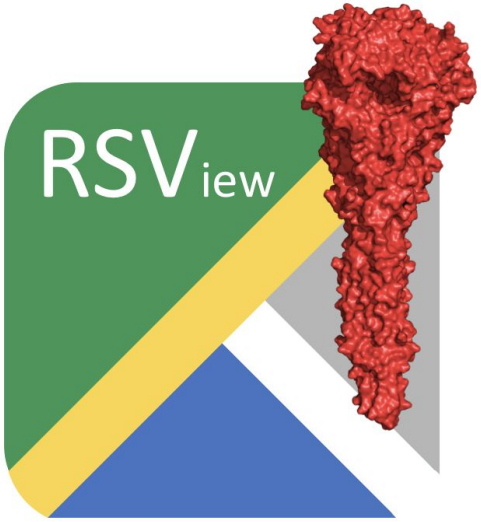


RSView

Jilliane, Kate, Katie

Respiratory syncytial virus (RSV) is an important cause of viral infections

In healthy adults and children, RSV causes cold-like symptoms lasting 1-2 wks

RSV can cause severe infections:

Infants

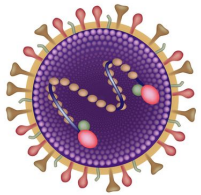
Immunocompromised

Elderly

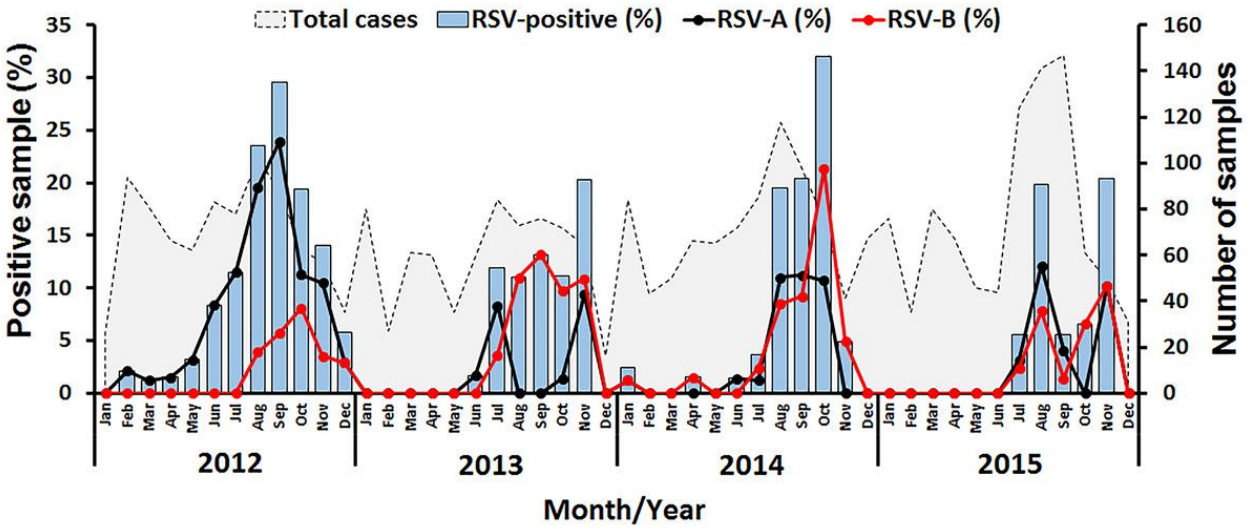
RSV causes high global burden of disease in children under 5

RSV-associated acute respiratory infection causes more than **3 million hospital admissions** and **~100,000 deaths** in children less than 5 each year.

RSV-immunity is short-lived and individuals can be re-infected throughout their lives.



RSV is a genetically diverse virus that circulates seasonally



There are 2 main subtypes - A and B - and several genotypes characterized by the sequence of the attachment glycoprotein, G.

Different viral subtypes and genotypes are thought to replace each other over time.

Evidence for this is generally based on isolated studies (such as Thailand, above).

Thongpan, I. *PeerJ* 2017

Goals of project

1. Investigate global patterns of RSV cycling using all RSV G protein sequences deposited in GenBank

    Can we see evidence for subtype cycling and genotype replacement at global and country scales?

Goals of project

1. Investigate global patterns of RSV cycling using all RSV G protein sequences deposited in GenBank

   Can we see evidence for subtype cycling and genotype replacement at global and country scales?

2. Assess if subtype and/or genotype affects disease severity as determined by deaths due to acute respiratory infection in children <5.

   Does subtype/genotype *cycling* affect disease severity? Are seasons immediately following a switch in the dominant strain worse?

Goals of project

1. Investigate global patterns of RSV cycling using all RSV G protein sequences deposited in GenBank

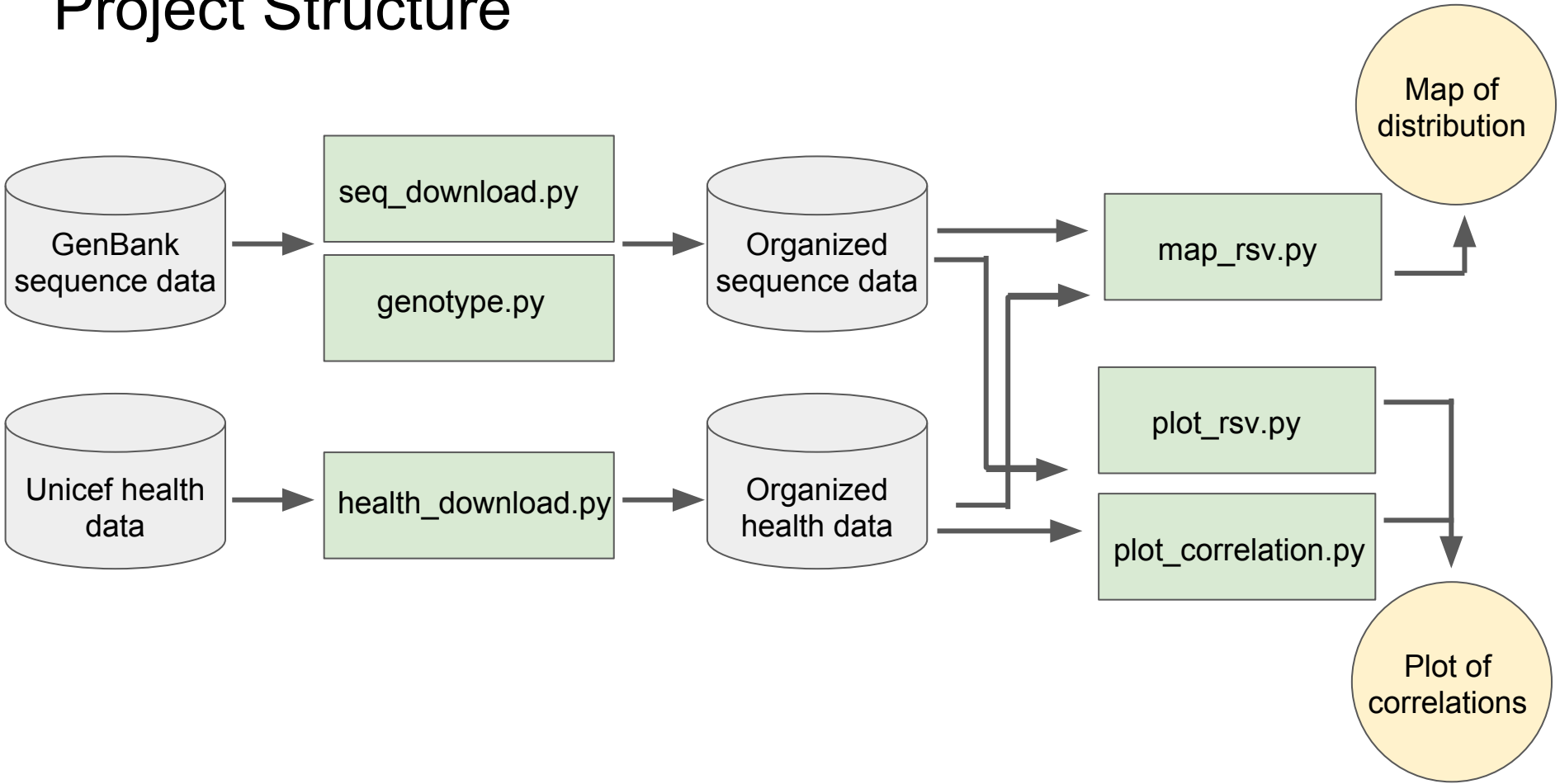
   Can we see evidence for subtype cycling and genotype replacement at global and country scales?

2. Assess if subtype and/or genotype affects disease severity as determined by deaths due to acute respiratory infection in children <5.

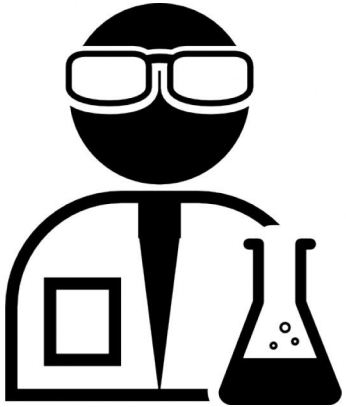
   Does subtype/genotype *cycling* affect disease severity? Are seasons immediately following a switch in the dominant strain worse?

3. Write a program that could be easily expanded to examine other viral sequences downloaded from GenBank.

Project Structure



User profile



Physicians and scientists studying treating patients with or studying RSV

Need to be able to clone git repos and have some comfort running python packages. Need not be able to code.

Future work could make this transferrable to other virus datasets given ability to code in python

Workflow

1. Download and organize sequence data
    a. Download sequences and metadata from GenBank
    b. Align sequences and call missing genotypes

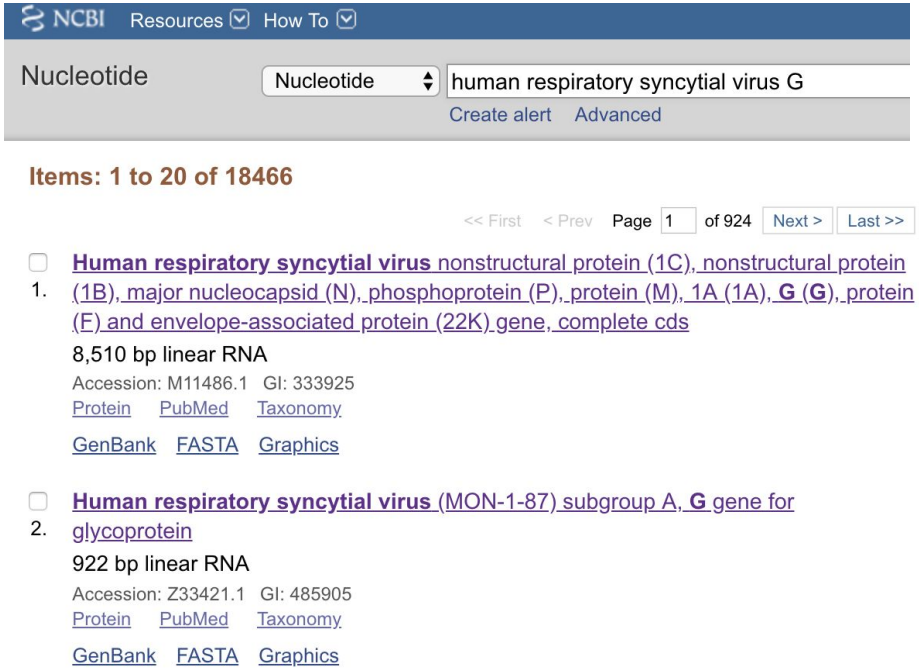
2. Download and organize health data

3. Map global distribution of RSV types interactively
    a. Allow user to look at dominant subtypes/genotypes in each country over time.
    b. Integrate with deaths from acute respiratory tract infections in children under 5.

4. Analyze and plot correlations between virus type and health outcomes
    a. More in-depth analysis of effects of subtype on disease severity

Data sources

GenBank:



Unicef:

Estimates of child cause of death, Acute Respiratory Infection 2018

Last update: February 2018

Stats on neonatal, post-natal, and under 5 deaths from acute respiratory infection.

Data from 2000-2016 for 194 countries

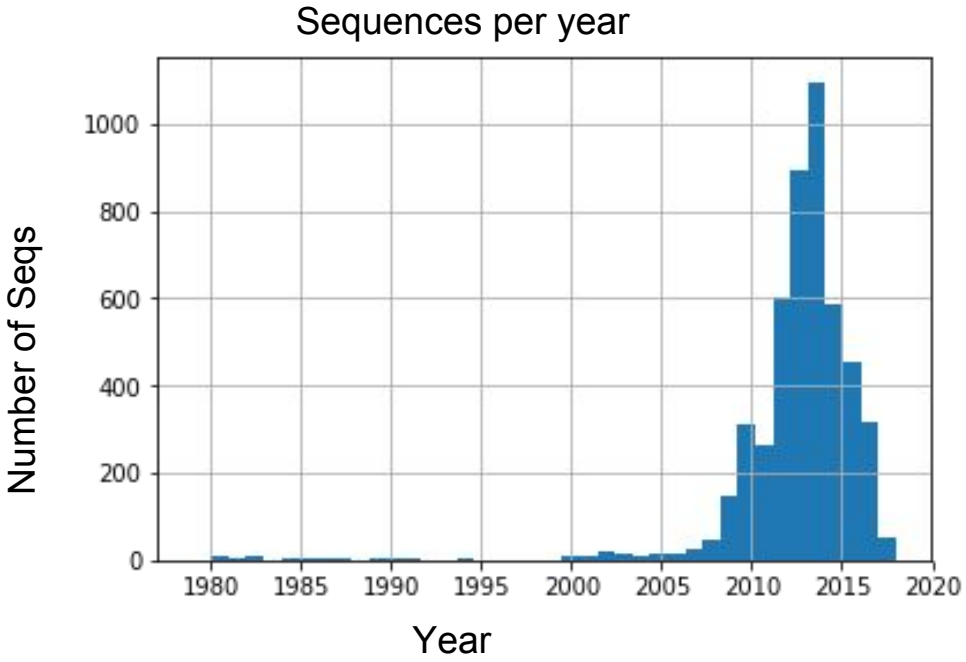
GenBank data is deposited in various forms from many scientists

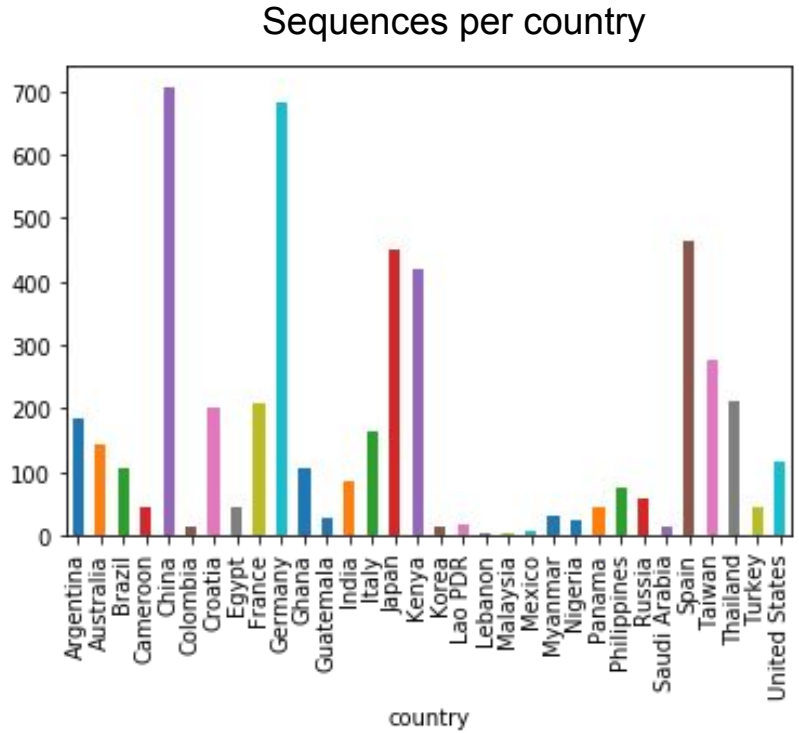
```
FEATURES             Location/Qualifiers
     source          1..15277
                     /organism="Human orthopneumovirus"
                     /mol_type="viral cRNA"
                     /isolate="B05"
                     /isolation_source="tracheal rinse"
                     /host="Homo sapiens"
                     /db_xref="taxon:11250"
                     /country="United Kingdom"
                     /collection_date="06-Jan-2005"
                     /note="subgroup: B; genotype: BA"
     gene            46..577
                     /gene="NS1"
     mRNA            46..577
```

Whole genome vs. full G vs. partial G

Different degrees of subtype and genotype information

Sequences are not distributed evenly





Sequences are only a small sample of circulating RSV - not surveillance

Lack of good RSV surveillance sequencing limits the power of our analyses

Bio.Entrez allows for downloading GenBank data, but messily

Output from Bio.Entrez is essentially lots and lots of nested dictionaries

Lots of string parsing to extract necessary information (`seq_download.py`)

After string parsing, ~90% sequences had subtypes and ~33% had genotypes

Used sequence alignments of all G protein sequences and sequences with known genotypes to (very hackily) call genotypes for another 33% of seqs (`genotype.py`)

Data is incomplete, but useable despite caveats

>18,000 RSV G sequences to download

~90% of sequences had subtype information from GenBank

~33% had genotype information from GenBank

Need to download and clean data

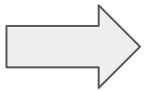
Call additional genotypes

Download and genotype demo.

Health Data Processing and Integrating Datasets

The data was mostly clean but we had to:

- Re-format numbers to create summary dataset
- Convert a variety of ‘country’ input styles to integrate with the genotype dataset



country converter

iso3	Country/area	year	nnd	pnd	neo9	post9	ufive9	rneo9	rpost9	rufive9	fneo9	fpost9	fufive9
	Global	2000	4,003,141	6,047,778	307,725	1,466,299	1,774,025	2	11	14	8%	24%	18%
	Global	2001	3,902,504	5,851,905	294,644	1,406,326	1,700,971	2	11	13	8%	24%	17%
	Global	2002	3,800,435	5,547,240	282,158	1,348,396	1,630,553	2	10	12	7%	24%	17%
	Global	2003	3,699,679	5,421,280	270,859	1,291,699	1,562,558	2	10	12	7%	24%	17%
	Global	2004	3,601,737	5,163,989	259,239	1,241,529	1,500,768	2	9	11	7%	24%	17%
	Global	2005	3,505,159	4,928,527	246,933	1,192,271	1,439,204	2	9	11	7%	24%	17%
	Global	2006	3,415,328	4,660,282	235,214	1,155,132	1,390,346	2	9	10	7%	25%	17%
	Global	2007	3,323,929	4,409,248	223,953	1,107,225	1,331,178	2	8	10	7%	25%	17%
	Global	2008	3,236,930	4,265,929	214,501	1,056,833	1,271,334	2	8	9	7%	25%	17%
	Global	2009	3,153,926	4,115,455	205,889	1,013,537	1,219,426	1	7	9	7%	25%	17%
	Global	2010	3,070,191	3,943,044	197,625	972,561	1,170,186	1	7	8	6%	25%	17%

Mapping demos

Basic maps:

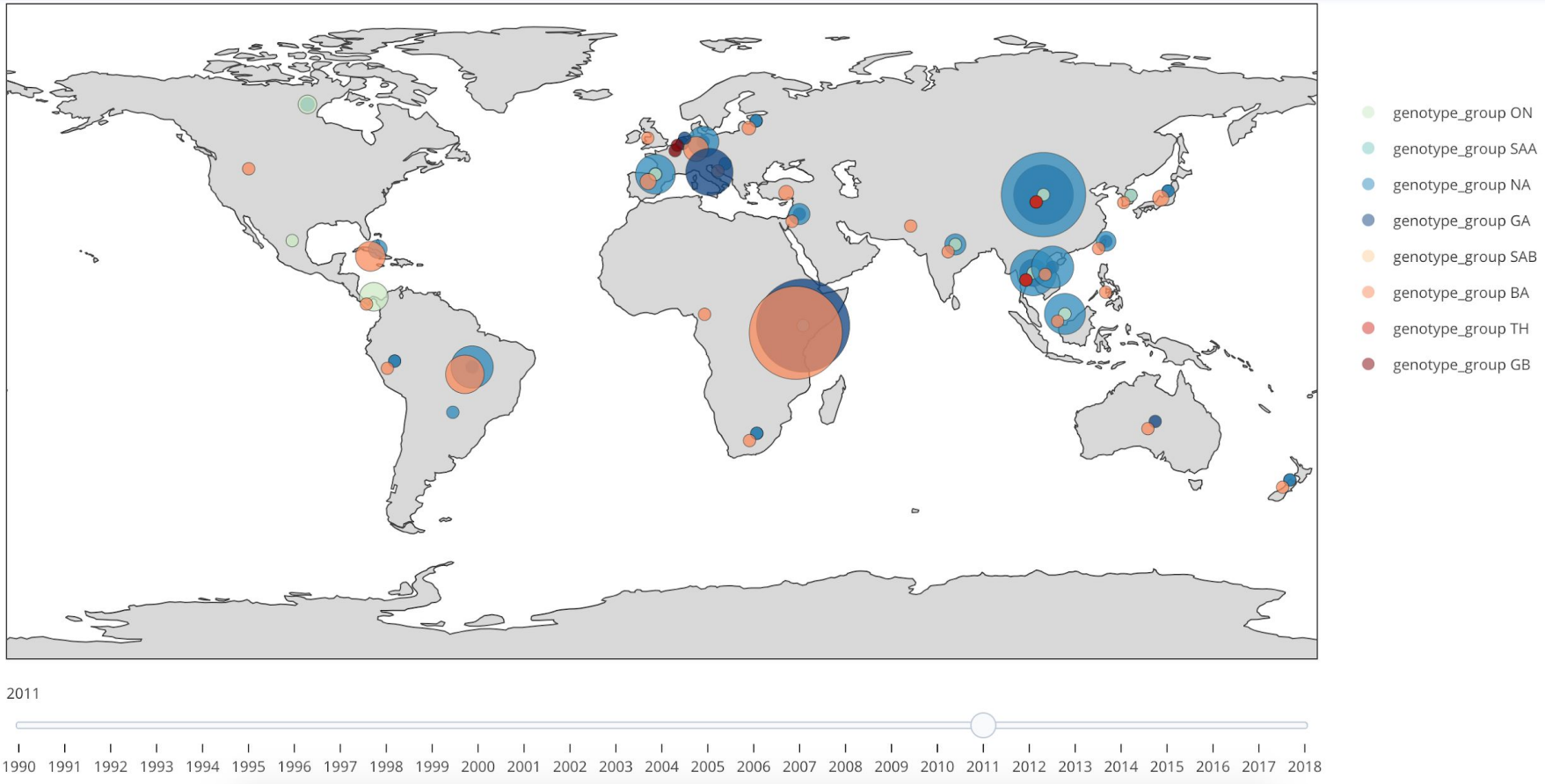
```
$ map_rsv.py subtype data

$ map_rsv.py genotype data
```

Optional arguments:

```
$ map_rsv.py subtype data --years all

$ map_rsv.py genotype data --genotype_level all
```

Analyze global distribution of RSV genotypes

```
$ map_rsv.py genotype data
```



Biological observations from maps

```
$ map_rsv.py subtype data
```

1. Subtype A/B replacement on regional, not global, scale
    a. Previous evidence from isolated studies only

2. No clear evidence of correlation between ARI and dominant subtype

3. Inferences limited by sampling
    a. Support for: Ratio of A to B for a given country
    b. No support for: Absolute number of A vs B between countries

Plotting demos

Health data:

```
$ plot_rsv.py all rufive9          (optional: --highlight_country)

$ plot_rsv.py country rufive9      (optional: --country)
```

Health vs Subtype data:

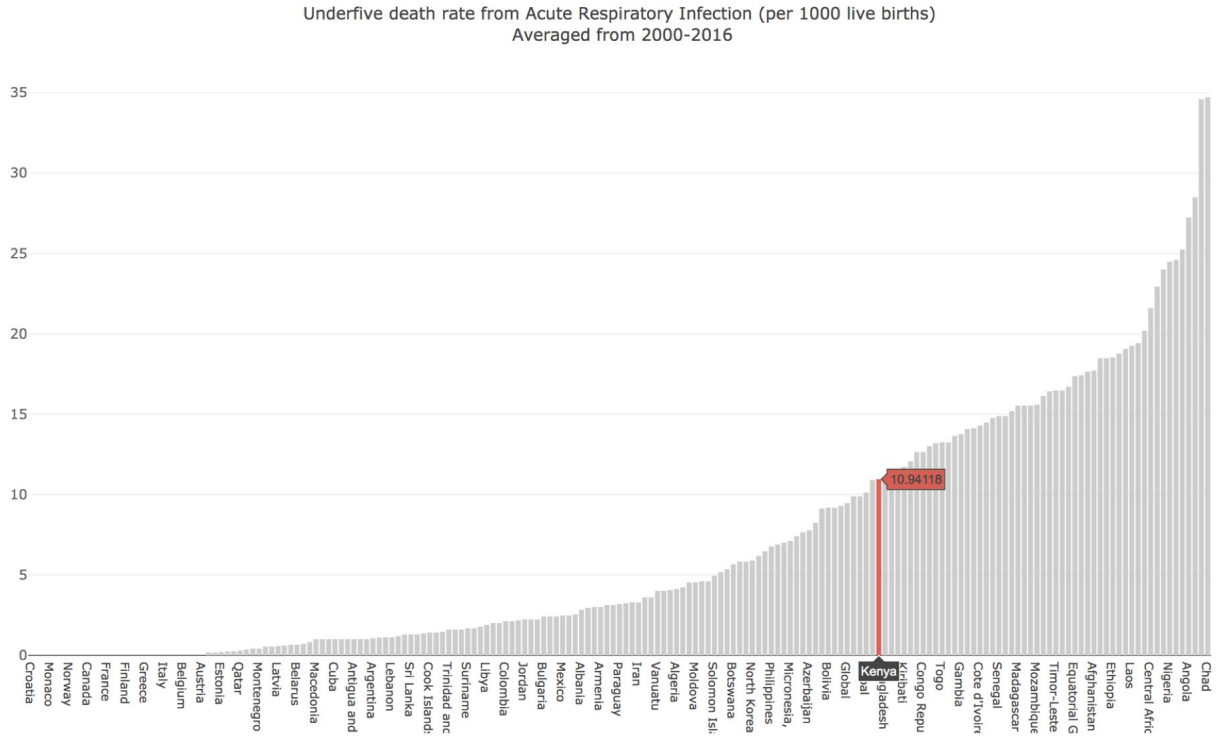
```
$ plot_correlation.py all rufive9

$ plot_correlation.py year rufive9
```

Health Data Variables

```
Specify which category of data to plot:
  nnd : Total Neonatal Deaths
  pnd : Total Post-Neonatal Deaths
  neo9 : Neonatal deaths due to Acute Respiratory Infection
  post9 : Post-neonatal deaths due to Acute Respiratory Infection
  ufive9 : Underfive deaths due to Acute Respiratory Infection
  rneo9 : Neonatal death rate from Acute Respiratory Infection (per 1000 live births)
  rpost9 : Post-neonatal death rate from Acute Respiratory Infection (per 1000 live births)
  rufive9 : Underfive death rate from Acute Respiratory Infection (per 1000 live births)
 fneo9: Percent Neonatal deaths due to Acute Respiratory Infection
 fpost9: Percent Post-neonatal deaths due to Acute Respiratory Infection
 fufive9: Percent Underfive deaths due to Acute Respiratory Infection
```

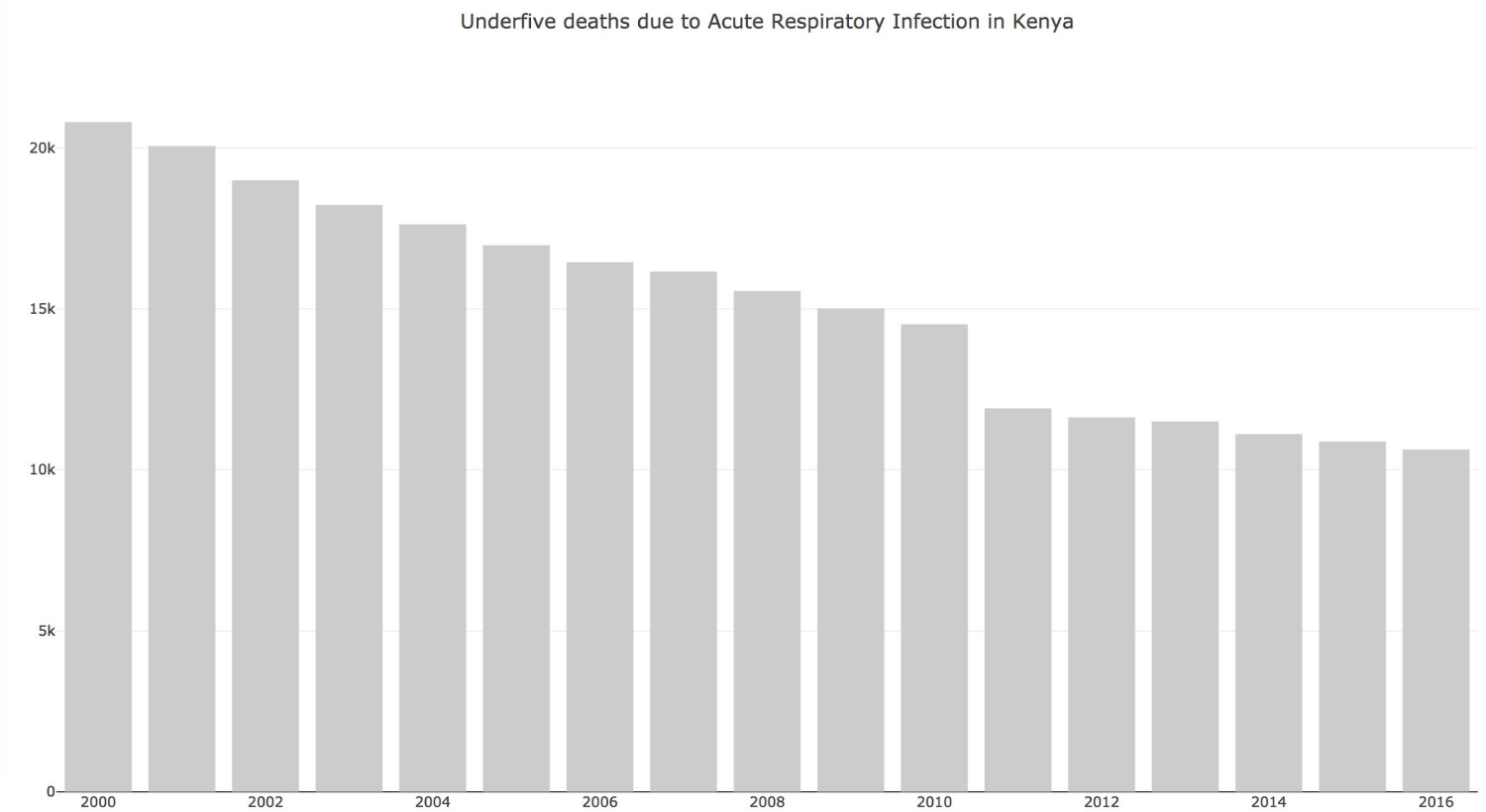
Analyze health impact of acute respiratory infections around the world

```
$ plot_rsv.py all rufive9          (optional: --highlight_country)
```



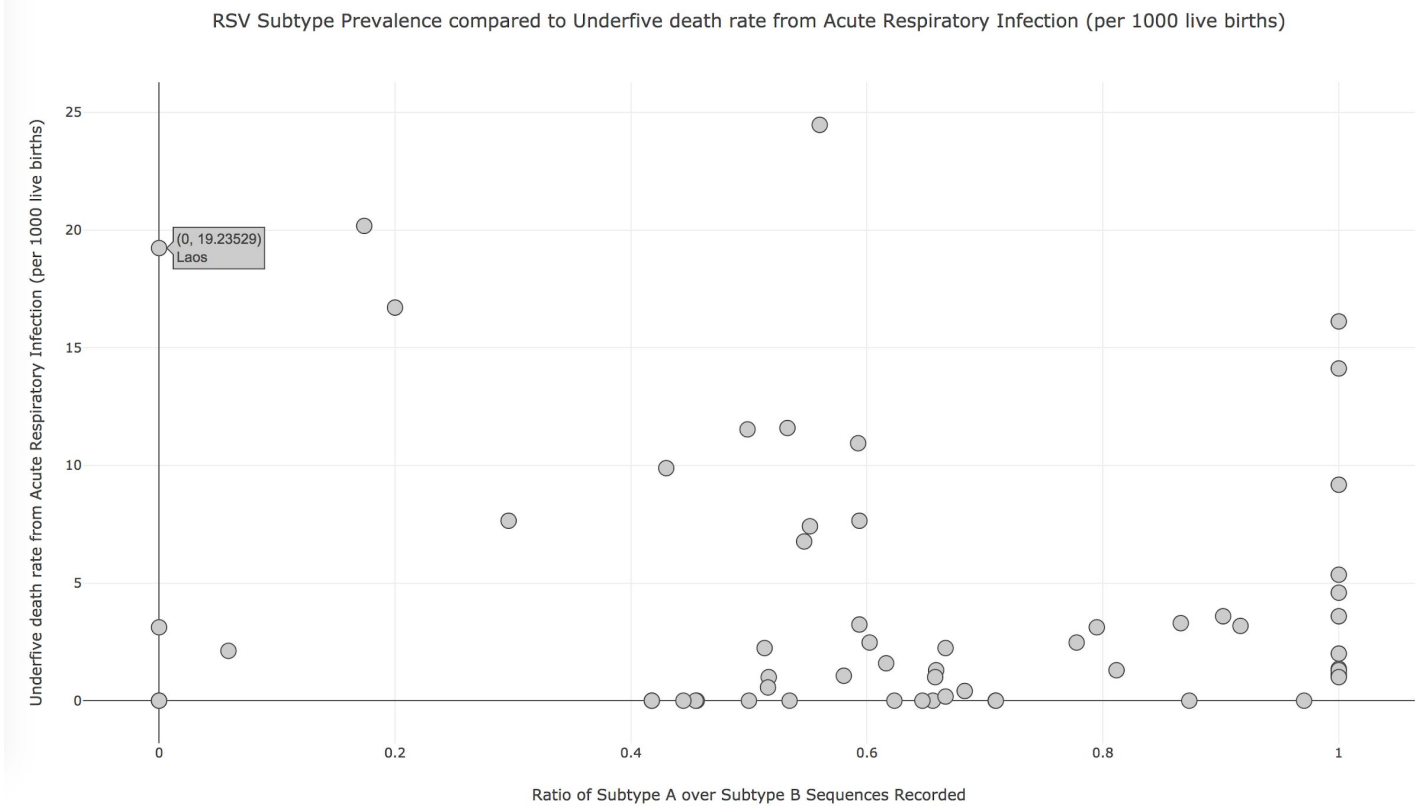
Analyze health impact trends in a specific country over time

```
$ plot_rsv.py country rufive9     (optional: --country)
```



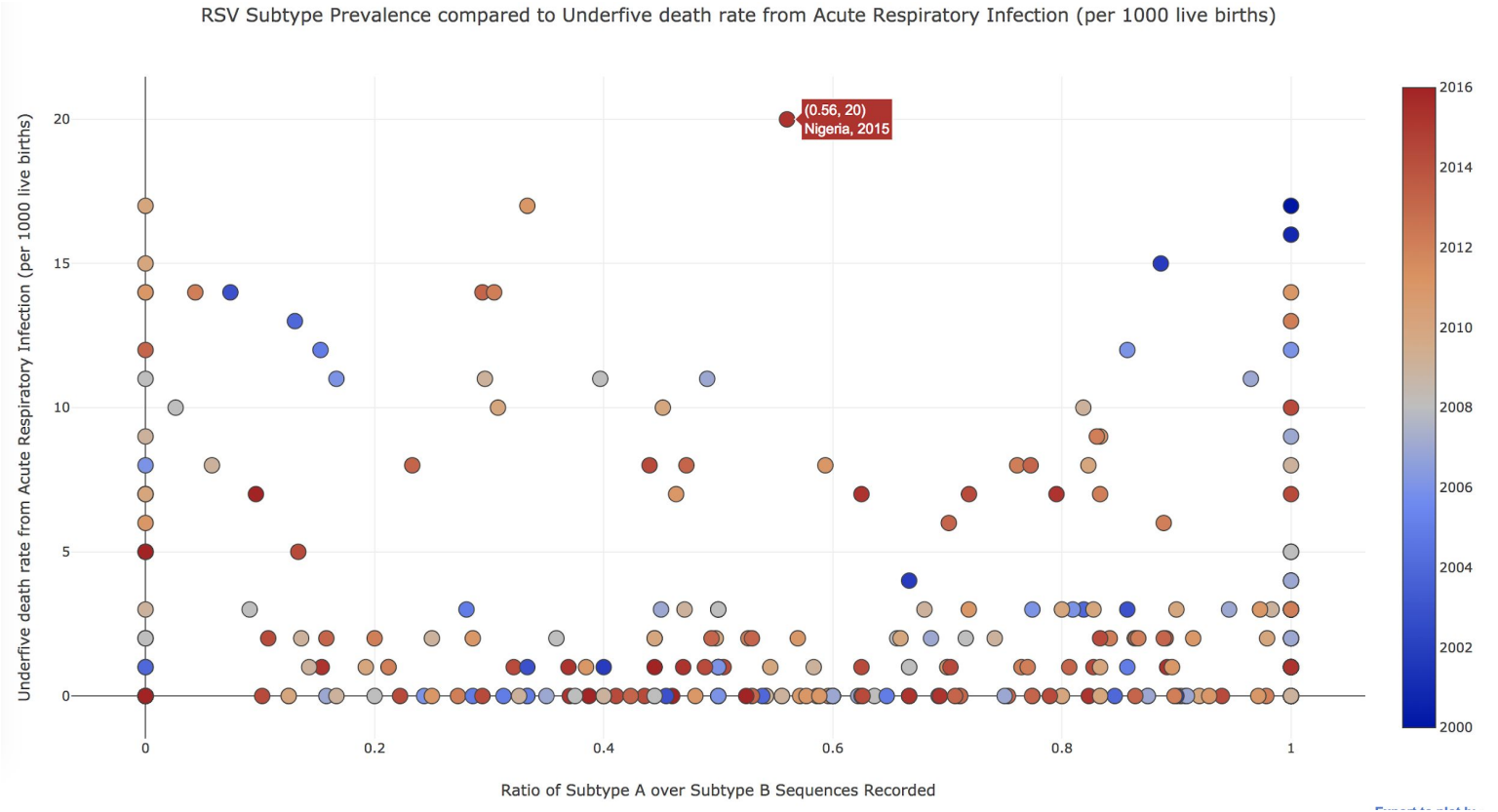
Analyze the correlation between RSV subtype prevalence and health impact

```
$ plot_correlation.py all rufive9
```



Analyze the correlation between RSV subtype prevalence and health impact

```
$ plot_correlation.py year rufive9
```



Future work (in the next week)

1. Update and flesh out documentation
    a. README.md
    b. Installation instructions
    c. Script comments
2. Make code PEP-8 compliant
3. Example usage iPython notebook
4. Finish testing
5. Add trendline and compute R squared for subtype/health correlations
6. Beautify display of health data on map
    a. Show data visually rather than in text

Future work (post-quarter)

1. Improve genotype classification
2. Integrate more specific health outcome data
3. Investigate correlation between subtype cycling and disease severity
4. Address sampling issues
5. Expand RSView to other viruses