



I N S T I T U T E F O R D E F E N S E A N A L Y S E S

Genomics Analysis of the Covid-19 Pandemic (Presentation)

Katherine I. Fisher-Aylor
Izzy Chaiken
Emily D. Heuring
Felicia D. Sallis-Peterson

September 2022

Approved for public release;
distribution is unlimited.

IDA Document NS D-33260

Log: H 22-000411



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

Copyright Notice

© 2022 Institute for Defense Analyses
730 East Glebe Road, Alexandria, Virginia 22305-3086 • (703) 845-2000.

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (Feb. 2014).

Executive Summary

The 2019 Covid-19 pandemic is the most data-rich pandemic in human history to date, offering us an unprecedented chance to learn from it and prepare for future pandemics. Here, we report on our work studying SARS-CoV-2 genetic samples from fall 2019 to summer 2021. Along with this real-world data, we created a pandemic simulator, Simdemic, to model foundational principles of genome changes during a pandemic. Using our simulator along with the SARS-CoV-2 genetic data, we found that sub-sampling below 1% of the true case count gives a skewed estimate of viral diversity, offering decision makers in the government an actionable lower bound for future sampling methods. We also examined the use of canonical principles of population genetics applied to a viral population. When this was unsuccessful, we found success in alternate methods, which we present here as recommendations for a path forward to better quantify future pandemics. Finally, we tested a combination of methods to predict future SARS-CoV-2 mutations, explaining roughly two-thirds of the Delta strain's mutations, and we discuss methods that could help refine these predictions in future studies.



Genomics Analysis of the Covid-19 Pandemic

STD CRP

Katherine Fisher

Izzy Chaiken

Emily Heuring

Felicia Sallis-Peterson

10/1/21

Institute for Defense Analyses

4850 Mark Center Drive • Alexandria, Virginia 22311-1882

I'm going to tell you about my CRP from this year, where I analyzed the genomics of the Covid-19 pandemic.

In honor of this occasion, my zoom background is a transmission electron microscope image showing SARS-CoV-2, the virus that causes COVID-19. Virus particles are emerging from the surface of human cells cultured in the lab. This image is from the National Institute of Allergy and Infectious Diseases (NIAID).

I'd also like to mention that Izzy Chaiken from ITSD worked with me on this CRP.

Finally, please do ask questions throughout this presentation. You have all been with me during this pandemic, watching cutting-edge science emerge in real time. It's been frustrating and it's been tough and sad, but it's also been an interesting example of how science works.

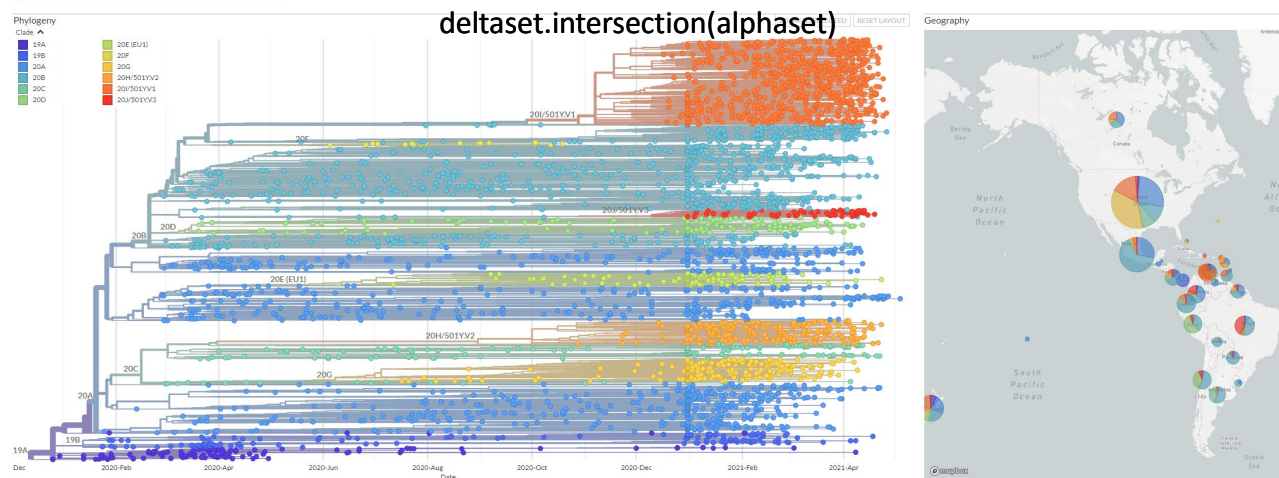
High throughput genomic sequencing has resulted in unprecedented insights into how SARS-CoV-2 is evolving

This is the most data-rich pandemic that has ever existed.

Genomic epidemiology of novel coronavirus - Global subsampling

Built with nextstrain/ncov. Maintained by the Nextstrain team. Enabled by data from [GISAID](#).

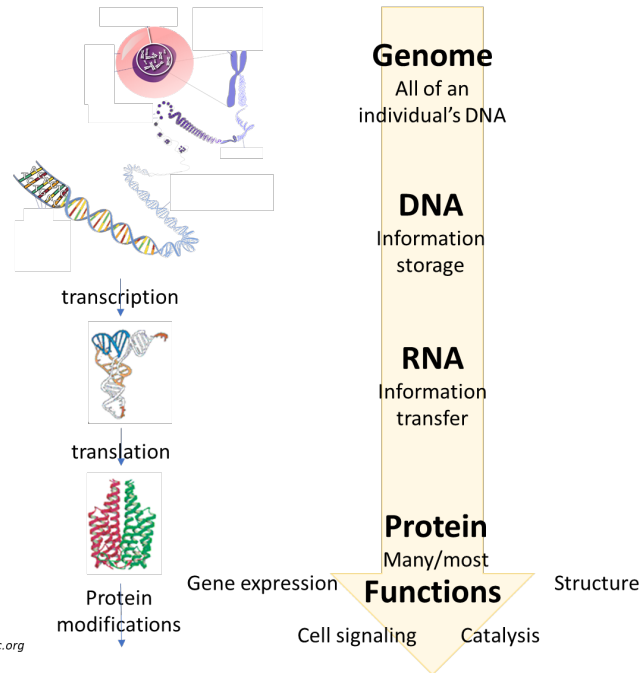
Showing 3890 of 3890 genomes sampled between Dec 2019 and May 2021.



High-throughput genomic sequencing has resulted in unprecedented insights into how SARS-CoV-2 is evolving. In short, we are witnessing the most data-rich pandemic that has ever existed.

So, I decided to see what I could do with all of the viral genetic sequences that were being collected. Specifically, I wanted to know if the genetic sequences could help tell us how many cases of Covid-19 there truly were in the United States, and whether we could predict worrisome future variants.

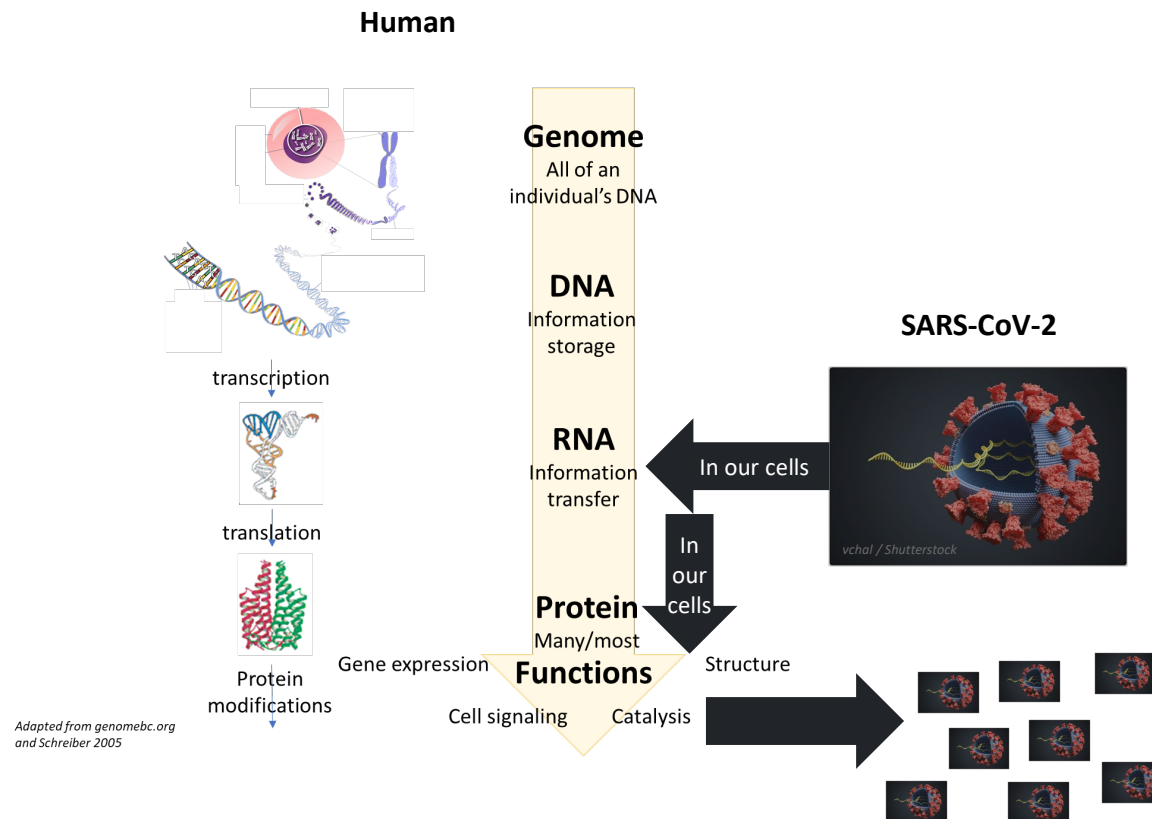
Background: genes and function



Adapted from genomebc.org
and Schreiber 2005

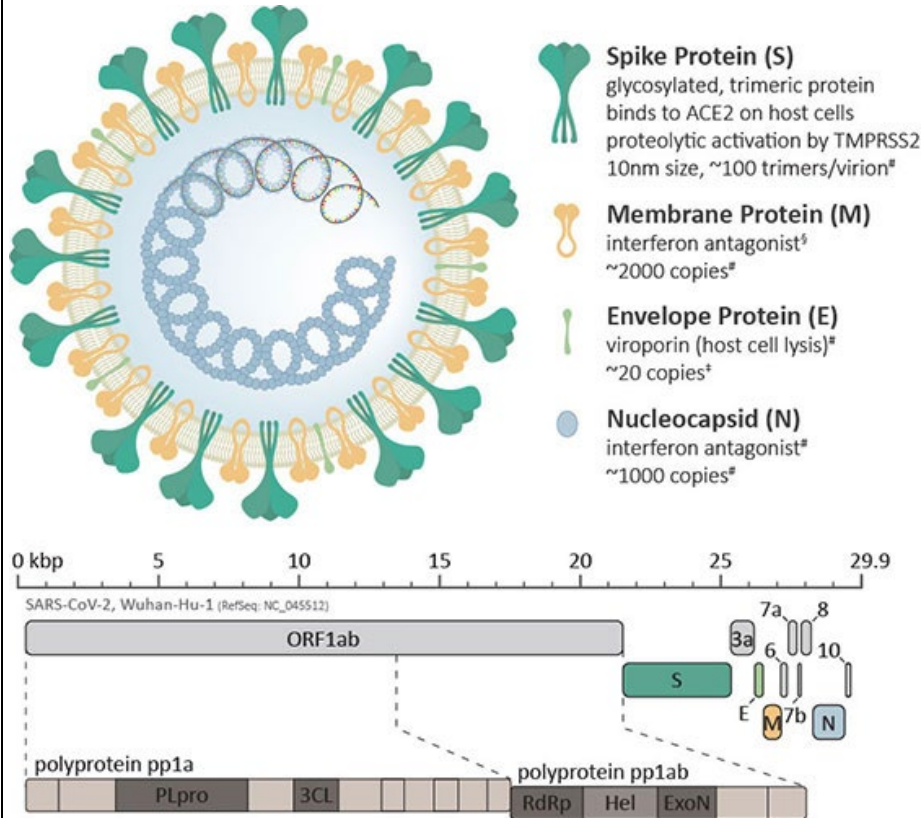
Let me start with a little bit of background. If you recall from my past talks, in organisms like us, genes in the DNA are transcribed into RNA, which are then translated into proteins. So, genes ultimately code for proteins, and proteins do most of the “stuff” in the organism.

Background: SARS-CoV-2 genes and function



SARS-CoV-2 is a virus whose genes are made of RNA. It gains access to our cells and then inserts its RNA with ours. This RNA along with ours is translated, by our cellular machinery, into proteins. These proteins then assemble into more viral particles, as the cell is essentially cannibalized into a zombie viral factory.

Background: the SARS-CoV-2 genome



- The genome has been sequenced (we know the genome sequence)
- We know which parts of the genome are genes
- We have a reasonable idea about what these genes do

Example: gene S (Spike; turquoise) makes a surface protein on the virus that it uses to gain access to human cells via our Ace2 receptors. *This is the protein the vaccines are made against.*

Moderna, Pfizer, and J&J all encode/are the full-length spike mRNA with modifications to stabilize its shape (Baden et al. 2021; Polack et al. 2020)

The SARS-CoV-2 genome is sequenced, which means that we know all of its genetic material.

We also know, partially from work in this virus and partially from studying SARS1 and MERS in the early 2000s, what most of the genome encodes for.

Here, I am showing you the short SARS-CoV-2 genome in the bottom left. It is only 30,000 bases in length and only has 12 genes. The functions of four of these genes are shown in the viral diagram in the upper left.

For example, notice the turquoise spike gene and protein. This is a gene I've been keeping an eye on because it makes the protein that the virus uses to get into our cells—and this is what all of the vaccines we've taken are against. This was a good choice for a vaccine because it will be hard for the virus to mutate away from the vaccine without mutating away from its ability to infect us.

References: Baden, Lindsey R. et al. 2021. "Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine." *The New England Journal of Medicine* 2021, 384: 403–416. <https://doi.org/10.1056/NEJMoa2035389>; Polack, F. P. et al. 2020. "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine." *The New England Journal of Medicine* 2020, 383: 2603–2615. <https://doi.org/10.1056/NEJMoa2034577>.

Background: the data (GISAID: 2M samples as of August 2021)

Metadata:
collection/processing
of the sample

```
In [5]: metadf.loc["EPI_ISL_424237", :]  
  
Out[5]: covv_virus_name          hCoV-19/USA/WA-UW-1739/2020  
covv_collection_date          2020-03-21  
covv_location                 North America / USA / Washington  
covv_clade                    GH  
covv_lineage                   B.1  
pangolin_lineages_version     2021-01-20  
covv_gender                   unknown  
covv_patient_age              unknown  
covv_outbreak                 unknown  
covv_passage                   Original  
covv_seq_technology            unknown  
covv_subm_sample_id             
covv_provider_sample_id         
all_prot_mutations            NaN  
covv_subm_date                2020-04-12  
covv_assembly_method            
covv_coverage                   
Name: EPI_ISL_424237, dtype: object
```

RNA sequence
(here, written as DNA
after sequencing)

```
kfisher@STDDEVSVR4:~/kfisher/CAPE_Covid19/Data/GeneticData/NextStrain/Dec2020/coVsurver_msa_1210$ grep "EPI_ISL_424237" -Al msa_1210.fasta  
>hCoV-19/USA/WA-UW-1739/2020|EPI_ISL_424237|2020-03-21|NorthAmerica  
-----C  
GCAGCCGATCATCAGCACATCTAGGTTTGTCCGGGTGTGACCGAAAGGTAAG---ATGGAGAGCCTTGCCTGGTTTCAACGAGAAAAACACAGTCCAACTCAGTTTGCCTGTTTACAGGTTTCGCGACGTGCTCGTACGTG  
---GCAGAACTCGAAGGCATTTCAGTACGGTCTAGTGGTGAGACACTTGGTGTCTTGTCCCTCATGTGGGGGAAATACCACTAGGCTTACCGCAAGG---TCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGCCATAGTT  
TTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAGCTTCATGCACCTTGTCCGAACAACTGGACTTTTATTGACACTAAGAGGGGTGTATAC-----  
GAACGTTCCTG-AAAAGAGCTATGAATTGCAGACACCTTTTGAATTAATTTGGCAAGAAATTTGACACTCTTCAATGGGGRAATGTCCAAATTTTGTATTTCCCTTAAATTCATAATCAAGACTATTCAACCAAGGGTTGAAAA  
GAAGGTGCCACTACTTGTGGTTACTTACCCC-AAAATGCTGTTTAAATTTTATTGTCCAGCATGTCACAATTCAGAAAGTAGG----ACCTGAGCATAGTCTTGCCGAATAC-----CATATGAATCTGGCTTGAAGACC  
AAAAAGAGAAAGTCAACATCAATATTGTTGGTGACTTTAACTTAATGAAGAGATGCCAATTTTGGCATCTTTTCTGCTTCCACAAGTGCTTTTGTGGAACTGTGAAAGG-TTTGATTATAAGCAATTCACAAATTT  
TACAGAGGCGCGCTATAACAATACTAGATGGAATTTACAGATATTCACTGAGACTCATTGATGCTATGATGTTCCACATCTGATTGGCTACTAACAACTAGTTGTAAATGGCTACATTACAGGTGGTGTGTTTCAGTTGACTT  
TTTATCTCAACCTGTGCTTGTGAAATTTGCGGTGGACAAATTTGCACCTGTGCAAGGAAATTAAGGAGAGTGTTCAGACATCTTTAAGCTTGTAAATAAATTTTGGCTTTGTGTGCTGACTCTATCATTTATTGGTGAGCT  
CTGGTGATTACACCACTTAGAACCACTACTAGTGAAGCTGTGGAAGTCCATTGGTTGGTACACCAG-TTTGTAATTAACGGGCTTATGTTGCTCGAAATCAAGACACAGAAAAGTACTGTGCCCTTGACCTAATATGATG  
CTCGGTACAGAAATAATGAGTTTCGCTGTGTTGGCAGATGCTGTATATAAACTTTGCAACCACTAT-----CTGAATTACTTACACCACTGGGCATTGATTAGATGAGTGGAGTATGGCTACATACTAC  
GCTCTTCAACCTGAAGAGAGCAAGAGAAGATTGTTAGATGATGATAGTCAACAACTGTTGGTCAACAAGAC-GGCAGTGAGGAC-AATCAGACAACTACTATTCAACAATTTGTGAGGTTCAACCTCAATTAGAGATGG  
AATAAGGCTACTAACATGCCATGCAAGTTGAATCTGATGATTACATAGCTACTAATGGACCACTTAAAGTGGGTGGTAGTTGTGTTTAAAGCGGACACAATCTTGCTAAACACTGTCTTCATGTTGTGCGGCCAAATGTTAAC  
TGACAACTTGTTCAGGCTTTTGGAAATGAAGAGTG-AAAAGCAAGTTGAACAAAAGATCGCTGAGATTCTAAGAGAGGAAGTTAAGCCATTTATACTGAAAGT-----AAACCTTCAGTTGAACAGA-GAA  
GGTGATGTTGTTCAAGAGGGTG-TTTTAACTGCTGTGGTTATACCTACTAAAAAGGCTGGTGGCACTACTGAAATGCTAGCG-AAAAGCTTTGAGAAAAGTGCCACAGACAATTATATAACCACTTACC-----  
TCT-GTGTGGAACTAAGGCTAGTTTCATATACAGGTAATATAGGGCTATTAATAGCAGAGGGGTGGTGGTATAGGCTAGATTTTCTTTACAGGCT-AAAACACTGTAGGGTC-ACCTTATCAGACA
```

And for my fellow data enthusiasts, this is a look at what my raw data look like. I have metadata for the collection and processing of the sample. Then I have the RNA sequence (here, converted to DNA because of how it is sequenced), which with modern bioinformatics tools and some ingenuity, I can compare to other viral sequences. Bioinformatics is a lot of substring matching, really.

Motivating questions

- Can we use genetic sequences to determine the number of Covid-19 cases in the United States?
- Can we predict emerging variants?

I'm going to break this talk down into two parts. For the first part, I attempt to use the genetic sequences to find the true case count of Covid-19 in the United States. For the second part, I attempt to use the genetics of the pandemic to predict emerging SARS-CoV-2 variants.

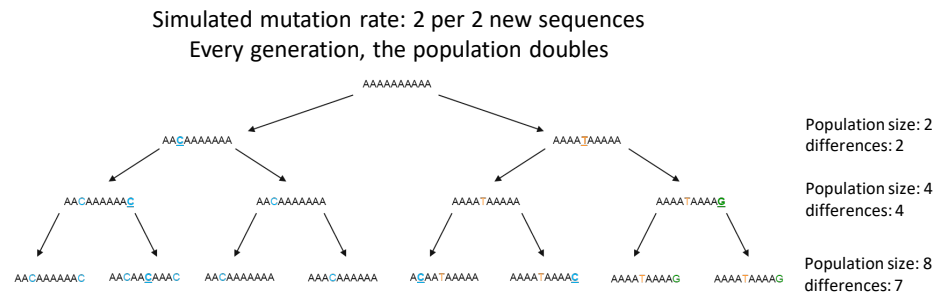
Motivating questions

- Can we use genetic sequences to determine the number of Covid-19 cases in the United States?
- Can we predict emerging variants?

Don't get too excited about this first part—I ultimately failed at this, but learned some valuable lessons that are generalizable to viral genomics and the state of sequencing in the United States.

Background: Population Genetics Analysis

The number of mutations in a population increases with the population size



The reason I pursued this line of research was because of the field of population genetics. An underlying principle of this field is that the number of mutations in a population increases with the population size. In this simple family tree, you can see that with a steady random mutation rate, the number of different mutations, which are colored, are growing at a proportionate rate as the population size itself.

An analysis that is frequently performed in vertebrate species, for example, is to go into an area with a population you don't know much about, sample animals at random, and use their genetic diversity to determine how many members of a population there are in the area. So, I was hoping to do the same with viral diversity and Covid-19 case numbers.

For more information about population genetics, you can read: Okazaki, Atsuko, Satoru Yamazaki, Ituro Inoue, and Jurg Ott. 2021. "Population Genetics: Past, Present, and Future." *Human Genetics* 140 (2), pp. 231–240. <https://doi.org/10.1007/s00439-020-02208-5>.

Background: Population Genetics Analysis

There are multiple ways to measure the number of mutations

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)

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

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

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): frequency of two (i^{th} and j^{th}) sequences
 π_{ij} : number of differences between the two sequences (i and j)

I don't want to get too bogged down in the math here, but after studying several different population genetics estimators, I settled on two. Theta is a canonical genetic diversity estimator that appears most commonly used to estimate population size, and pi is a second method that I settled upon because it makes different underlying assumptions. This work was aided by Felicia Sallis-Peterson and Isaac Chappelle, both of STD.

For more information about using mutations to estimate population size (historically in mammals), you can read: Fu, Yun-Xin. 1994. "Estimating Effective Population Size or Mutation Rate Using the Frequencies of Mutations of Various Classes in a Sample of DNA Sequences." *Genetics* 138, pp. 1375–1386.

...and this paper: Wang, Jinliang. 2005. "Estimation of Effective Population Sizes from Data on Genetic Markers." *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* 360 (1459), pp. 1395–1409. <https://doi.org/10.1098/rstb.2005.1682>.

You can further explore the math and ideas behind the pi estimator here: Nei, Masatoshi, and Fumio Tajima. 1981. "DNA Polymorphism Detectable By Restriction Endonucleases." *Genetics* 97, pp. 145–163.

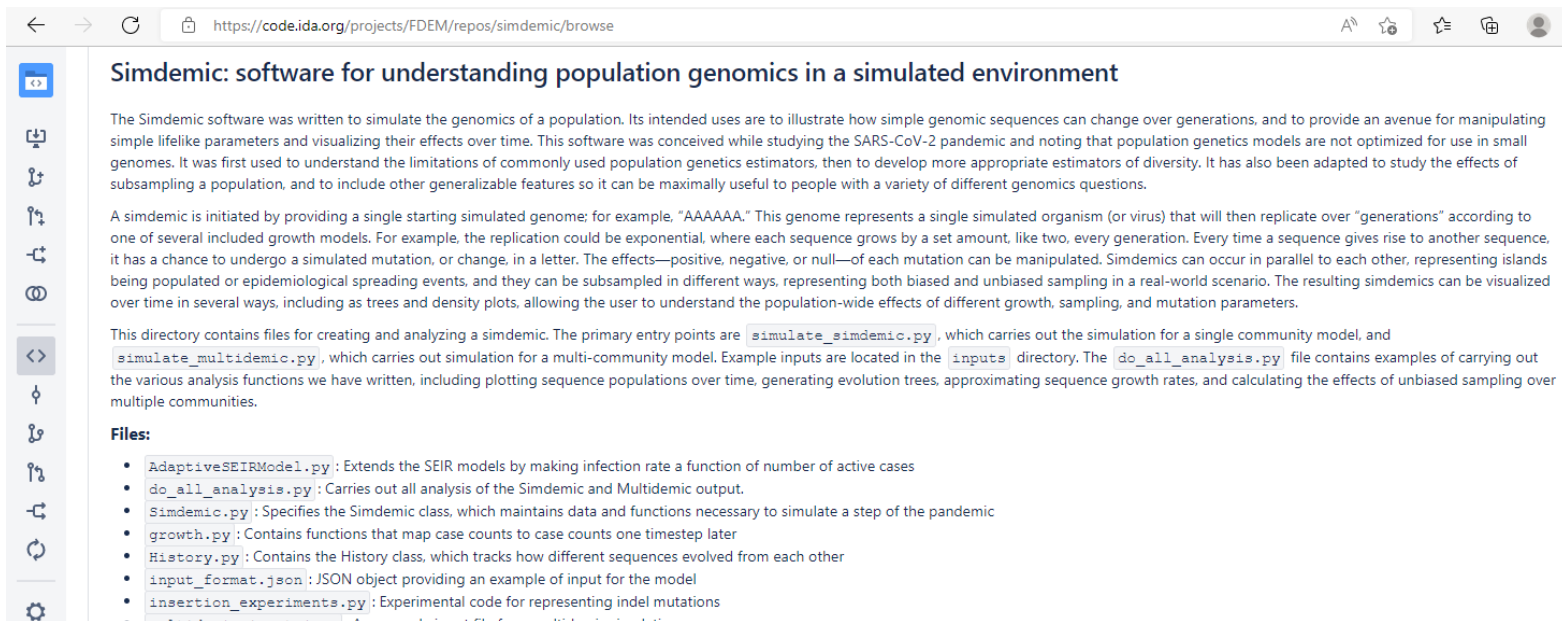
You can further explore the math and ideas behind the theta estimator here: Watterson, G. A. 1975. "On the Number of Segregating Sites in Genetical Models without Recombination." *Theoretical Population Biology* 7, pp. 256–276.

... here: Ferretti, Luca, and Sebastián E. Ramos-Onsins. 2015. "A Generalized Watterson Estimator for Next-Generation Sequencing: From Trios to Autopolyploids." *Theoretical Population Biology* 100C, pp. 79–87. <https://doi.org/10.1016/j.tpb.2015.01.001>.

...and here, to understand coalescent theory, which is behind the Wu-Watterson theta estimator: McVean, Gil; Philip Awadalla, and Paul Fearnhead. 2002. "A Coalescent-Based Method for Detecting and Estimating Recombination from Gene Sequences." *Genetics* 160, pp. 1231–1241.

New tool: Simdemic pandemic simulator

<https://code.ida.org/projects/FDEM/repos/simdemic/browse>



The screenshot shows a web browser window displaying the Simdemic project page. The browser's address bar shows the URL <https://code.ida.org/projects/FDEM/repos/simdemic/browse>. The page title is "Simdemic: software for understanding population genomics in a simulated environment". The main content area contains a description of the software, its intended uses, and a list of files. The left sidebar shows a navigation menu with icons for code, issues, pull requests, and other project features.

Simdemic: software for understanding population genomics in a simulated environment

The Simdemic software was written to simulate the genomics of a population. Its intended uses are to illustrate how simple genomic sequences can change over generations, and to provide an avenue for manipulating simple lifelike parameters and visualizing their effects over time. This software was conceived while studying the SARS-CoV-2 pandemic and noting that population genetics models are not optimized for use in small genomes. It was first used to understand the limitations of commonly used population genetics estimators, then to develop more appropriate estimators of diversity. It has also been adapted to study the effects of subsampling a population, and to include other generalizable features so it can be maximally useful to people with a variety of different genomics questions.

A simdemic is initiated by providing a single starting simulated genome: for example, "AAAAAA." This genome represents a single simulated organism (or virus) that will then replicate over "generations" according to one of several included growth models. For example, the replication could be exponential, where each sequence grows by a set amount, like two, every generation. Every time a sequence gives rise to another sequence, it has a chance to undergo a simulated mutation, or change, in a letter. The effects—positive, negative, or null—of each mutation can be manipulated. Simdemics can occur in parallel to each other, representing islands being populated or epidemiological spreading events, and they can be subsampled in different ways, representing both biased and unbiased sampling in a real-world scenario. The resulting simdemics can be visualized over time in several ways, including as trees and density plots, allowing the user to understand the population-wide effects of different growth, sampling, and mutation parameters.

This directory contains files for creating and analyzing a simdemic. The primary entry points are `simulate_simdemic.py`, which carries out the simulation for a single community model, and `simulate_multidemic.py`, which carries out simulation for a multi-community model. Example inputs are located in the `inputs` directory. The `do_all_analysis.py` file contains examples of carrying out the various analysis functions we have written, including plotting sequence populations over time, generating evolution trees, approximating sequence growth rates, and calculating the effects of unbiased sampling over multiple communities.

Files:

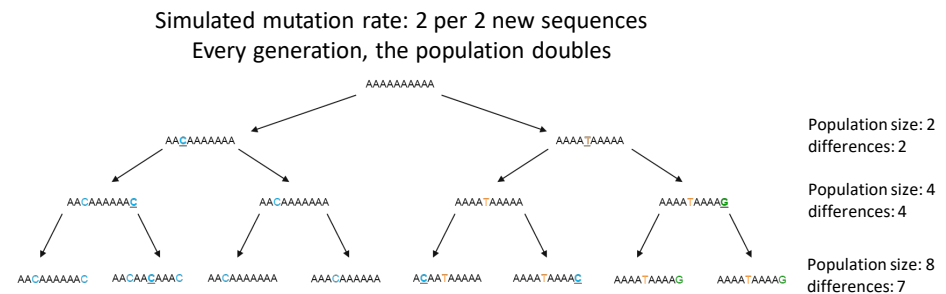
- `AdaptiveSEIRModel.py`: Extends the SEIR models by making infection rate a function of number of active cases
- `do_all_analysis.py`: Carries out all analysis of the Simdemic and Multidemic output.
- `Simdemic.py`: Specifies the Simdemic class, which maintains data and functions necessary to simulate a step of the pandemic
- `growth.py`: Contains functions that map case counts to case counts one timestep later
- `History.py`: Contains the History class, which tracks how different sequences evolved from each other
- `input_format.json`: JSON object providing an example of input for the model
- `insertion_experiments.py`: Experimental code for representing indel mutations

Now I'd also like to introduce you to one of the main products of my CRP: the Simdemic software package. With help from Izzy Chaiken, I built a pandemic simulator in order to study simplified elements of a pandemic with control that we could never find in the real world.

New tool: Simdemic pandemic simulator software tool

Simdemic's algorithm, simplified:

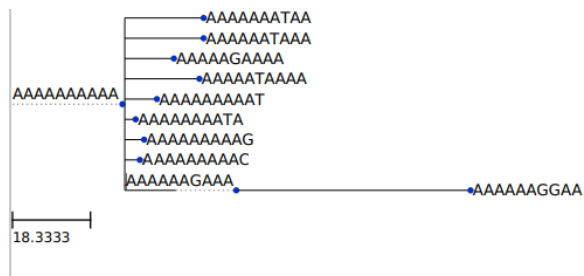
1. Ingest a DNA sequence provided by the user.
2. Use one of several growth models to calculate the number of sequences there will be in the next "generation"
3. Create "mutations" in the next generation randomly, or based on a provided model specifying which positions would cause deleterious, beneficial, or neutral mutations (probability score).
4. Continue these steps until the user-provided number of generations has passed.
5. Return the final "population" to the user for further analysis.



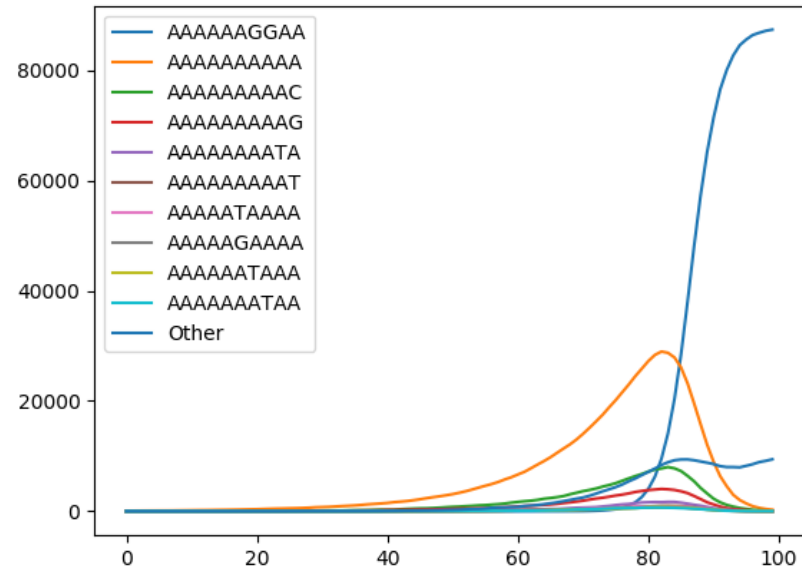
Single Community Simdemic – Sequence Counts

The original sequence, AAAAAAAAAA, is eventually dominated by the sequence AAAAAAGGAA

Simdemic evolutionary tree



Fauxdemic sequence counts over time



This simulator allows you to specify a starting “genomic sequence,” a growth model, and a mutation model. We also added in different abilities as the CRP progressed and we explored other elements of pandemics. Here, I am showing one simple pandemic simulation. On the left is an evolutionary tree visualization of the pandemic, and on the right is the prevalence of different mutations that occur during this simulation over the generations.

Pi vs. Theta



Pandemic simulator 1.0
Seed: AAAAAAAAAA
Growth: exponential, 2
Generations: 10

Mutation rate:
1 every 2 sequences

```
sequences 1024
unique 726
[('AAAAACAAA', 10), ('AAAAAGCAA', 9), ('AAATACTACA', 7), ('AAATACAAA', 7), ('AAAAAGCAAT', 6), ('AAAAAGCAA', 6), ('AAGAAATAT', 6),
('AAAAACACC', 4), ('AAATAACACA', 4), ('AAAAACTACA', 4), ('TAAAAACAAA', 4), ('TATATACCAA', 4), ('TAACAGCAA', 4), ('GTAACACAAA', 4),
('CAATACAAA', 4), ('AAAAACTAAA', 4), ('GAAAAACAAG', 4), ('GAAAAACGAG', 4), ('AAGAAAAATA', 4), ('AACACCCAAA', 4), ('AAAAATTGAG', 4),
('AAAAAGAGAG', 4), ('AAGAAAAACG', 4), ('AAAAATAACG', 4), ('AAGAAATCAA', 3)]
seq length 10
seg sites 10
theta s.s. 1.331877305568086
pi 5.413810483870968
```

Mutation rate:
1 every 5 sequences

```
sequences 1024
unique 237
[('AAAAAAAAA', 198), ('AAAAATAAAA', 68), ('AAAAAAAAG', 36), ('AATAAAAAA', 27), ('AAGAAAAAA', 21), ('AAAAAATAA', 21), ('AAAAAATAA', 18),
('AAAAACAAA', 17), ('AATAAAAAAG', 16), ('AAAAAACACA', 15), ('GAAAAAAA', 15), ('AAAAATAAA', 15), ('ATAAAAAA', 14), ('AATAAAAAA', 13),
('AAAAAACA', 13), ('AAAAAATAA', 11), ('AAAAAACAAA', 10), ('AAGAGAAAA', 10), ('TAAAAAAA', 9), ('AAGAAAAA', 9), ('CAAAAAA', 8),
('AATAAAAAA', 8), ('AAAAATAAAA', 8), ('AAAAAGAAA', 8), ('AAAAAACAAA', 8)]
seq length 10
seg sites 10
theta s.s. 1.331877305568086
pi 2.4517866416177907
```

Mutation rate:
1 every 10 sequences

```
sequences 1024
unique 49
[('AAAAAAAAA', 774), ('CAAAAAAAA', 34), ('AACAAAAA', 17), ('AAAAATAAA', 12), ('AACAAATAA', 12), ('AAAAAAGAA', 11), ('AAAAAACA', 9),
('AAAAAATAA', 8), ('ACAAAAA', 8), ('AAAAAACAA', 8), ('AAAAACAGAA', 8), ('AAAAGAAAA', 7), ('AAAAAAGA', 7), ('AAAAACAAA', 7),
('ATAAAAAA', 6), ('AAAAAAGAAA', 6), ('AAAAACAAA', 6), ('AAAAAAG', 5), ('AATAAAAAA', 5), ('AAGAAAAA', 5), ('GAAAAA', 5),
('AAAAACAAA', 5), ('AAAAAATAA', 5), ('AAGAAAAA', 5), ('AACAAATGAA', 4)]
seq length 10
seg sites 10
theta s.s. 1.331877305568086
pi 0.5885853916177908
```

In a small genome, the number of locations with mutations saturate very quickly, meaning that the infinite sites assumption and resulting metrics, like theta, are inappropriate – but pi is useful.
(SARS-CoV-2 is 30,000 bp – vertebrate genomes are in the Billions of bp)

Using the Simdemic software, I was able to analyze how the two different population genetics estimators, θ and π , performed in a small genome. In short, θ fails and π succeeds. These three different rows show three simulated pandemics with different variability of sequences—at the top is the population with the highest rate of variability and at the bottom is the lowest rate of variability. An appropriate estimator will be sensitive to variation.

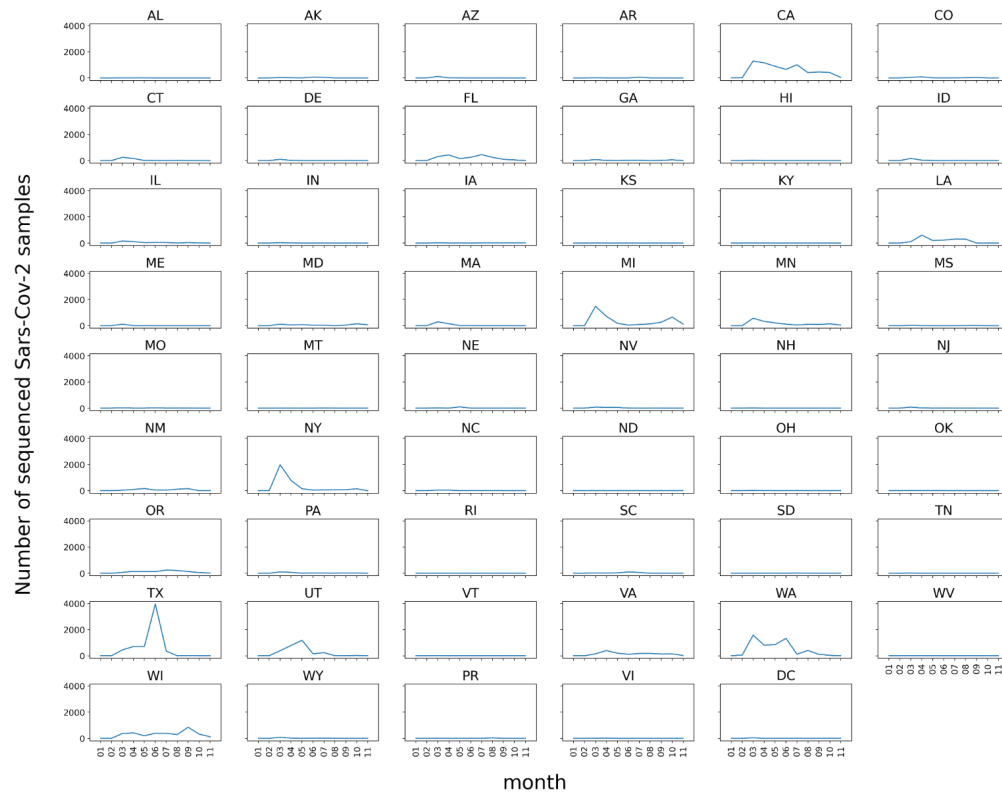
When measuring the final population of 1,000 entities, θ is the same in each of the three populations, meaning it is no good. However, π does vary, suggesting it is sensitive to the amount of variability. When we drilled down on the reasons for this difference, it turned out that θ essentially makes an assumption that mutations can occur at an infinite number of sites. This makes sense for vertebrate species with huge genomes (ours is 3 billion base pairs), but it makes no sense for a tiny viral genome—or this ludicrously tiny simulated example.

Motivating questions

- Can we use genetic sequences to determine the number of Covid-19 cases?
 - MAYBE: but population genetics tools need to be adapted to small genomes
- Can we predict emerging variants?

So, the answer to the first part is a bold “maybe” except for the fact that....

The U.S. sequenced relatively few SARS-CoV-2 samples (data shown for 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

We had a data problem in the United States for the duration of the pandemic. Especially in the first two thirds of the pandemic, the United States was woefully under-sampled in terms of the number of viral sequences it collected.

If you look at the rate of sample collection compared to population size on the right, you will see what our rate of sample collection looks like compared to two countries that really did have a good effort during the pandemic.

If you look at these line graphs, you will see, state by state, the number of samples collected each month in 2020. Many states had fewer than 100 samples collected for all of 2020, and some had none at all. There are a few spikes of sample collection here and there, and they are associated with a handful of individual academic papers. Each paper did a good job of collecting samples in a particular city or state to measure whatever they were trying to get a handle on—for example, one measured an early superspreader event in Manhattan. These sequences were shared with the community, which is how I obtained them, but they of course do not tell us about the pandemic as a whole in the United States.

For me, this was particularly annoying because this is NOT a matter of capability—scientists from this country invented the modern DNA sequencers and most modern genomics algorithms. We have capable people in academic institutions and biotech throughout the country, and anyone who is trained in any sort of genomics analysis, like me, can learn how to do this type of analysis. There are even more people who know how to collect samples and sequence them. There must be thousands or tens of thousands of people who can do this. But we just didn't—there was no country-wide organized effort to collect fairly distributed viral samples.

Motivating questions

- Can we use genetic sequences to determine the number of Covid-19 cases?
 - MAYBE: but population genetics tools need to be adapted to small genomes
 - NO: sampling in the U.S. throughout the pandemic has been sporadic at best
 - There are also some more unknowns (viruses per patient, variation within a patient, etc.)
- Can we predict emerging variants?

So, unfortunately for me, even though I learned how to adapt population genetics metrics to a smaller genome, this was a nonstarter because of under-sampling in the United States. I will also point out that there are some other unknowns that would also have made this difficult, such as knowing how many copies of SARS-CoV-2 are in the average patient. I hold out hope that we might be able to, with better sequencing, pursue this method in future pandemics because having a new method to determine the true number of cases in a pandemic would be invaluable. Current methods rely on testing the population, and generally that is done in a very biased way (i.e., waiting for sick patients coming in). A method that operates based on genetic variation would not be as biased by this problem, assuming enough samples were collected and the samples were taken from geographically/socially different sources.

Ultimately, our colleagues at IDA and elsewhere in the community used different methods like antibody titers to better estimate how many true cases of Covid-19 there were relative to the official case counts in the United States. So, in the end, we got the answers we needed and learned something valuable about pandemic genomics, during the first time in the world we were able to truly study them.

Simdemic: Biased Sampling

- We have added an ability to down-sample and create biased samples of a simulated pandemic.
- The simulation provides us the ability to truly know the “ground truth” of every sequence and its prevalence, which is lacking in a true pandemic situation.

- Sample n percent of sequences from one

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

S is the set of all sequences in the population, n_s is the count of sequence s in the sample, n is the sample size, and p_s is the true proportion of sequence s in the population.

- Create a test statistic by taking unbiased samples, then counting how many of those have higher loss

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. **Early estimates suggest a drop in proper estimating capacity at 1%.**

Since we have very fine control over every element of our simulated pandemic, this includes knowing the entire population and its entire evolutionary history. We decided to exploit this by adding a down-sampling and biased sampling ability to the Simdemic software. We ran out of time before we could fully analyze it, but we roughly estimate that it is safe to sample as little as 1% of a pandemic without getting too biased a determination of the true pandemic's genomics. Note that this is roughly what the UK and Denmark ended up doing during 2020, while we were several logs worse than that in terms of our sample collection.

I would like to follow up on this work in the future to help determine whether a pandemic is under-sampled and if so, how much more data needs to be collected to get a fair measure of the pandemic.

Conclusions, part 1

- Many current population genetics estimators are not suited for analysis of small genomes (like viruses)
 - The π metric is the only one we studied that was useful
- Sequencing efforts in the U.S. were insufficient for genomics analysis
- We present “Simdemic”: a pandemic simulator to aid in the further study of these principles
 - Specifiable genomes
 - Multiple growth models
 - Multiple mutation rates and viability estimators
 - Tree and graph-based visualizations

So, for part 1, I found that although many classic population genetics estimators are not suitable for use on small viral genomes, π is worth pursuing in the future. I also found that the lack of a centralized sequencing effort really hurt the United States in terms of understanding how the pandemic was unfolding here.

Finally, we produced what I hope will be a useful teaching and exploratory tool for analyzing pandemics and population genetics principles in the future.

Motivating questions

- Can we use genetic sequences to determine the number of Covid-19 cases in the U.S.?
- Can we predict emerging variants?

The second half of this talk will concern emerging variants. This is all near and dear to all of us, because all of you, my friends, are stuck here with me watching cutting-edge science emerge in real time. Right now, we are all familiar with the Delta variant, and before that, there was the “UK variant” (now alpha) the “Brazil variant” (now gamma) and so on and so forth.

So, I was interested to see if I could use a different aspect of the genome to determine which types of characteristics might emerge in the future.

Background: conservation

High “conservation” means that, compared to related species, a particular nucleotide is the same. Low conservation means that the nucleotide changes over evolutionary history.



Let me start with some background information. First is the concept of conservation, which is very central to genomics analysis. A conserved region is one that is not observed to vary over the course of evolutionary history. In humans, for example, we share 99% of our genome with chimpanzees. However, as you go back along our evolutionary tree and start looking at the genomes of organisms that are more and more distantly related to us, you will still find some genes that are the same. RNA polymerase II, the enzyme that is required for converting our genes into RNA, is very similar, not just throughout the animal kingdom, but also down through plants, fungus, even bacteria. So, we would say that the gene for RNA polymerase II is extremely highly conserved. For another example, there are some regions of the genome that are unique to humans. Many of these relate to brain size and characteristics. These regions of the genome would be called unconserved because they are not shared with any species, even closely related ones.

Now with those analogies out of the way, observe this section of the SARS-CoV-2 genome. At the top, I am showing the gene for spike, and below it, I am showing some areas within spike that have a known function. Below that is conservation. In the individual green density plots, I am showing 56 different viruses that affect a variety of different species, from SARS1, to bat viruses, to turkey coronavirus, to a type of viral bronchitis. Each of these viral genomes has been lined up against SARS-CoV-2 and then their conservation scores are shown at each nucleotide location. Basically, where you see a lot of green is where these viruses stay the same over the course of evolution. You can see that this covers the right half of the spike protein. Notably, there is less conservation, or more differences, under the ACE2 binding domain. This is presumably because each species has a slightly different structure of ACE2 receptor and the virus has to change to adapt to its current species.

Just above all of the green tracks is a red and blue track. This is a metric that combines all of the below green tracks, nucleotide by nucleotide, and gives an overall conservation score for SARS-CoV-2.

To review the concept of conservation, you can read one of the seminal founding discussions of the concept here: Zuckerkandl, Emile, and Linus Pauling. 1965. "Evolutionary Divergence and Convergence in Proteins." *Evolving Genes and Proteins. A Symposium Held at the Institute of Microbiology of Rutgers: the State University with Support from the National Science Foundation*, edited by Vernon Bryson and Henry J. Vogel, New York, Academic Press Inc., pp. 97–166.

... and these two more recent publications:

Asthana, Saurabh; Mikhail Roytberg, John Stamatoyannopoulos, and Shamil Sunyaev. 2007. "Analysis of Sequence Conservation at Nucleotide Resolution." *PLOS Computational Biology* 3 (12), Article e257, pp. 2559–2568. <https://doi.org/10.1371/journal.pcbi.0030254.g001>.

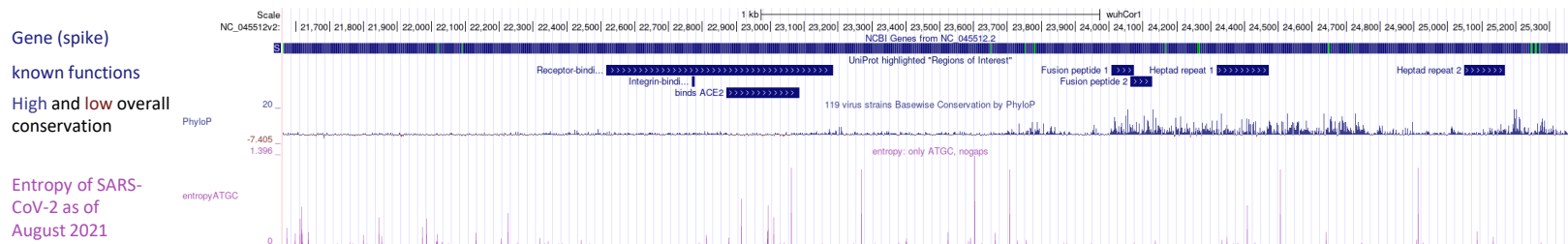
Cooper, Gregory M., and Christopher D. Brown. 2008. “Qualifying the Relationship between Sequence Conservation and Molecular Function.” *Genome research* 18 (2), pp. 201–205. <https://doi.org/10.1101/gr.7205808>.

... as well as a specific discussion of SARS-CoV-2 conservation here: Srinivasan, Suhas, Hongzhu Cui, Ziyang Gao, Ming Liu, Senbao Lu, Winnie Mkandawire, et al. 2020. “Structural Genomics of SARS-CoV-2 Indicates Evolutionary Conserved Functional Regions of Viral Proteins.” *Viruses* 12 (4). <https://doi.org/10.3390/v12040360>.

To learn more about the SARS-CoV-2 genome as well as this genome browser, you can read: Fernandes, Jason D., Angie S. Hinrichs, Hiram Clawson, Jairo Navarro Gonzalez, Brian T. Lee, Luis R. Nassar, et al. 2020. The UCSC SARS-CoV-2 Genome Browser.

Background: entropy

High “entropy” means that, within a population, a particular nucleotide is highly variable.
An entropy of zero means that the nucleotide is never observed to vary.



Measures of observed nucleotide variance....

over evolution: conservation

within a species: entropy

The second piece of background information concerns the concept of entropy. I can only assume that this is an embarrassing use of the term, so I apologize—I didn't come up with it myself. In this particular biological case, “entropy” means “variance,” which is sort of like disorder, I suppose.

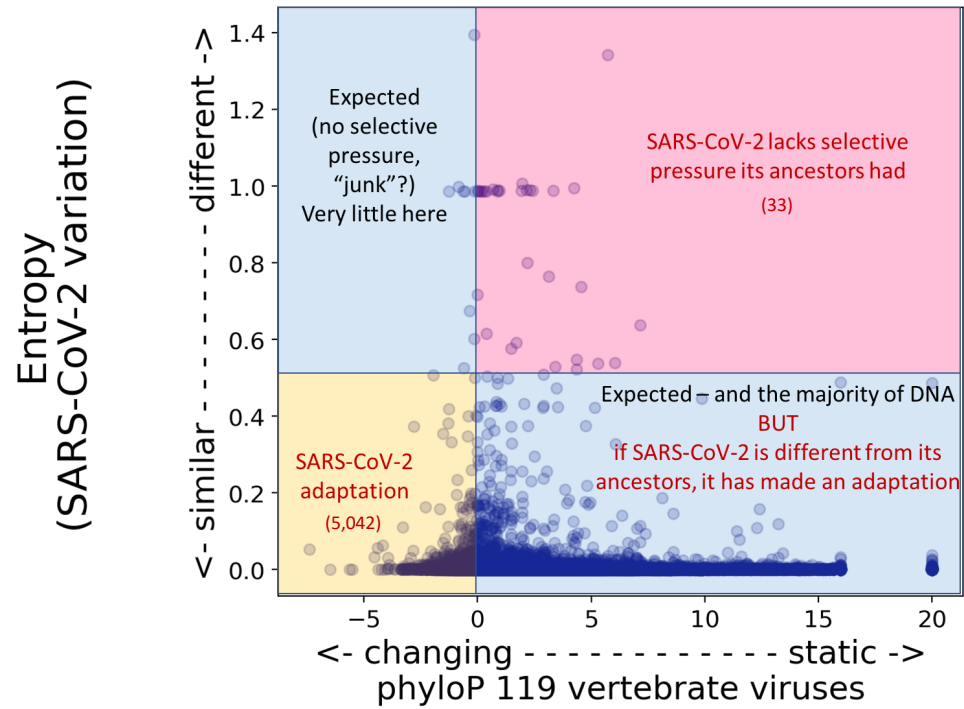
Similar to the pi metric from the previous section, entropy measures variation within a population. However, while pi is a number that is applied to the whole population regardless of where the differences occur, entropy measures variation nucleotide by nucleotide along the entire genome. A high entropy score means that a base pair is observed to have high variability. It basically doesn't matter what nucleotide occurs at that position. A low entropy score, however, means that variation is rarely observed—that nucleotide is so important to the survival of the virus that it is rarely able to change.

So, both conservation and entropy are different measures of observed changes at specific positions in a genome. Conservation measures changes over evolution, and entropy measures observed changes within a current population of the same species.

You can read more about the concept of entropy in genomics here: Schmitt, A. O., and H. Herzel. 1997. “Estimating the Entropy of DNA Sequences.” *Journal of Theoretical Biology* (1997), pp. 369–377.

...and here, relating to SARS-CoV-2 earlier in the pandemic: Ghanchi, Najia Karim, Asghar Nasir, Kiran Iqbal Masood, Syed Hani Abidi, Syed Faisal Mahmood, Akbar Kanji, et al. 2021. “Higher Entropy Observed in SARS-CoV-2 Genomes from the First COVID-19 Wave in Pakistan.” *PloS one* 16 (8), e0256451. <https://doi.org/10.1371/journal.pone.0256451>.

Entropy vs. conservation for the entire genome



The reason these concepts are important is because they can both help us predict what is likely to change in the future. Conserved elements can tell us what seems to be generally required for being a coronavirus. Elements with low entropy can tell us what the current human SARS-CoV-2 appears to require for survival.

In this plot, I am showing for each position in the SARS-CoV-2 genome what the conservation and entropy scores. Naively, we expect to see conservation and entropy track each other. Positions that have high conservation are important for viral survival over evolution, and should generally have low entropy because they will not have much variation in the current pandemic. The blue boxes represent these expected cases.

In the red square, conversely, are 33 elements with high conservation and high entropy. This means that although these positions are very important to viral survival historically, they have had a lot of changes in the current pandemic. These represent areas where the virus lacks the selective pressure its ancestors had.

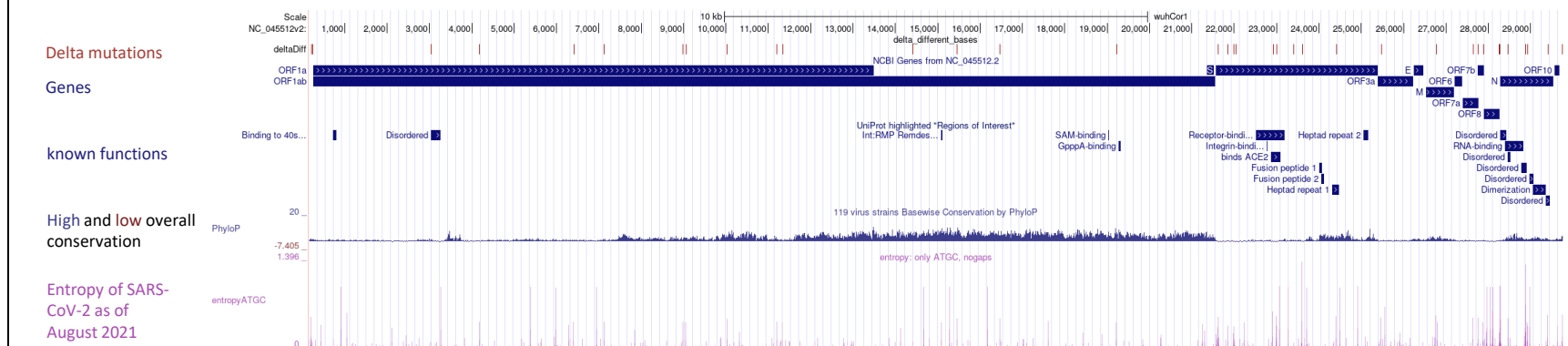
In the yellow square are roughly 5,000 positions where there is low conservation and low entropy. This means that although these positions are not vital to viral survival historically, they nevertheless do not tend to vary in the current pandemic. These locations might represent new SARS-CoV-2 adaptations.

Finally, entropy and conservation are ignorant to what the actual nucleotides are. It is possible for an element to be both highly conserved during evolution and invariant in the current pandemic—and yet also be a different element in the current pandemic compared to what it was in history. These elements represent a third group where SARS-CoV-2 has made an important adaptation.

Trying to predict the Delta variant

1. Pull all observed instances of Delta variant from GISAID data (B.1.617.2)
2. Align the sequences with each other
3. Calculate the “consensus” (average) genome of Delta, including deletion of base pairs
4. Determine where Delta differs from the current canonical sequence of SARS-CoV-2

This results in 46 mutations that characterize Delta



Now, since it is near and dear to us, so to speak, I am going to focus on the Delta variant for this section.

To find all of the elements that make Delta what it is, I pulled all observed instances of the Delta variant from the GISAID data, which amount to 344,845 individual viral sequences. I aligned these sequences with each other and then created a ‘consensus genome’ for Delta by taking the most prevalent base pair at each location, including deletions. Then I compared the Delta consensus sequence with the current canonical sequence of SARS-CoV-2 that the genomics community has agreed on.

This results in 46 nucleotides that characterize Delta. I am showing them here across the entire SARS-CoV-2 genome. The Delta mutations are in red at the top. The 12 SARS-CoV-2 genes are below that, and then are the known functional regions of each protein. Below that are the conservation and entropy scores for SARS-CoV-2. At first glance, you’ll see that there are a reasonable number of mutations in the spike gene – yikes! A highly mutated Spike protein could cause the virus to evade our vaccine immunity.

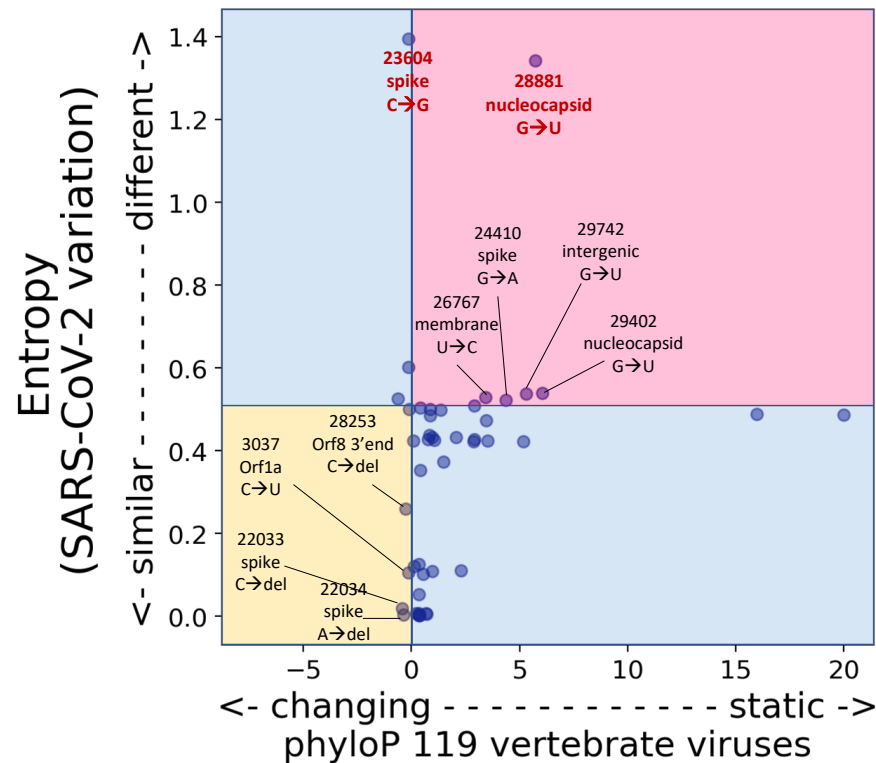
If we went back in time to before Delta existed, how many of the 46 mutations would we have been able to predict?



So, let's go back in time, data-wise, and see if we could have predicted the 46 mutations that make Delta what it is. Here, I am using a combination of entropy, conservation, and knowledge of mutations that occurred in past major strains. To do this, I first made a consensus sequence of Delta by computing the most prevalent base for each position for each Delta-strain sequence in GISAID. Next, I took all of the SARS-CoV-2 sequences that occurred before Delta emerged and computed the entropy of these pre-Delta sequences. Finally, I made a consensus sequence for each past significant strain and calculated the mutations that had already occurred in them. I would predict the most likely places for future mutations would be regions with high entropy as well as regions that had previously mutated. I would also predict that some, but not all, of the places where SARS-CoV-2 shows conservation with related viruses would not be mutated.

Entropy/conservation predicts 2-10 Delta mutations

(there would have been 0 overpredictions for the 2 / 5,065 overpredictions for the 10)



Using entropy and conservation, we would have predicted two mutations with high accuracy based on their extremely high entropy scores.

We might also have been able to predict 8 other Delta mutations in the red and yellow squares, although those would have had a very high over-prediction or “false positive” rate because there were 5,000 other locations in the yellow box for the whole genome.

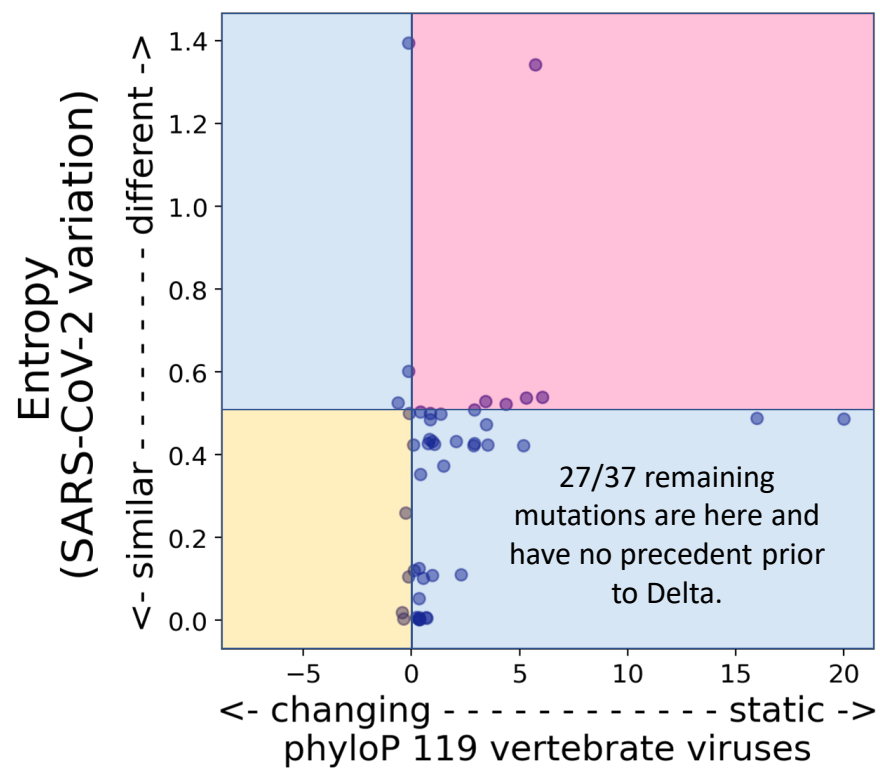
The blue boxes are particularly interesting to keep an eye on in the future. They may represent places where selective pressure on SARS-CoV-2 has changed, causing mutations to occur in regions that were previously conserved.

Using past variants predicts 9 Delta mutations

(there would have been 87 overpredictions for these 9)

Location	Gene	Mutation	variants	Identified previously?
241	intergenic	C → U	α, β, γ	
3037	Orf1a	C → U	α, β, γ	Y
14408	Orf1b	C → U	α, β, γ	
23403	spike	A → G	α, β, γ	
28253	Orf8	C → del	β	Y
28881	nucleocapsid	G → U	α, γ	Y
23604	spike	C → G	α	Y
28271	Intergenic/Orf9	A → del	α	
22917	spike	U → G	B.1.429	

If we looked at past variants of SARS-CoV-2 to predict new variants, we would have predicted an additional four (or six) Delta mutations. There are 87 total mutations between the alpha, beta, and gamma variants, and nine of these are shared with the Delta variant. However, this only explains 14 of the 46 Delta mutations. Is there anything we could have done to predict the rest?



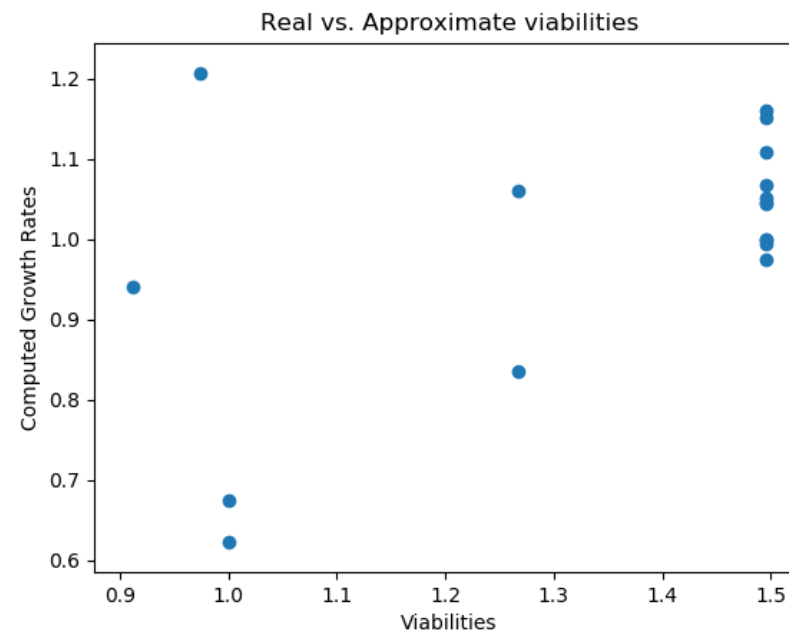
Going back to the conservation vs. entropy plot, 27 of the 37 remaining mutations are in the lower right quadrant. This was the place where the nucleotide identity itself had to be analyzed on top of the conservation and entropy scores.

I looked to see if there was any hint prior to Delta that these mutations would crop up, and so far, I have failed to find them. So far, it looks like these 27 mutations occurred spontaneously and were retained because they happened to be extremely advantageous to SARS-CoV-2.

Simdemic: Viability scores

Viability is a parameter built into each sequence that determines how likely a sequence is to infect people.

Computing growth rate in the case counts from one generation to the next allows us to approximate these viabilities.

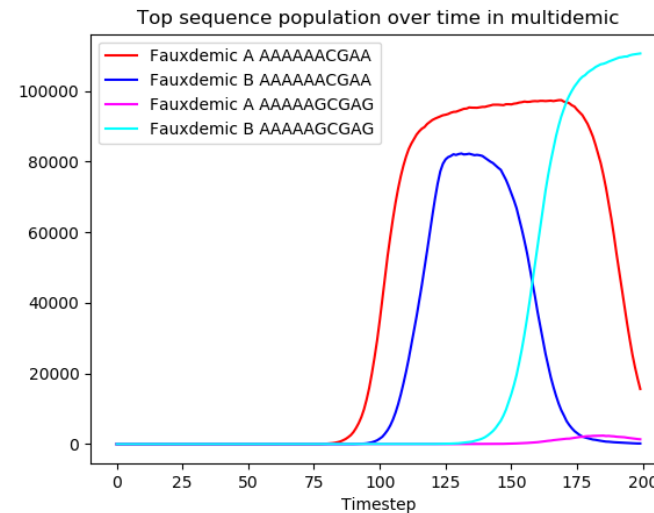


In order to better understand the spread of mutations with both positive and negative selective pressure, we have encoded a viability score for a Simdemic's mutations. This can be randomly generated or specified by the user.

Here, we show that Simdemic's encoded viability scores correlate with the subsequent viral growth rates. This mimics the real-life example of Delta evolving a set of highly beneficial mutations and then taking over the SARS-CoV-2 population.

Simdemic: “Multidemic” system for understanding founder effects and viral spread

- We want to be able to simulate multiple communities with their own dynamics of case spread and counts of different sequences
- Solution: Run multiple simdemics in parallel
- Each timestep, each community transmits a predefined proportion of cases to other communities
- Some shared info, such as viability of each sequence and evolution history



- Simdemic A sees the first rise of AAAAAACGAA.
- Sequence AAAAAGCGAG becomes dominant in Simdemic B, but not Simdemic A.
- Examining viability scores plus evolutionary history tells us how much effect a mutation has compared to chance in determining why a particular variant predominates.

Finally, we have added a “multidemic” capability to the Simdemic software. This capability mimics the spread of random entities to a new self-contained area. This mimics superspreader events and founding events, but it also allows us the ability to decouple the effects of random chance from viability in whether a particular mutation will predominate in a population.

Here, I am showing that in two different linked Simdemic simulations, different strains predominate in each case, even though they have the same viability scores as each other.

Conclusions, part 2

- High entropy, but not low conservation, predicts some upcoming mutations with high accuracy.
- One third of Delta's mutations can be explained by having previously evolved in SARS-CoV-2 variants.
- Two thirds of Delta's mutations appear to have evolved in highly conserved, highly invariant regions with no prior precedent in other variants. **Constantly monitoring a pandemic via sequencing may be the only way currently to detect the emergence of mutations like this.**
- We present “Simdemic”: a pandemic simulator to aid in the further study of these principles
 - Specifiable genomes
 - Multiple growth models
 - Subsampling/biased sampling
 - Tree and graph-based visualizations
 - **Multiple mutation rates and viability estimators**
 - **Outbreak events/founder effects**

In conclusion, high entropy, but not low conservation, predicts some upcoming mutations with high accuracy.

A third of Delta's mutations can be explained by having previously evolved in other SARS-CoV-2 variants. As far as predictive biology goes, this is probably fairly good.

However, the remaining two thirds of Delta's mutations appear to have evolved with no prior precedent either in evolutionary history or in the history of the Covid-19 pandemic. Constantly monitoring a pandemic via sequencing may be the only way currently to detect the emergence of mutations like this. Given the lack of sequencing in the United States during the bulk of the pandemic, we are lucky that other countries did this, or we would not have noticed the emergence of Delta at all.

Finally, we have added capabilities to the Simdemic software to study emerging mutations, both beneficial and detrimental. This is done in a way to enable decoupling of the effects of mutations and the effects of chance from founder effects.

Future directions

- Use genomic data to predict case number (would rely on future research)
 - Finding the number of viruses on average per person.
 - Knowing virus variability per person – and understanding to what extent a SARS-CoV-2 sequence is a consensus sequence of all of a person's variants. (Dr. Elodie Ghedin at the NIH is working on this.)
- Predicting mutations with combinatorics
 - Some mutations will only be beneficial if other mutations have occurred.
 - Since we have data from the beginning of the pandemic, we could examine which pairs of sequences show this behavior.
- Add useful features to Simdemic
 - Outbreak/bottleneck simulation (requires two populations that only occasionally interact).
 - Add additional useful metrics and visualizations.

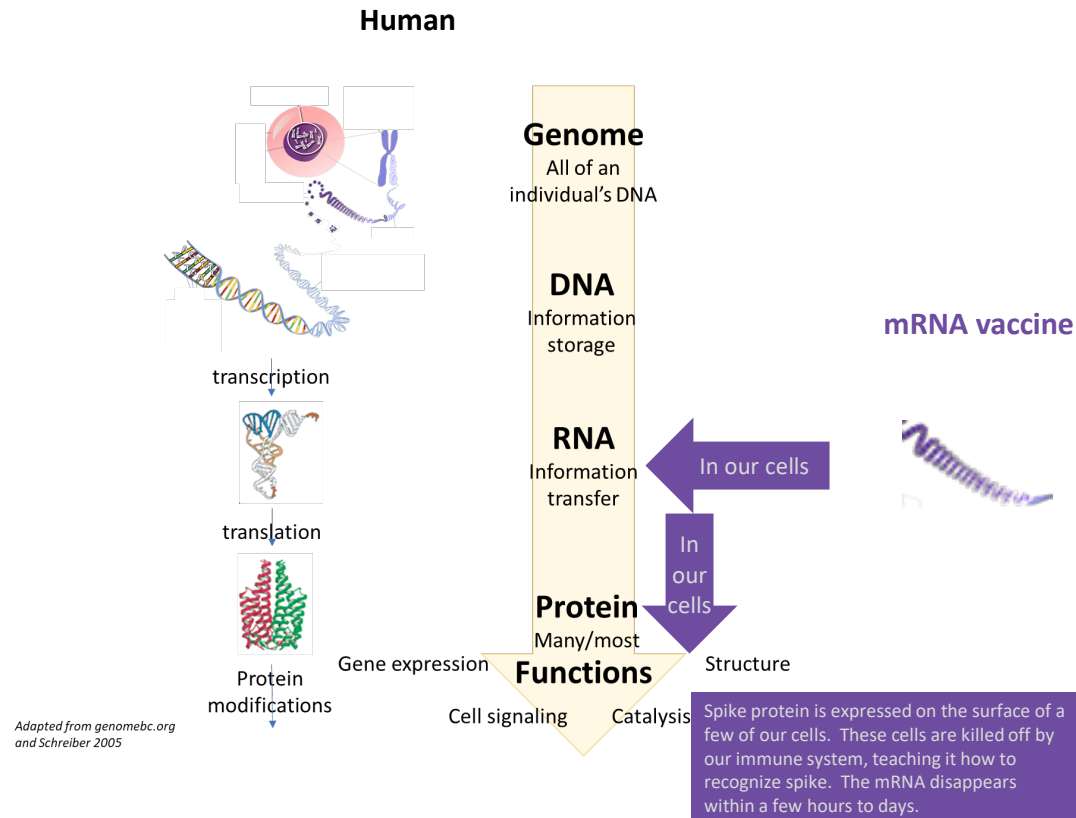
In the future, I would love to be able to predict the number of cases using genomic data. However, this will rely on some laboratory research. First, I would need to know the number of virus that are present in each patient, as well as the variability of this number. Second, I would need to know how much variance occurs within a patient. The sequences we are analyzing likely correspond to a consensus sequence of a group of variants that occur within a single patient rather than a single viron's genome. I have spoken with academics who mentioned working on some of these problems, so I am hopeful that I will have more information in the future to improve my predictive capabilities.

I would also like to predict mutations using combinatorics. Some mutations will only be beneficial once other mutations have occurred, meaning that the likelihood of mutation of a particular place in the genome is contingent on mutations first occurring in other places. For example, if a virus jumps into a species with a faster immune system than its current host, it might then mutate to reproduce faster, allowing it to evade the new immune system. This would be an example of the mutations coding for quicker replication becoming more likely once the mutations allowing the virus to jump species occurred. Another example is that one mutation might be disadvantageous since it would change the structure of the protein it codes for, but having a specific second mutation in combination with the first might allow the original protein structure to be preserved. These two mutations would only be beneficial when they co-occur. Since we have data from the beginning of the pandemic, we could determine whether there are any pairs of sequences which show this behavior. In addition to better helping us to predict future mutations, knowing which mutations rely on others might tell us more about the ultimate function these genomic regions code for.

Finally, I would like to work on a Simdemic 2.0 to add additional capabilities and continually improve it.

Backup slides

Background: SARS-CoV-2 genes and function



Background: Population Genetics Analysis

- It is possible to use genetic variation to estimate overall population size *in vertebrate species* (Wang 2005 Philosophical transactions of the Royal Society of London B, 360, 1459)
- Similar types of estimations have been attempted in viruses (Rousseau et al. 2017)
- Is it possible to apply these methods to the Covid-19 pandemic, and if not, what are the bottlenecks?

Reference: Rousseau, Elsa et al. 2017. “Estimating Virus Effective Population Size and Selection without Neutral Markers.” *PLOS Pathogens* 13 (11). <https://doi.org/10.1371/journal.ppat.1006702>

Computational challenges

- Source of data

- Large, continuously updating [Addressed in 2020](#)
- Permission-locked

- Exploratory analytic tools

- Memory-intensive [Addressed in 2020](#)
- Long-running
- Must be validated for viral datasets

- Tool and pipeline development

- Need to understand the implications of using tools for viruses; pandemics [This CRP](#)

Scientific challenges

- Sampling biases?

- Any analysis only tells you about the characteristics of your dataset. Whether this is generalizable depends on how representative the data are.

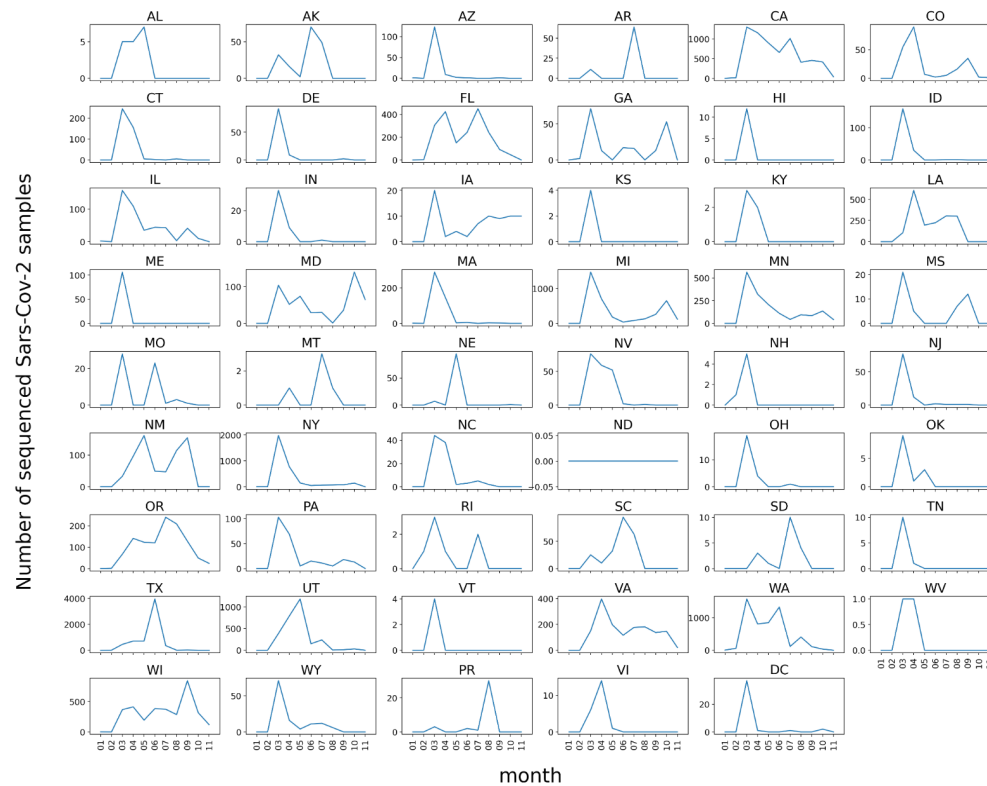
This CRP

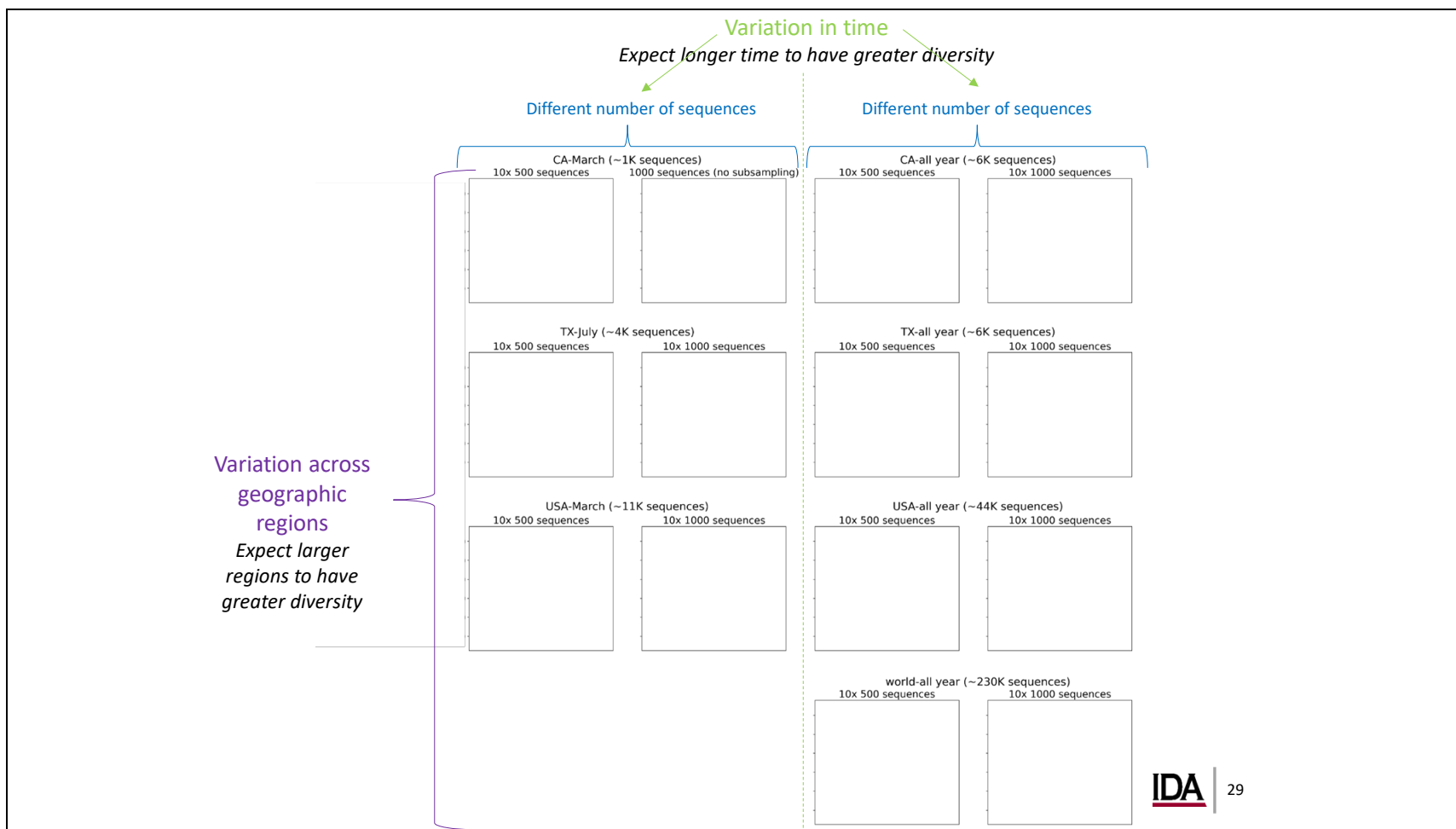
- Known unknowns/unknown unknowns due to this being the first data-rich pandemic of this type

- How many viruses per person?
- How much genetic variation of viruses within a person?
- Can we predict upcoming variants? This CRP
- **Staying open-minded for other surprises**

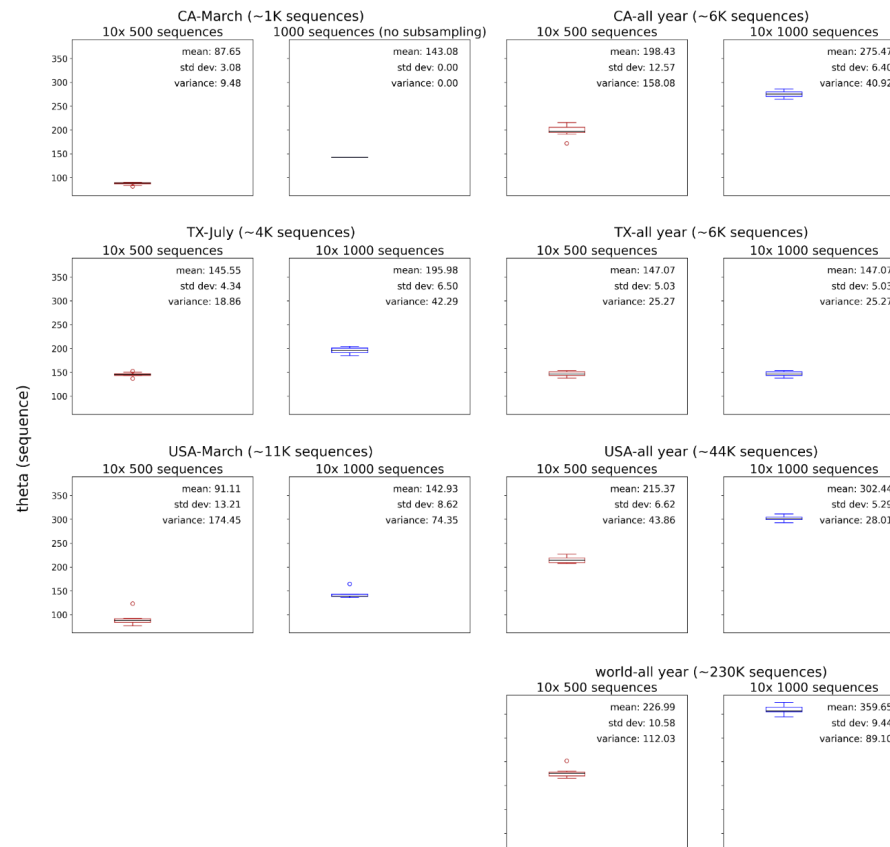
The U.S. sequenced relatively few SARS-CoV-2 samples

Y-axis changes from plot to plot

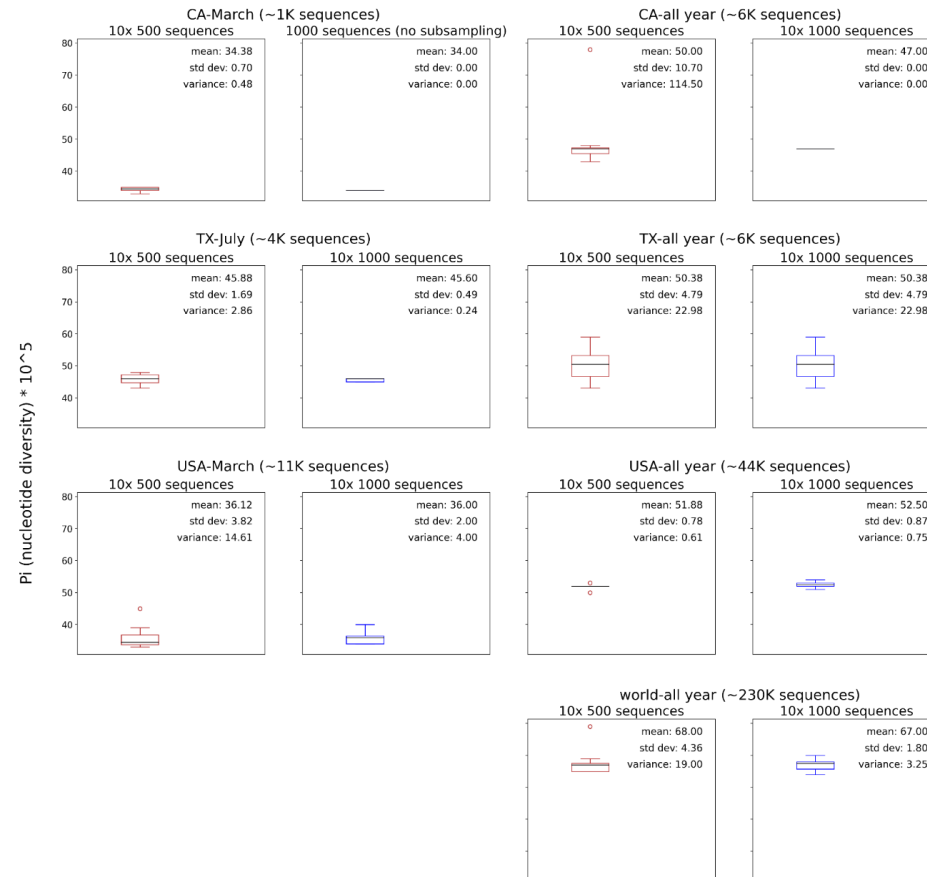




Observation: Theta is unstable to sample number



Observation: Pi is stable to sample number



Pi vs. Theta

AAAATAAAA
 ACAAAAAAA
 AATAAACAA
 AAAAAAGAAA
 AAAAAA
 AAAAAAAC
 AAACAAAAA
 GAAAAA
 AAAAAATA
 AAAAAATA

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 in at least one sample

This takes a long time to saturate

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

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

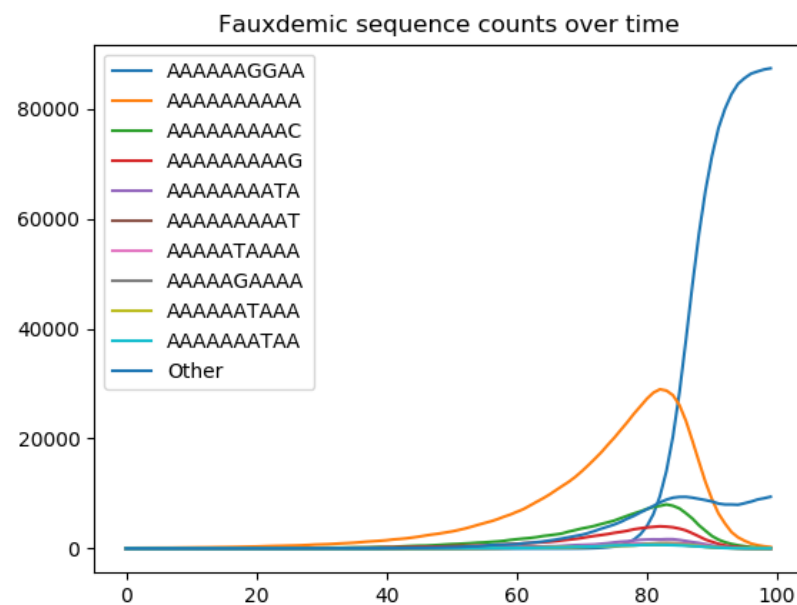
S: number of segregating sites (places that have a mutation somewhere in the population)
 n: number of samples
 μ: mutation rate (for the bounds of the whole region or genome in question)
 N: population size
 i (left): index of summation (which sample you are on)
 i, j (right): frequency of two (i^{th}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

Simdemic Population Simulator

- We want to generate data to test our methodologies
- Develop a program to simulate steps of a pandemic
- Given an initial sequence, population, and growth model, track evolution over time
- Produce genomes, their populations over time, and evolution history

Single Community Simdemic – Sequence Counts

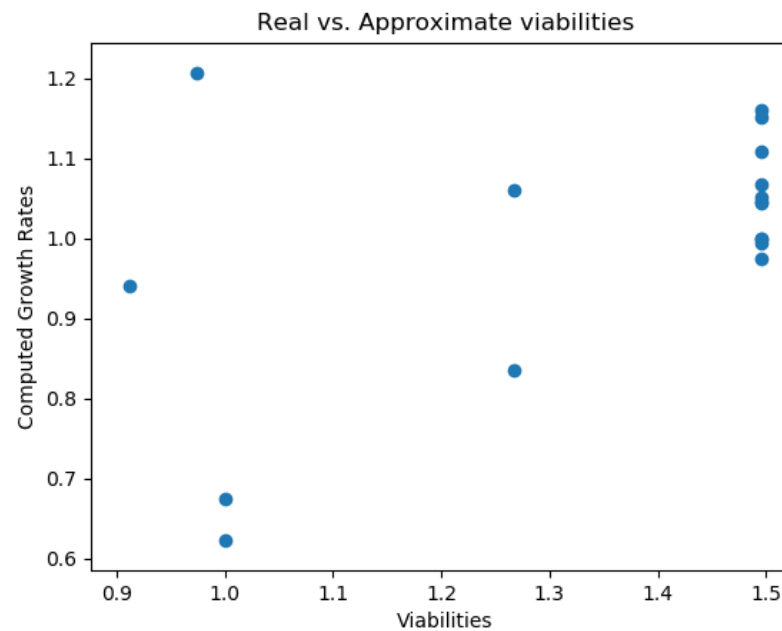
The original sequence, AAAAAAAAAA, is eventually dominated by the sequence AAAAAAGGAA



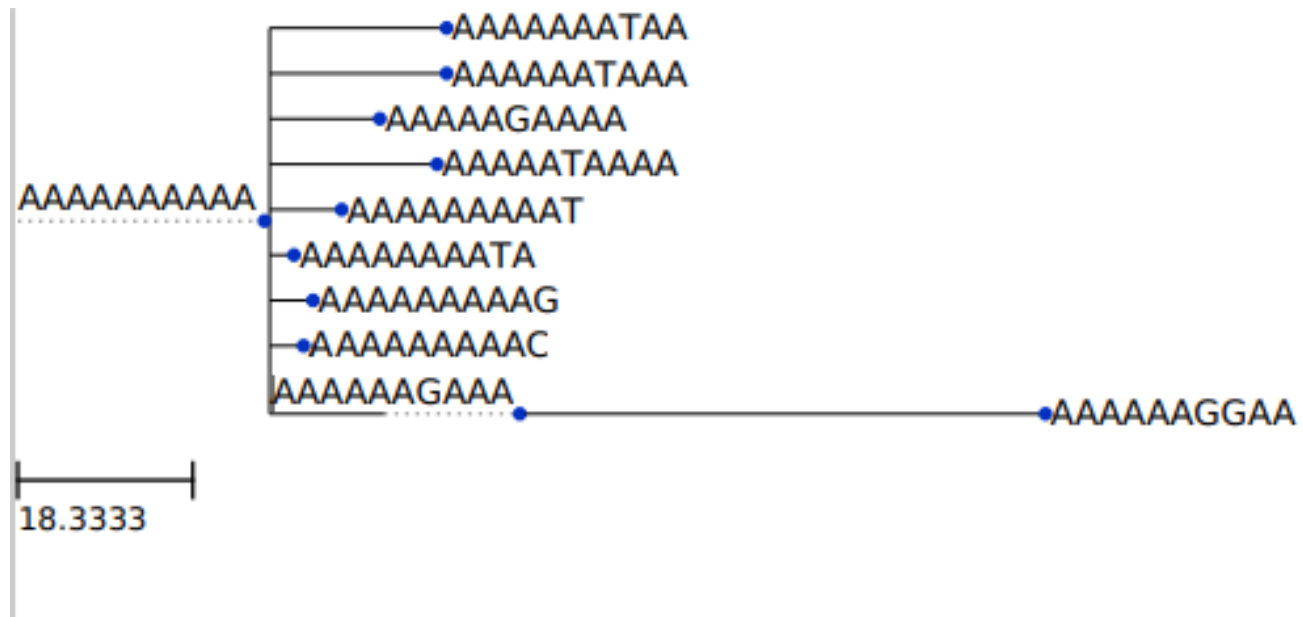
Computing Growth Rates Approximates Viability

Viability is a parameter built into each sequence that determines how likely a sequence is to infect people.

Computing growth rate in the case counts from one generation to the next allows us to approximate these viabilities



Single Simdemic Evolution Tree

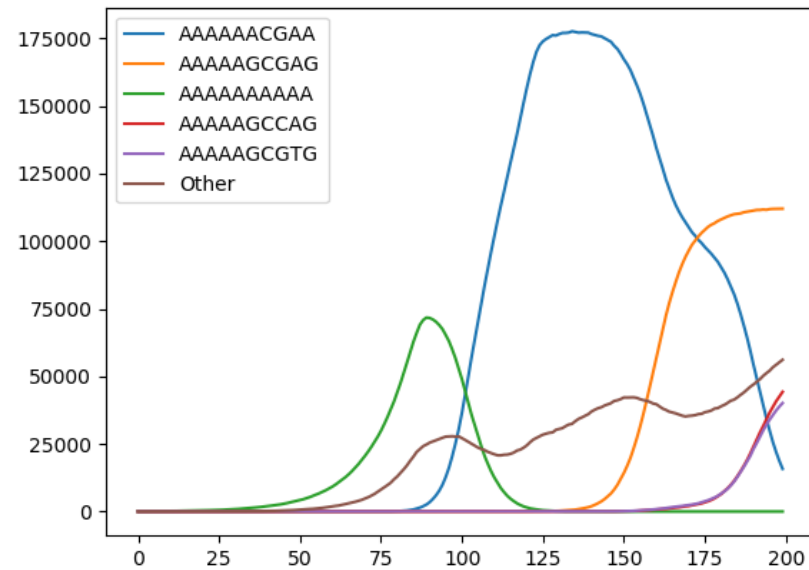


Multidemic

- We want to be able to simulate multiple communities with their own dynamics of case spread and counts of different sequences
- Solution: Run multiple simdemics in parallel
- Each timestep, each community transmits a predefined proportion of cases to other communities
- Some shared info, such as viability of each sequence and evolution history

Two Community Sequence Counts

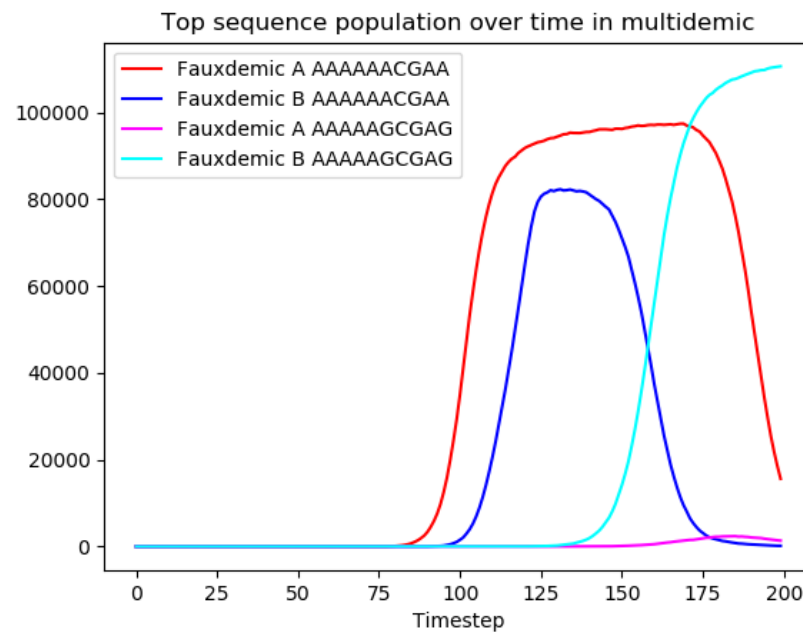
Longer timeframe allows for third set of
sequences to dominate at the end



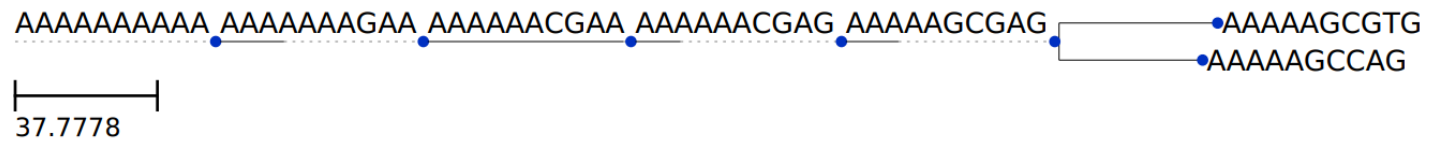
Top Sequences by Community

Simdemic A sees the first rise of AAAAAACGAA

Sequence AAAAAAGCGAG becomes dominant in Simdemic B, but not Simdemic A



Multidemic Evolution Tree

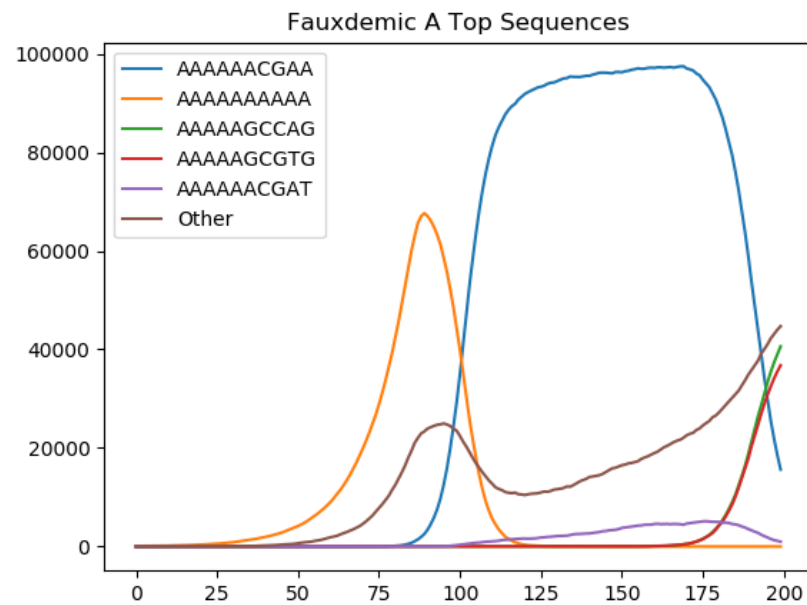


AAAAAGCGAG later mutates into two sequences, AAAAAGCGTG and AAAAAGCCAG

Simdemic A Most Populous Sequences

We see that these two sequences dominate AAAAAACGAA at the end of the simulation

Both have the same viability as their parent sequence, but they become dominant in Simdemic A because of random chance



Biased Sampling

- Want to know how bad a biased sampling is
- Sample n percent of sequences from one
- Compute loss:

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

S is the set of all sequences in the population, n_s is the count of sequence s in the sample, n is the sample size, and p_s is the true proportion of sequence s in the population.

- Create a test statistic by taking unbiased samples, then counting how many of those have higher loss

Biased Sampling Results

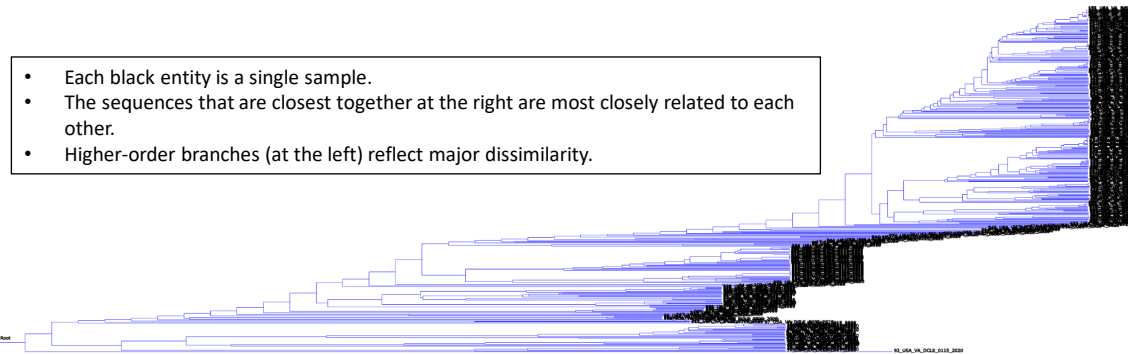
```
{'A': 0.1, 'B': 0.0} 0.0  
{'A': 0.090000000000000001, 'B': 0.01} 0.0  
{'A': 0.080000000000000002, 'B': 0.02} 0.0  
{'A': 0.070000000000000002, 'B': 0.03} 0.002  
{'A': 0.060000000000000002, 'B': 0.04} 0.123  
{'A': 0.050000000000000002, 'B': 0.05} 0.829  
{'A': 0.0400000000000000015, 'B': 0.0600000000000000005} 0.003  
{'A': 0.0300000000000000013, 'B': 0.07} 0.0  
{'A': 0.0200000000000000001, 'B': 0.08} 0.0  
{'A': 0.0100000000000000001, 'B': 0.09} 0.0  
[0.0, 0.0, 0.0, 0.002, 0.123, 0.829, 0.003, 0.0, 0.0, 0.0]
```

Past work: Variant analysis (2020)

- Viruses mutate relatively quickly
 - *SARS-Cov-2 is mutating at a rate of ~2 nucleotides a month (half the rate of influenza) (reviewed by Callaway 2020).*
- Viral mutations **may or may not** affect infectivity and pathogenicity
 - Tracking mutations may help predict which strains will become worrisome
- Analyzing strains can help us understand viral diversity, spread, etc.
- There is not a consistent methodology for identifying and declaring new SARS-CoV-2 strains

Reference: Callaway, Ewen. 2020. “The Coronavirus Is Mutating – Does It Matter?” *Nature* 585: 174-177 (2020).
<https://doi.org/10.1038/d41586-020-02544-6>.

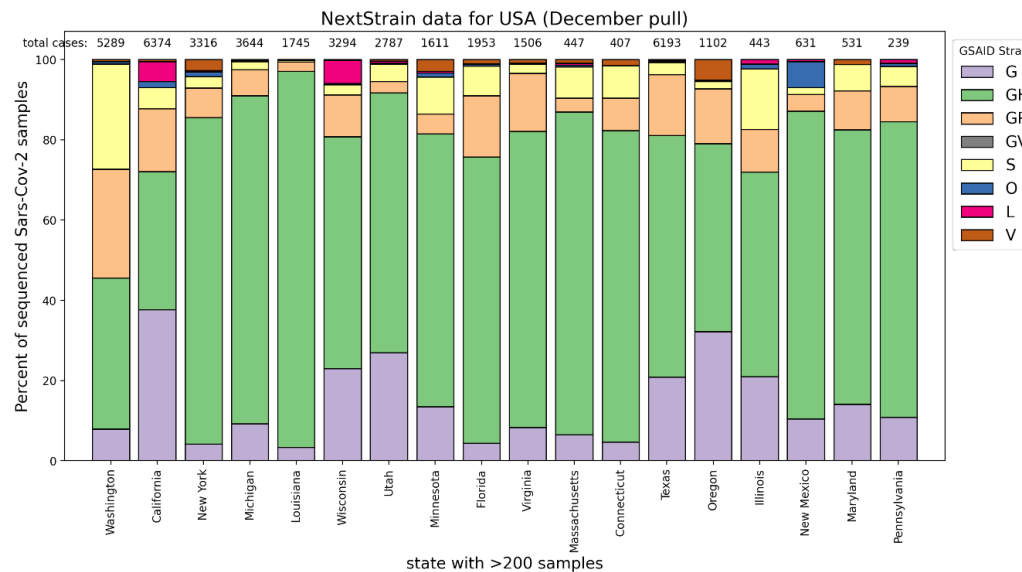
Preliminary Variant Analysis



- This phylogenetic tree shows the genetic (nucleotide) similarities between all SARS-Cov-2 samples that were sequenced in VA in spring 2020 (March 6 – May 29).
- Of the 433 samples sequenced in VA, there are 365 unique entities.

Strains vs. lineages

- Early on, NextStrain/GISAID defined several SARS-CoV-2 “strains” based on specific, studied mutations.

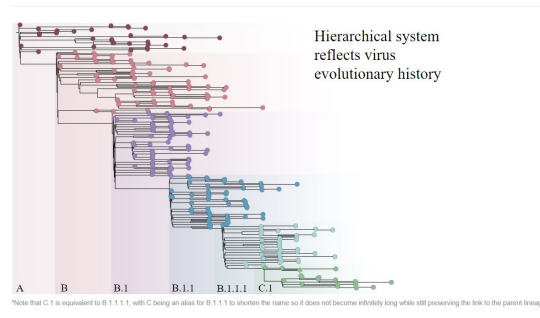


Strains vs. lineages

PANGO (Rambaut et al. 2020), is the algorithm behind the new lineages making the news e.g., “the UK variant”

Lineages reflect:

- 1) A virus's lineage
- 2) A local outbreak
- 3) A new mutation (in that lineage)



Lineages DO NOT reflect:

- 1) Whether a mutation is meaningful
- 2) Whether a mutation is unique (the same mutation can occur in different lineages)

Is this cluster of sequences a new lineage?

- Monophyly and cluster together on the global tree
- Consistent support values on the base node of the lineage (e.g. transfer bootstrap or ultrafast bootstrap)
- Epidemiological support (e.g. location, travel history)
- Introduction into a novel geographic region
- Evidence of circulation in that region (i.e. internal nodes within the lineage)
- A defining SNP

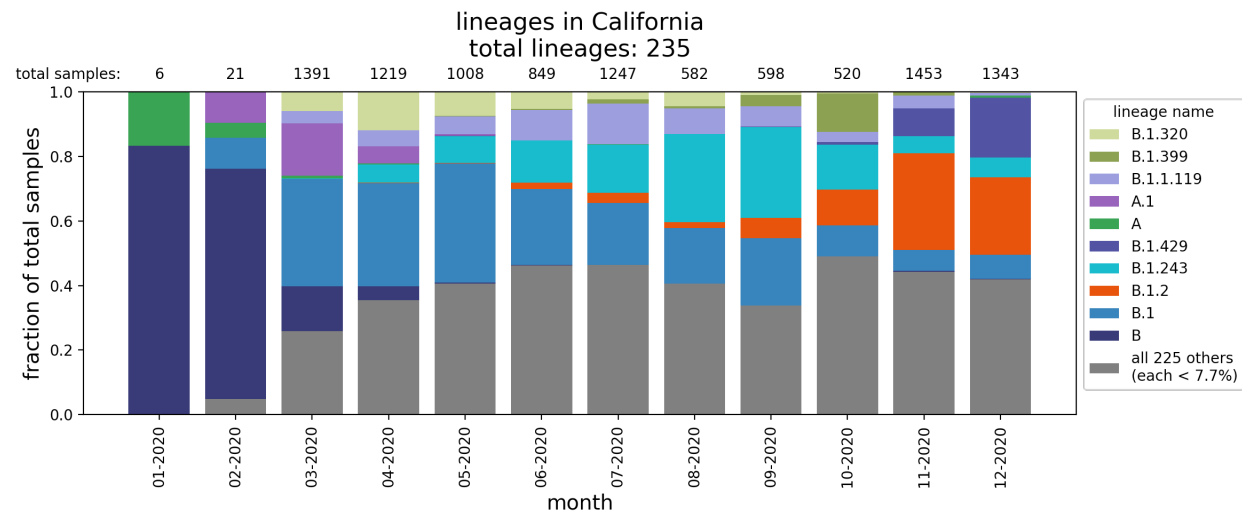
Reference: Rambaut, A. et al. 2020. “A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology.” *Nature Microbiology* 5(11), November 2020: 1403-1407. <https://doi.org/10.1038/s41564-020-0770-5>.

Strains vs. lineages

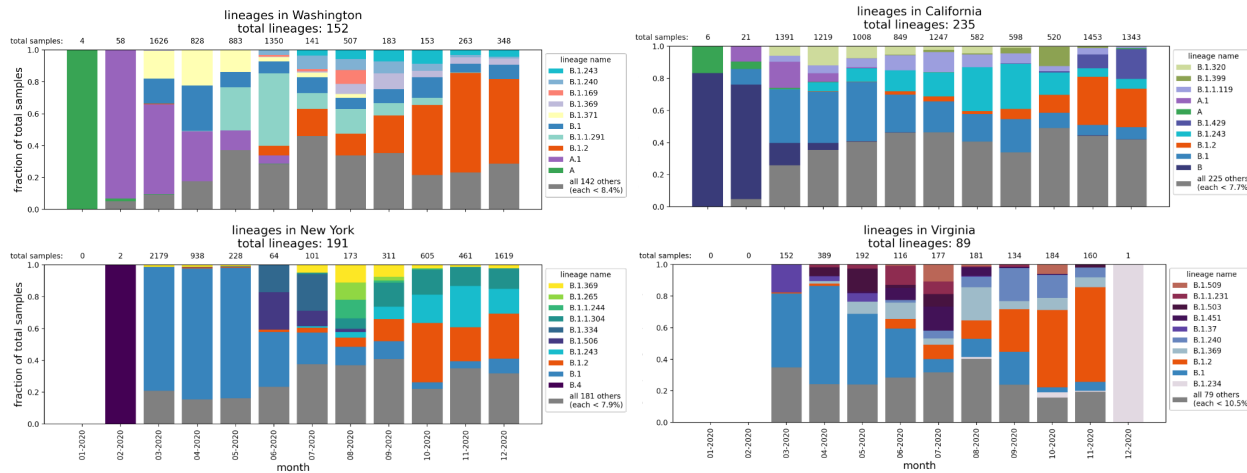
GISAID strains	PANGO lineages
Hypothesis-driven “I think this mutation is special, so it gets a name”	Agnostic “I see evidence of a new mutation spreading locally, so it gets a name”
One or two mutations define strain	Ancestry + mutation + local outbreak defines lineage
Tens total	Thousands total

Like the field, we have chosen to focus on PANGO lineages

Lineages over time in the U.S.



SARS-CoV-2 lineages have ebbed and flowed over the course of the pandemic in the U.S.



- It is easy to determine what a virus's genetic sequence is. It is more difficult to tell what the sequence means.
Lineage names reflect sequence, not meaning
- Things that may cause a lineage to become predominant:
 - Founder effect – the reduction in genetic variation when a small subset of a large population is used to establish a new population
 - One or repeated superspreader events with that lineage
 - Small number statistics/undersampling/biased sampling
 - A mutation with a biological mechanism that facilitates higher infectivity/spread

Past analysis: SARS-CoV-2 variant analysis conclusion (2020)

- Variants are numerous. This is okay and expected.
- PANGO provides a useful syntax for identifying emerging variants. Variants can be tracked to determine *with follow-up experiments* whether they are concerning.
- Variants come and go in the U.S. as they have in different countries ***and have for the entire pandemic.*** *We weren't looking for variants in the U.S. for most of the pandemic like countries with named lineages were.*

Past analysis: SARS-CoV-2 variant analysis conclusion (2020)

- **Public concern about variants is a byproduct of watching science in progress**
 - The general public – and people spreading the news – are not used to watching science in progress
 - The scientists doing the research are not used to having the public so interested in their work and are not used to explaining the process
 - Watching a typically messy scientific process leads to odd conclusions from the public

Behavior	Clarification
Naming variants has completely changed during the pandemic	Science develops organically, especially on the cutting edge. This does not mean the scientific community is incompetent or that the problem is intractable
Many lineages are being tracked	Tracking a variant/mutation does not necessarily mean anything scary about the effects of the mutation
Agnostic categories are getting names: "B.1.1.17" "the UK variant"	A name does not confer danger

REPORT DOCUMENTATION PAGE					<i>Form Approved OMB No. 0704-0188</i>	
<small>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</small>						
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)	