INSTITUTE FOR DEFENSE ANALYSES

Understanding Viral Population Genomics: Tools and Lessons for Future Pandemics

(Poster)

Katherine I. Fisher-Aylor
Izzy Chaiken
Emily D. Heuring

September 2022

IDA Document NS D-33249

Log: H 22-000397

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305-3086

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

This work was conducted by the Institute for Defense Analyses Central Research Program, project C2277 "SARS-CoV-2 Paper." The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For More Information

Katherine I. Fisher-Aylor, Project Leader
kfisher@ida.org, 703-845-6902

Leonard J. Buckley, Director, Science and Technology Division
lbuckley@ida.org, 703-578-2800

Rigorous Analysis | Trusted Expertise | Service to the Nation

BACKGROUND

The Covid-19 pandemic is the most information-rich pandemic in history.

For the first time, we were able to obtain entire genome sequences of the pathogen in real time during the pandemic.

Although sampling of genomes was spotty in places, we had an unprecedented opportunity to understand principles of viral population genomics on an extremely large scale.

During the first year of the pandemic, we wanted to know:

- **How many cases are there?** Is it possible to use viral genomes to estimate this?
- **How much undersampling is too much?** Many places in the U.S. collected very few SARS-CoV-2 samples. How much can we undersample before we get a skewed picture of the pandemic?

POPULATION GENETICS PRINCIPLES

Mutation numbers increase with population size.

Simulated population

exponential growth; 2 mutations every 2 new sequences

Therefore, it is possible to infer population size from mutation number.

In the field of population genetics, there are several canonical ways to measure mutations in a population (genetic variation).

Theta
number of sites with a difference
~normalized by population size

$$\theta = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

$$\theta = 2N\mu$$

Pi
average pairwise differences

$$\pi = \sum_{ij} x_i x_j \pi_{ij}$$

$$\pi = \frac{sum\ of\ pairwise\ differences}{number\ of\ pairs}$$

S: number of segregating sites (places that have a mutation somewhere in the population)
n: number of samples
μ: mutation rate (for the bounds of the whole region or genome in question)
N: population size
i (left): index of summation (which sample you are on)
i, j (right): two sequences
$x_i x_j$: frequency of two (i^{th} and j^{th}) sequences
π_{ij}: number of differences between the two sequences (i and j)

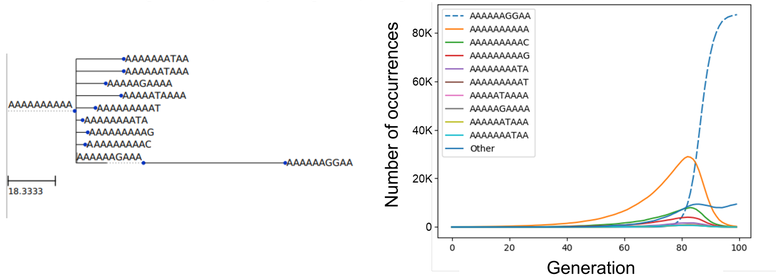
We analyzed the appropriateness of these canonical estimators for measuring SARS-CoV-2 genetic variation during the 2020 Covid-19 pandemic.

SIMDEMIC: SOFTWARE FOR MODELING POPULATIONS OF GENOMES

- **We created this software package to model a growing and mutating population of DNA sequences.**
- **Simdemic allows specification of mutation rate, growth model, genome size, selective pressure, and founder events.**
- We have used Simdemic to create an explicitly known population from which we can infer how real-world methodology affects our conclusions.
  - Assess different estimators of genetic variation
  - Learn the limits of subsampling
- In the future, we hope to use Simdemic to refine predictions of driving mutations, explore the accuracy of different epidemiological growth models, and understand the impact of founder events (as opposed to selective pressure) on a population.

Example simulation

100 generations; extremely small 'genome'; SEIR growth model; one founder event; first half conserved



RESULTS

Theta is an inappropriate estimator of SARS-CoV-2 genetic variation; Pi is appropriate

Simulated data

Simdemic shows theta is the same regardless of number of mutations; pi is not

Mutation rate: 1 every 2 sequences

```
sequences  1024
unique     726
[('AACAAACAAA', 10), ('AAAAAGCAAA', 9), ('AAATACTACA', 7), ('AAAATACAAA', 7), ('AAAAAGCAAT', 6), ('AAAAGGCAAA', 6), ('AAAGAAATAT', 6),
('AAAAAACACC', 4), ('AAATAACACA', 4), ('AAAAACTACA', 4), ('TAAAAACAAA', 4), ('TATATACCAA', 4), ('TAACAGCAAA', 4), ('GTAACACAAA', 4),
('CAAATACAAA', 4), ('AAAAACTAAA', 4), ('GAAAAACAAG', 4), ('GAAAAACGAG', 4), ('AAGAAAAATA', 4), ('AACAACCAAA', 4), ('AAAAATTGAG', 4),
('AAAAAGAGAG', 4), ('AAAGAAAACG', 4), ('AAAAATAACG', 4), ('AAGAATCAAA', 3)]
seq length 10
seg sites  10
theta s.s. 1.3318772805568886
pi         5.415810435707958
```

Mutation rate: 1 every 5 sequences

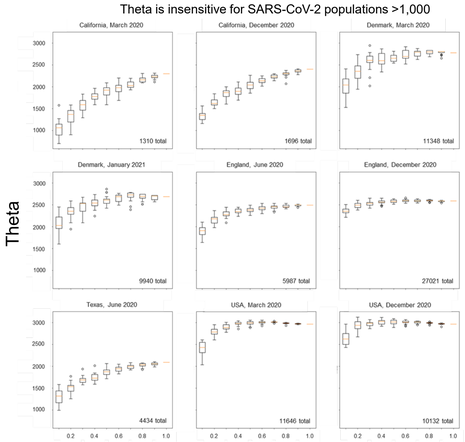
```
sequences  1024
unique     237
[('AAAAAAAAAA', 198), ('AAAAATAAAA', 68), ('AAAAAAAAAG', 36), ('AATAAAAAAA', 27), ('AAAGAAAAAA', 21), ('AAAAAAATAA', 21), ('AAAAAAATA', 18),
('AAAACAAAAA', 17), ('AATAAAAAAG', 16), ('AAAAAAACA', 15), ('GAAAAAAAAA', 15), ('AAAAATAAAA', 15), ('ATAAAAAAAA', 14), ('AAATAAAAAA', 13),
('AAAAAAAAAC', 13), ('AAAAAAAAT', 11), ('AAAAAAACAA', 10), ('AAAGAAAAAA', 10), ('TAAAAAAAAA', 9), ('AAAAGAAAAA', 9), ('CAAAAAAAAA', 8),
('AAATAAAGAA', 8), ('AAAATAAAAA', 8), ('AAAAAGAAAA', 8), ('AAAAACAAAA', 8)]
seq length 10
seg sites  10
theta s.s. 1.3318772805568886
pi         2.4517866416177907
```

Mutation rate: 1 every 10 sequences

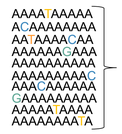
```
sequences  1024
unique     49
[('AAAAAAAAAA', 774), ('CAAAAAAAAA', 34), ('AACAAAAAAA', 17), ('AAAAATAAAA', 12), ('AACAAATAAA', 12), ('AAAAAAGAAA', 11), ('AAAAAAAAAC', 9),
('AAAAAAATA', 8), ('ACAAAAAAAA', 8), ('AAAAAAACA', 8), ('AAAACAGAA', 8), ('AAAAGAAAAA', 7), ('AAAAAAAGA', 7), ('AAAAAACAAA', 7),
('ATAAAAAAAA', 6), ('AAAAAGAAAA', 6), ('AAAACAAAAA', 6), ('AAAAAAAAAG', 5), ('AATAAAAAAA', 5), ('AAAGAAAAAA', 5), ('GAAAAAAAAA', 5),
('AAACACAAAA', 5), ('AAAAAAAAAT', 5), ('AAGAAAAAAA', 5), ('AACAAATGAA', 4)]
seq length 10
seg sites  10
theta s.s. 1.3318772805568886
pi         0.5885053916177908
```

Real-world data

Theta is insensitive for SARS-CoV-2 populations >1,000



Underlying math



Number of nucleotides: 10
Sites with a difference: 10
Average pairwise difference: 2

This quickly saturates: every nucleotide in a small genome will soon have a difference at every position at least once

$$\theta = \frac{S}{\sum_{i=1}^{6n-1} \frac{1}{i}}$$

This takes a long time to saturate

$$\pi = \frac{sum\ of\ pairwise\ differences}{number\ of\ pairs}$$

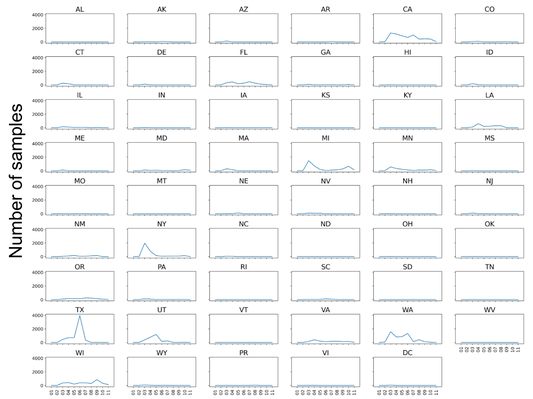
Subsampling a population below 1% skews results

GISAID SARS-CoV-2 samples per capita by country Jan-Nov 2020

	SARS-CoV-2 sequences	Population size	Ratio
USA	80 K	328 M	1 : 4,100
UK	184 K	67 M	1 : 364
Denmark	32 K	6 M	1 : 188

- No centralized effort to collect samples in the U.S.
- Many states have very few samples, making pandemic predictions extremely difficult.
- During the pandemic, we often had to extrapolate from well-sampled countries.

GISAID SARS-CoV-2 samples by state Jan-Nov 2020



Subsampling with Simdemic

Loss function: $\sum_{s \in S} \left(\frac{n_s}{n} - p_s \right)^2$

S: set of all sequences in the population
n_s: count of sequence s in the sample
n: sample size
p_s: true proportion of sequence s in the population

- **Preliminary work: 1% is the threshold for subsampling**
- **Denmark and the UK sampled ~1% of population, while U.S. sampled ~0.01%**
- **In the future, this can be used to estimate how undersampled a population is, and how many more samples would be necessary to get a fair assessment of a pandemic.**

GENO

Like other viruses,

Genome

S
SA
E. co
Bre
F
C
Z
I

C

- Caution mus canonical po to small geno
- Theta is a po variation in a Pi is an appr
- Subsampling results, which U.S. rate of s 19 pandemic
- More researc the relationsh size and hum like Simdemi understand p in small geno

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(Include area code)*