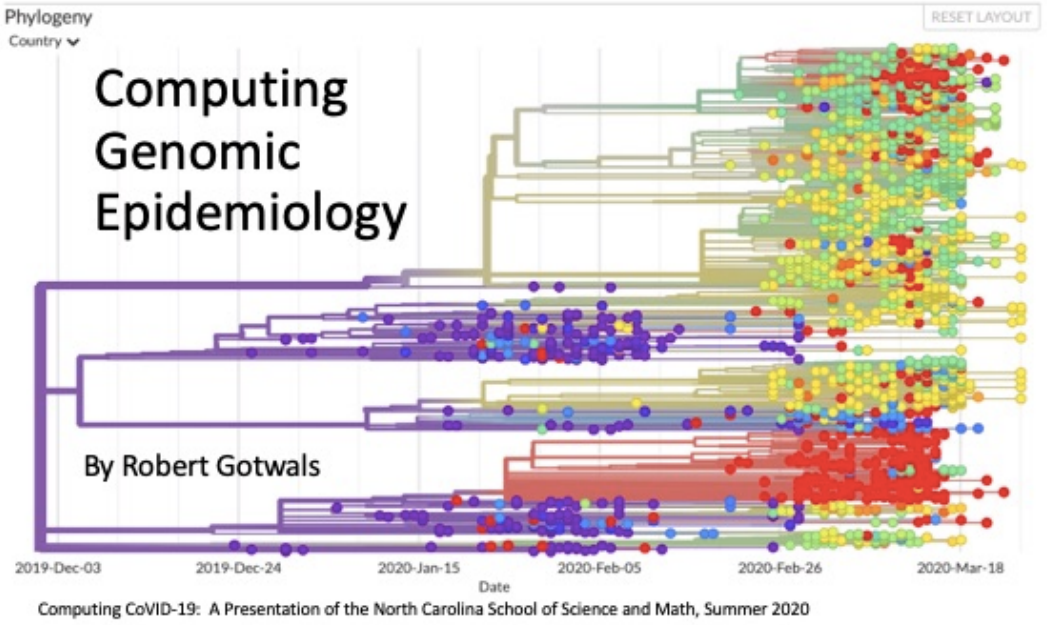

Computing CoVID-19: Summer 2020

Computing Genomic Epidemiology

Developer:
Robert Gotwals
May 21. 2020

1

Introduction

Virus, like all organisms, mutate over time. The description below, from the journal article "Mechanisms of viral mutation" ([1]) captures the essence of viral mutation. SARS-CoV-2 is, as you should know by now, an RNA virus:

> The remarkable capacity of some viruses to adapt to new hosts and environments is highly dependent on their ability to generate de novo (from scratch) diversity in a short period of time. Rates of spontaneous mutation vary amply among viruses. RNA viruses mutate faster than DNA viruses, single-stranded viruses mutate faster than double-strand virus, and genome size appears to correlate negatively with mutation rate.

Figure 1 shows a sample timeline of how viruses have evolved over time. In this figure, *paleoviruses* are those viruses that have gone extinct. The lab "mya" refers to millions of years ago. The term *hominoids* refers to humans, their fossil ancestors, and other anthropoid apes, which includes primates belonging to the family *Pongidae*, which includes the chimpanzee, bonobo, gorilla, and orangutan.

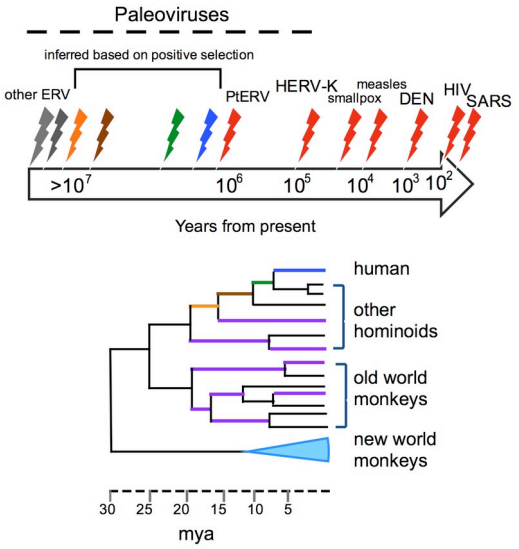

Figure 1: Viral mutations [3]

Notice from the phylogenetic tree that there are three major *clades*: humans and other hominoids; Old World monkeys (baboons and macaques); and New World monkeys (lemurs,spider monkeys, howler monkeys, etc.). Note that viral mutations for New World monkeys is relatively recent in evolutionary scale, if we are considering the three major clades on the phylogenetic tree.

By now you should have a fundamental knowledge of phylogenetic trees, and Figure 2 is provided as a reminder. To construct a phylogenetic tree, you take the sequences (either nucleotide/base pairs like A,C,T, and G; or the amino acid sequences) and align them. Then, based on the alignment, and using an algorithm such as the UPGMA (unweighted pair group method with arithmetic mean), you construct the phylogenetic tree. The ones shown in both figures are called *rooted* trees, because the root – on the far left – represents a common ancestor. In this lab, you will also see examples of unrooted, radial, and clock trees.

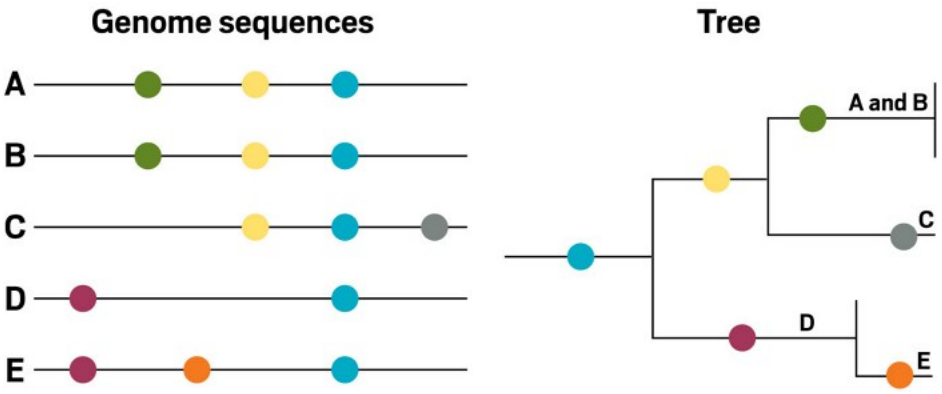

Figure 2: Genetic epidemiology [2]

The creation of a tree such as shown in Figure 2 is a standard method of performing what is known as *genomic epidemiology*, which is the focus of this lab. Figure 3 shows a generalized flowchart for how genomic epidemiology is conducted with bacterial studies, and the procedure for viruses is identical. Note that the typical products of a genomic epidemiology study is the creation of a phylogenetic tree and a map showing how the pathogen (whether bacterial or viral is inconsequential) across a region or across the world, as is the case with SARS-CoV-2.

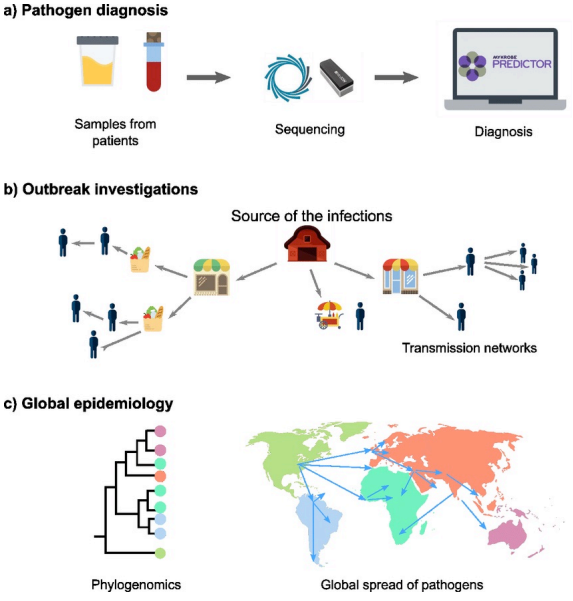

Figure 3: Genetic epidemiology of bacterial pathogens [5]

For this lab, you will be doing genomic epidemiology using a web-based tool called NextStrain ([6]). This tool uses data mostly from the GISAID Initiative ([7]), and is written using JavaScript and Python. This code is downloadable for you to conduct your own experiments,

but for this lab we will use the web-based online version located at https://nextstrain.org/.

2

Student Activity

NOTE! The majority of the steps for the activity will be demonstrated in the webinar. These instructions are meant only as short reminders of the steps you need to take to effectively modify the lead drug and test the new compounds!

Use the following questions as guides to using NextStrain to explore the SARS-CoV-2 viral evolution. Respond to these questions in the Canvas quiz for this lab!

For all of these questions, your data set selection should be "ncov" on the selection panel on the left-hand side of the NextStrain page. NOTE! There are different display options. I prefer the "Grid" panel options, but you will get better resolution with the "Full" panel option. Find the one that works best for you!

You will answer these questions using the Canvas lab/quiz page. They are provided here for your convenience.

1. Startup: go to https://nextstrain.org/ncov/global, the latest global analysis page of NextStrain.

2. Scenario 1: for this question, make sure your selection is on "global".

   (a) For a date range between Dec 2019 and January 2021, how many different viral genomes were sequenced?

   (b) Scroll down to the "Diversity" window under the map. Click at the very end of the diversity chart, locate Codon 24 in ORF (open reading frame) 10. From this, you should be able to determine how many base pairs (nucleotides, or As, Cs, Ts, and Gs) make up the sequence for this virus. So, how long is the SARS-CoV-2 virus in terms of base pairs/nucleotides? NOTE! It's a little tricky to hover the mouse over the right spot, but with a little patience, you should be able to do it!

(c) Look at the map. Hit the "Play" button and watch the transmission of the virus over time. Based on this, most of the transmissions in the US came from which continent – Asia (China) or Europe? Make sure you watch the animation for the entire time!

(d) Using the rectangular phylogenetic tree, click on the start of the tree on the very left-hand side of the tree. This should indicate December 24, 2019. What was the country of origin for this virus?

(e) Click on "unrooted" tree. Notice a lot of genomes branching off upwards and to the right. Click on the very light gray branch, near the root, the one with 2171 genomes, almost 70% of all the genomes sequenced. From what country did these come from (at least with 92% confidence!)

(f) Click on the "clock" tree option. How many new mutations ("subs" or substitutions) are being generated per year?

(g) Under "Branch Length", click on the "Divergence" tab. Notice that the x-axis is now in number of mutations, and the lengths of the lines are in divergence units, or number of mutations. Find the strain farthest to the right. How many mutations does that represent, and where (geographically) is that strain located?

3. Scenario 2: Clade analysis. As you should have learned from the earlier studies of phylogenetic trees, a clade is a grouping of related organisms. In Figure 4, there are three major clades shown for this hypothetical tree. These might correspond to mammals, vertebrates and invertebrates if we were describing animal taxonomies.

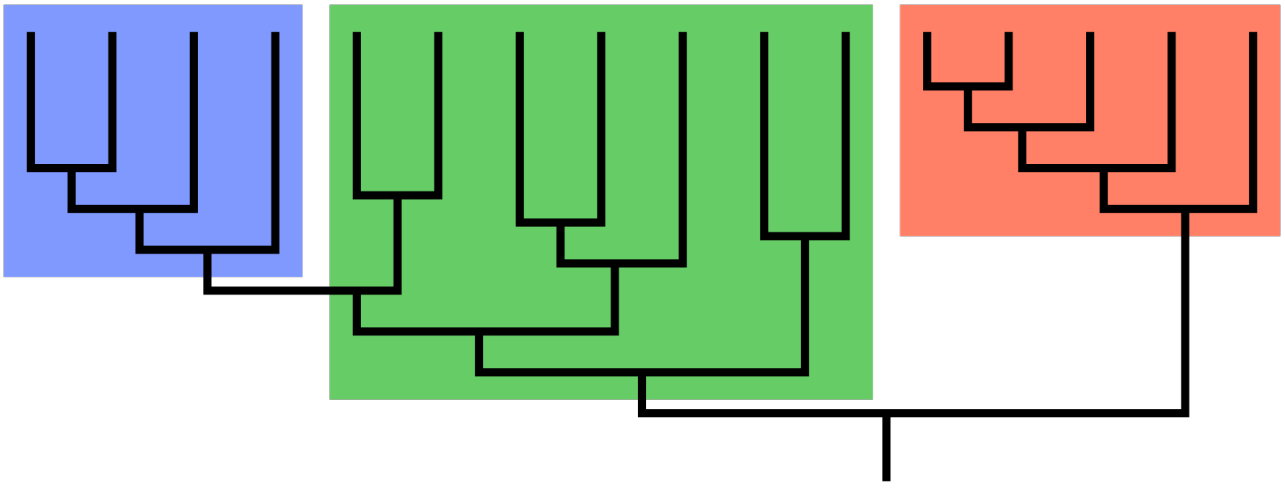

Figure 4: Example of clades [4]

(a) Genomic epidemiologists have divided the SARS-2-CoV-2 virus into how many clades?

(b) Color the tree by Genotype, using the S protein at position 614. What amino acid has changed at position 614 on this protein? Figure 5 provides a one-letter amino acid chart for your convenience!

(c) There is a new strain of virus, recently (December 2020-ish) on the spike protein at position 501. Color the tree by Genotype, using the S protein at position 501. What amino acid has changed at position 501 on this protein?

Amino acid	Three letter symbol	One letter symbol*
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Figure 5: One letter amino acid chart

4. Scenario 3: Searching by Genotype Use the "Color By" option and select "Genotype". If you recall, earlier we looked at a mutation on the S (spike) protein, and located at position 614. Do a genotype search by the S protein, and enter 614. Set the branch labels by clade.

   (a) You should note two major clades, one in blue and one in yellow/orange. What is the amino acid for the blue clade? What is the amino acid at the yellow clade?

References

[1] Sanjuán, Rafael, and Pilar Domingo-Calap. "Mechanisms of viral mutation." Cellular and molecular life sciences : CMLS vol. 73,23 (2016): 4433-4448. doi:10.1007/s00018-016-2299-6

[2] https://cen.acs.org/biological-chemistry/genomics/genomic-epidemiology-tracking-spread-COVID/98/i17, accessed May 24, 2020.

[3] https://en.wikipedia.org/wiki/Viral_evolution, accessed May 24, 2020

[4] https://en.wikipedia.org/wiki/Clade, accessed May 24, 2020

[5] Bacigalupe, Rodrigo. (2017). Population Genomic Analysis of Bacterial Pathogen Niche Adaptation.

[6] Hadfield, James et al. "Nextstrain: real-time tracking of pathogen evolution." Bioinformatics (Oxford, England) vol. 34,23 (2018): 4121-4123. doi:10.1093/bioinformatics/bty407

[7] Elbe, S., and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges, 1:33-46. doi:10.1002/gch2.1018 PMCID: 31565258