating to that specific virus. The folders holding this data in my GitHub are sub\_top\_virus\_taxname\_redo and pub\_top\_virus\_taxname\_redo. **I used a Python dictionary to iterate through the entire dataset, counting the total number of occurrences of submissions and publications for specific viruses throughout the years (the results of which can be found in the GitHub, under top\_names.PNG, top\_sub\_names.PNG, etc.)** This is the function def discover\_top\_names in tax\_virus\_analysis.py. I further used the functions sub/pub\_top\_name\_binder to return the values for those overall top names for each year. **I entered the data returned from each year by these functions into .csv files which you may also find on my GitHub (top\_cts\_total.csv, sub\_cts.csv, and pub\_cts.csv.** The latter two represent only the submission and publication networks, the former represents the total network. Essentially, we are exploring the overall trends of the most popular viruses in the network for each year of activity. I then manually entered these yearly results into Excel sheets, using this data to create the illustrative Flourish plots one may find in the ReadMe. With this visualization of the trends, I was able to discover increases in network activity in certain years corresponding to global health events.

Text

Description automatically generatedText

Description automatically generatedText

Description automatically generated

Clockwise from top left: Top viruses from both networks, top viruses in the publication network, top viruses in the submission network

**Public Health Events Discovered:**

The results of this analysis correlated with a few global virus events. The top 10 overall virus names in the network in our timeframe of 1992 -2018 were HIV, Dengue virus 2, Dengue virus 1, Dengue virus 3, West Nile virus, Hepacivirus C, Hepatitis C, Hepatitis B, Zika, and Hepatitis E. **Within this, the most prominent outbreaks of these viruses occurred with a worldwide Dengue outbreak from about 2005-2011, a West Nile virus outbreak from 2008-2012, and a worldwide outbreak of Zika virus from 2015-2017.** There was a substantial spike in Hepatitis C network activity from 2006-2008, although this appears to not be correlated with an outbreak, thus necessitating further investigation. During the Dengue outbreak, the most effected nations were India, Pakistan, Brazil, and Singapore. **You may find results under the ‘results’ folder on GitHub. You will find plots within .csv files. You may also find full exploration of these outbreaks and the countries affected by them under “recognizing\_outbreaks\_in\_network.docx.”**

**Questions & Exploration Process:**

* Is funding and research equitable? (ie does international attention and then funding only come when high profile / western countries begin to suffer?) Who is doing research? Is it global? Who benefits from research? Who suffers under outbreaks?
* How does the network react when an outbreak occurs? What factors lead to different levels of change?
* Does collaboration increase or decrease during global health events?
* How reliable is network activity in reflecting outbreaks?
* When outbreaks occur, does research on certain major pervasive viruses (HIV) significantly decrease when other times it does not? (See Dengue)