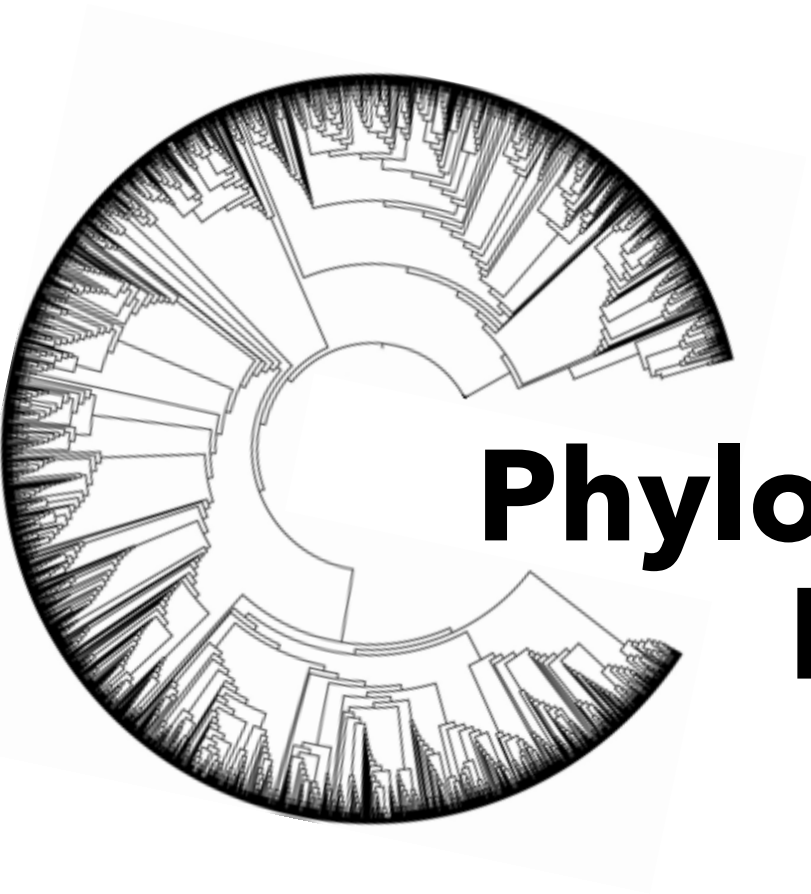
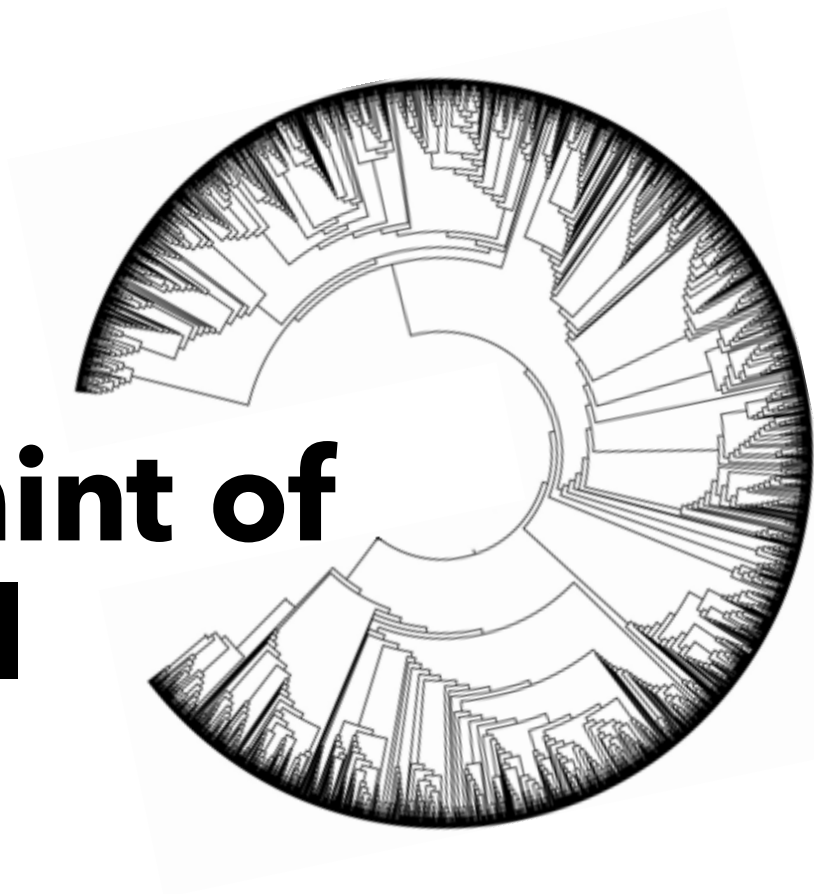


Phylogeny for the Faint of Heart: a Tutorial

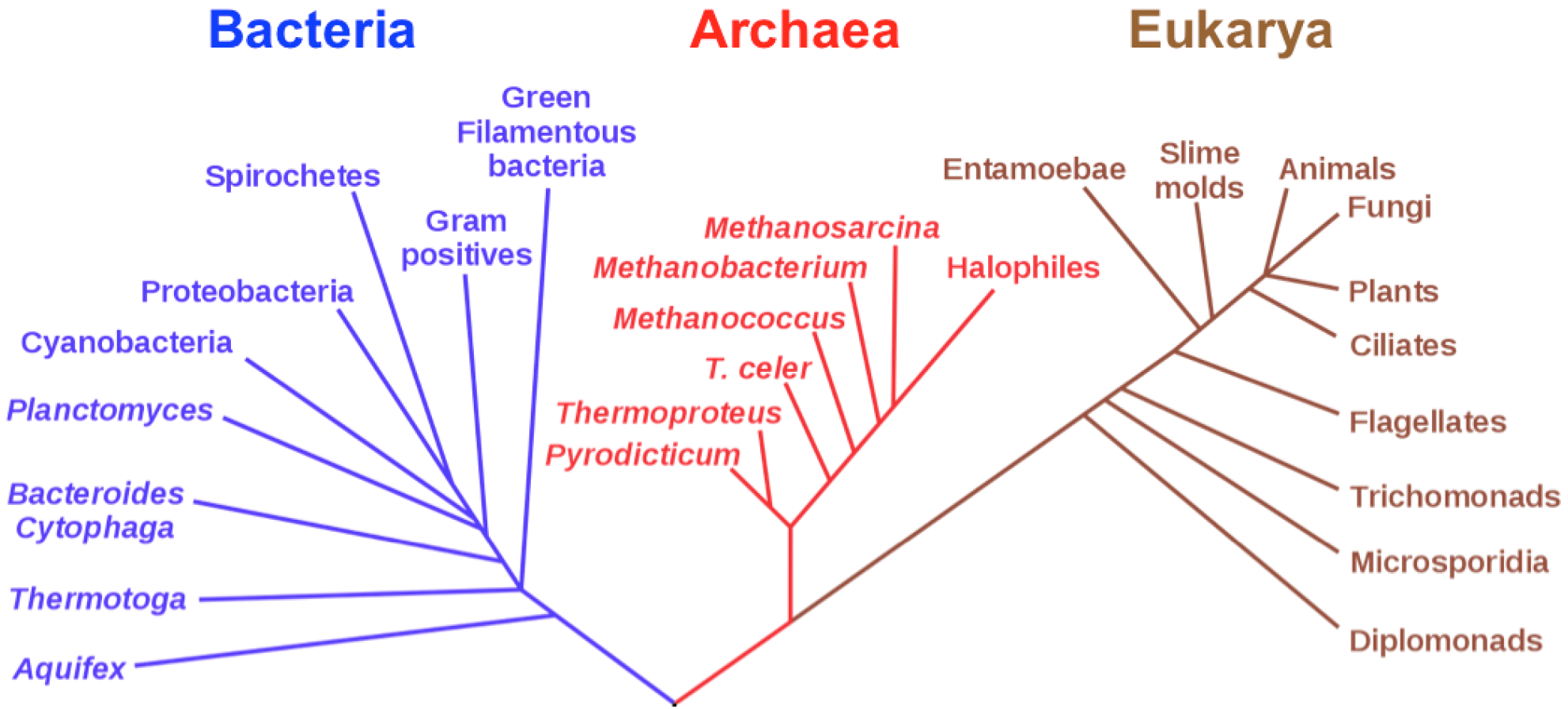


Sandra L. Baldauf
(2003)

Department of Biology, York University

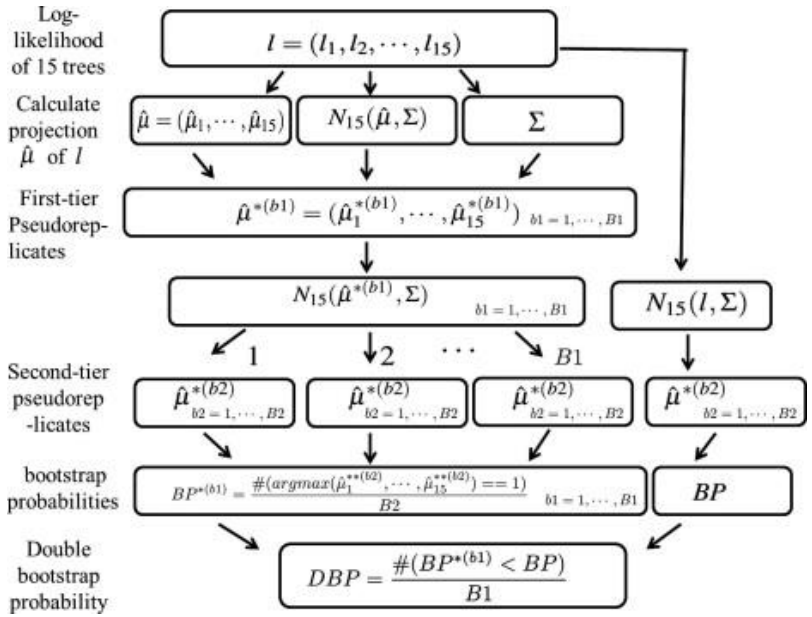
Presented by
Jenna Gold
and Paul Lyons

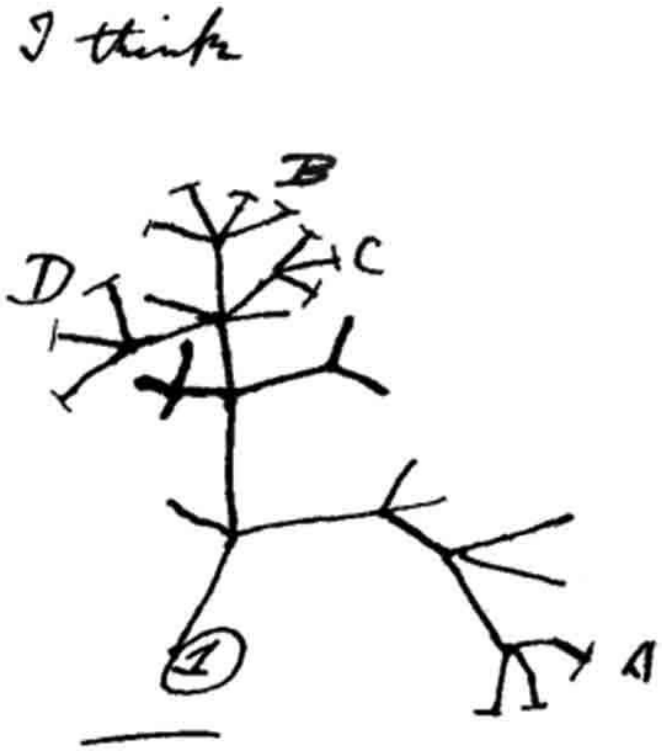
What is phylogenetics?

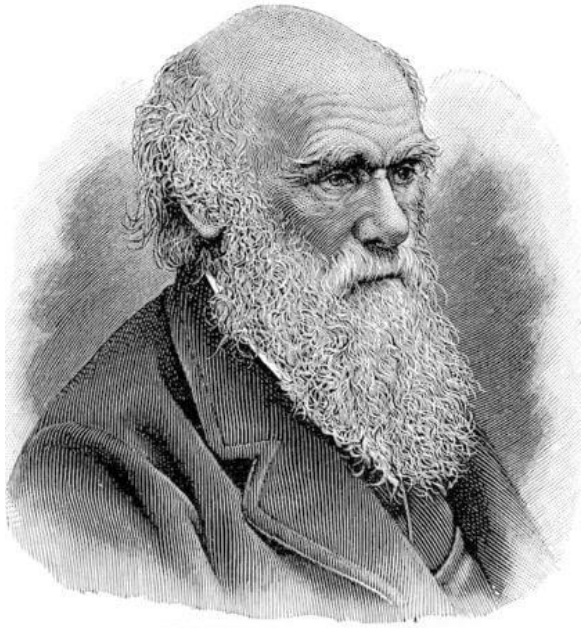


The comparison of DNA sequences to estimate the evolutionary past

Where did modern phylogenetics come from?

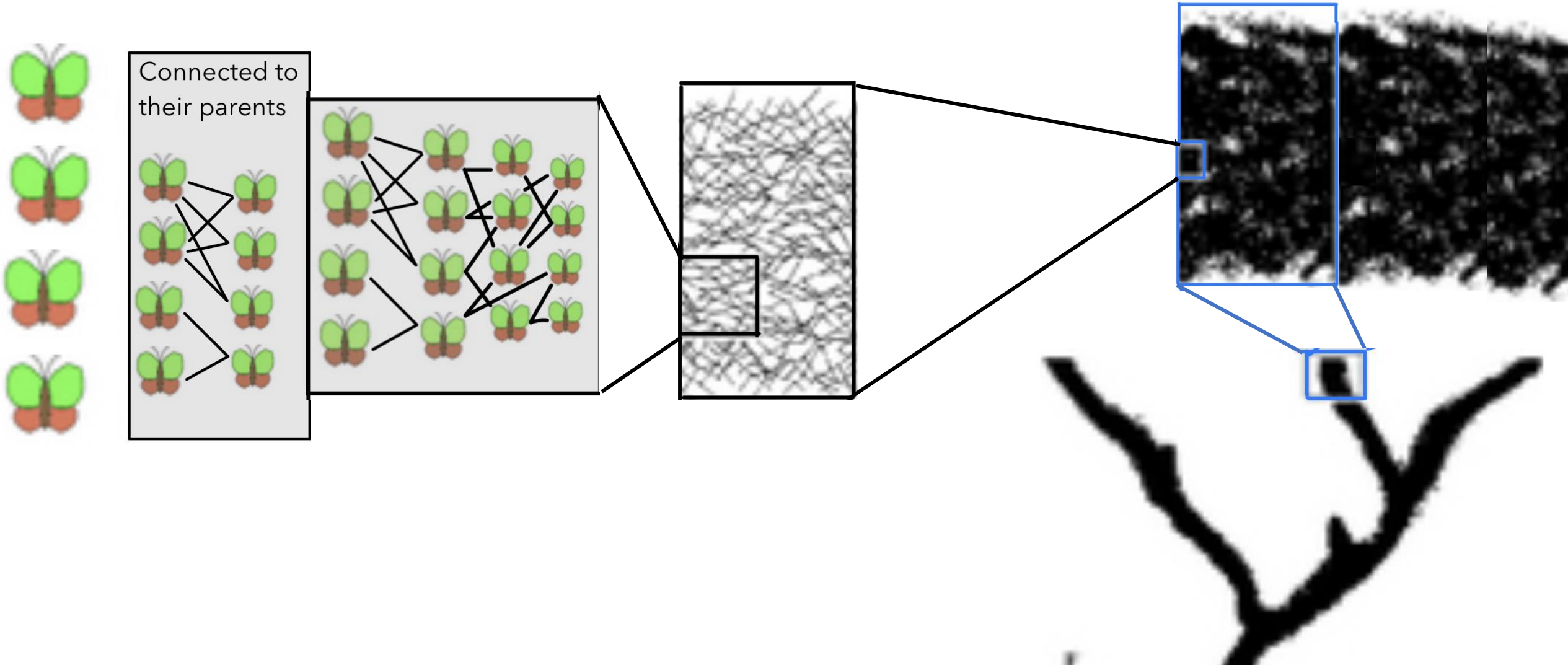




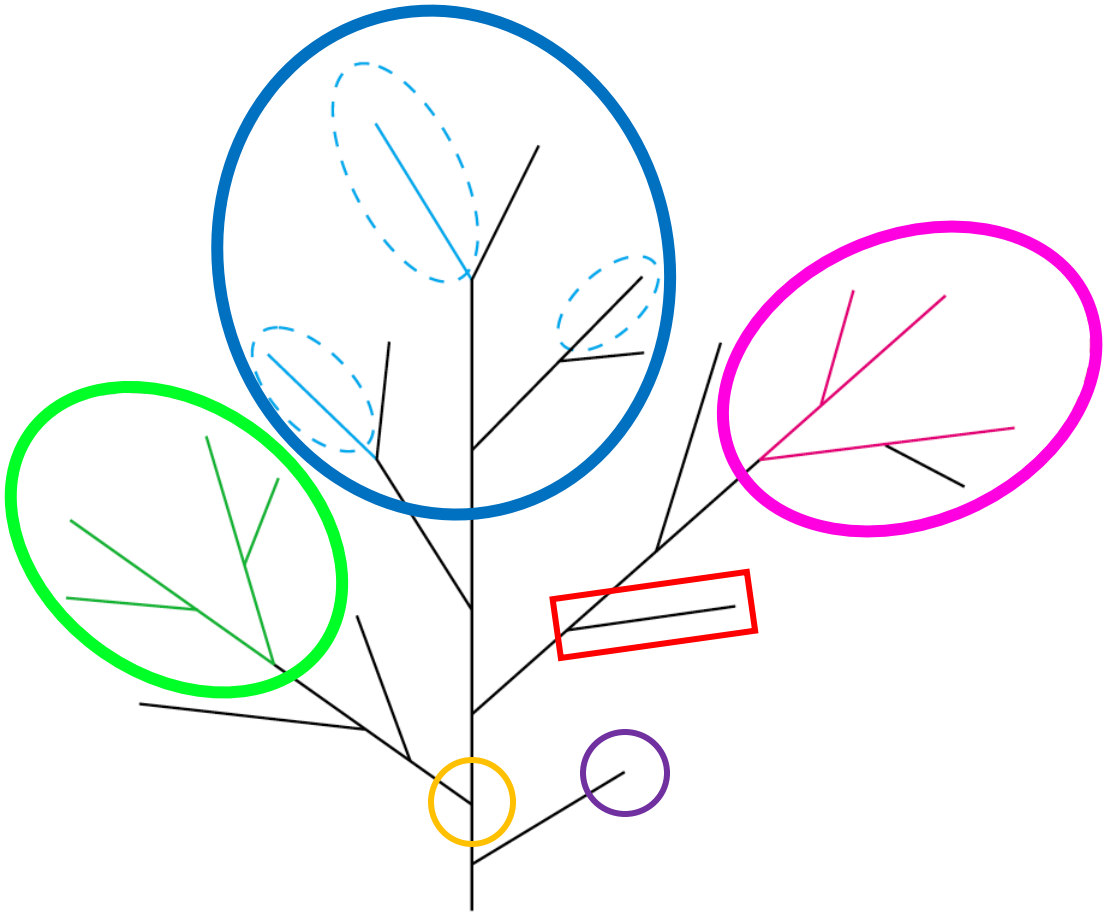


Charles Darwin.

How do pedigrees fit into phylogenetic branching?



How do we reference a tree?



TRENDS in Genetics

Monophyletic Group

Paraphyletic Group

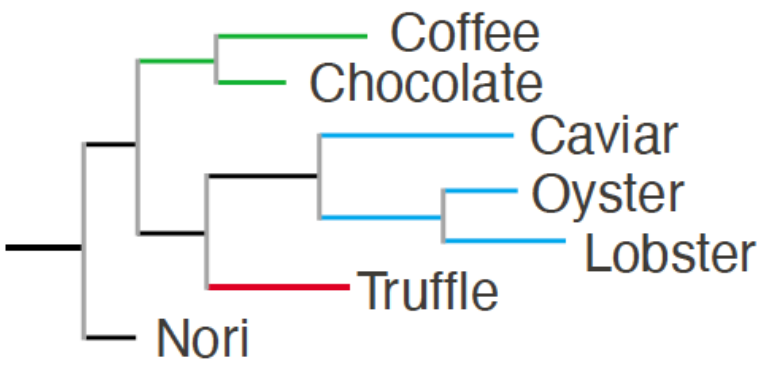
Polyphyletic Group

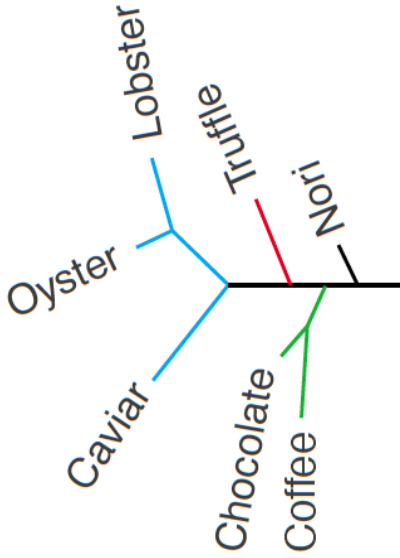
Branch

Internal Node

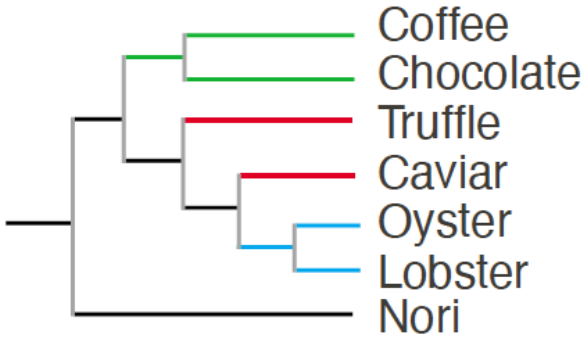
Terminal Node

Is there any format variation among phylogeny trees?

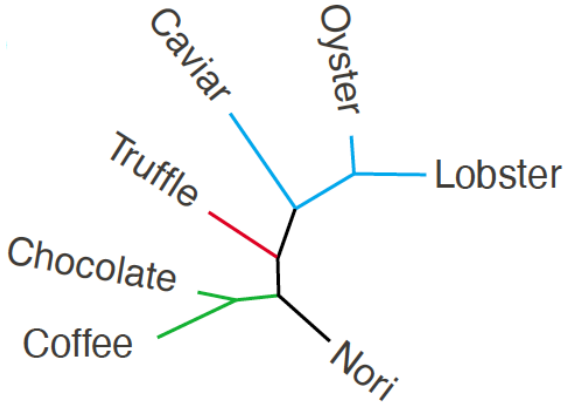




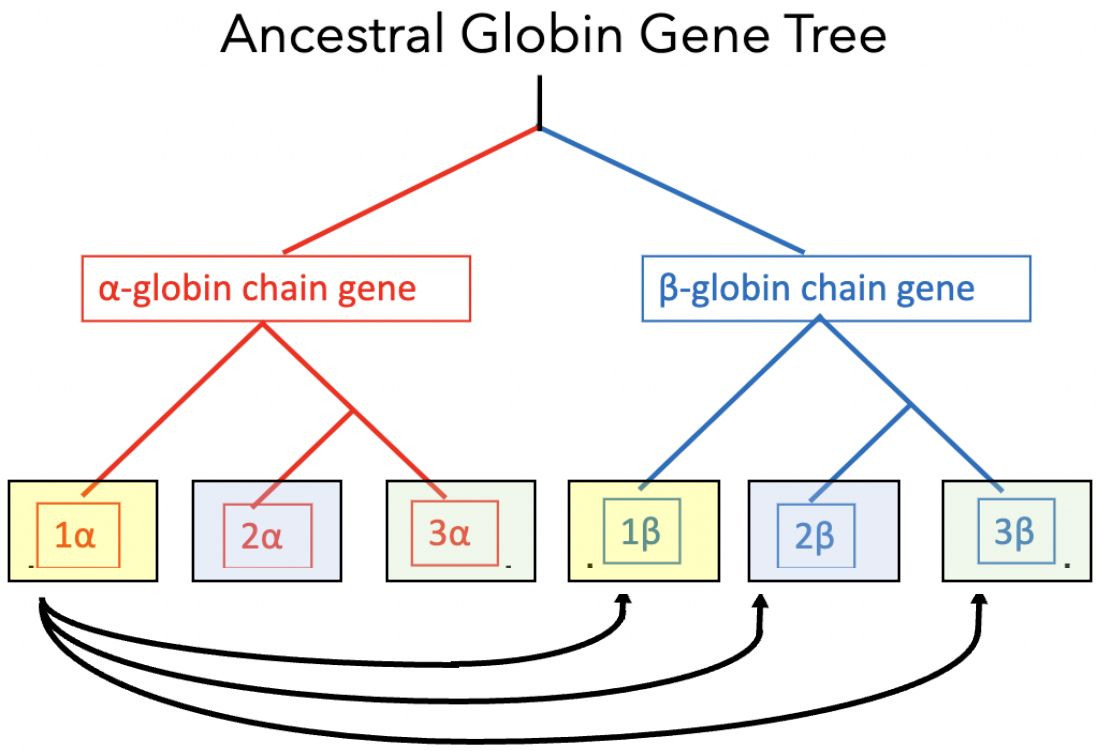
Cladograms:

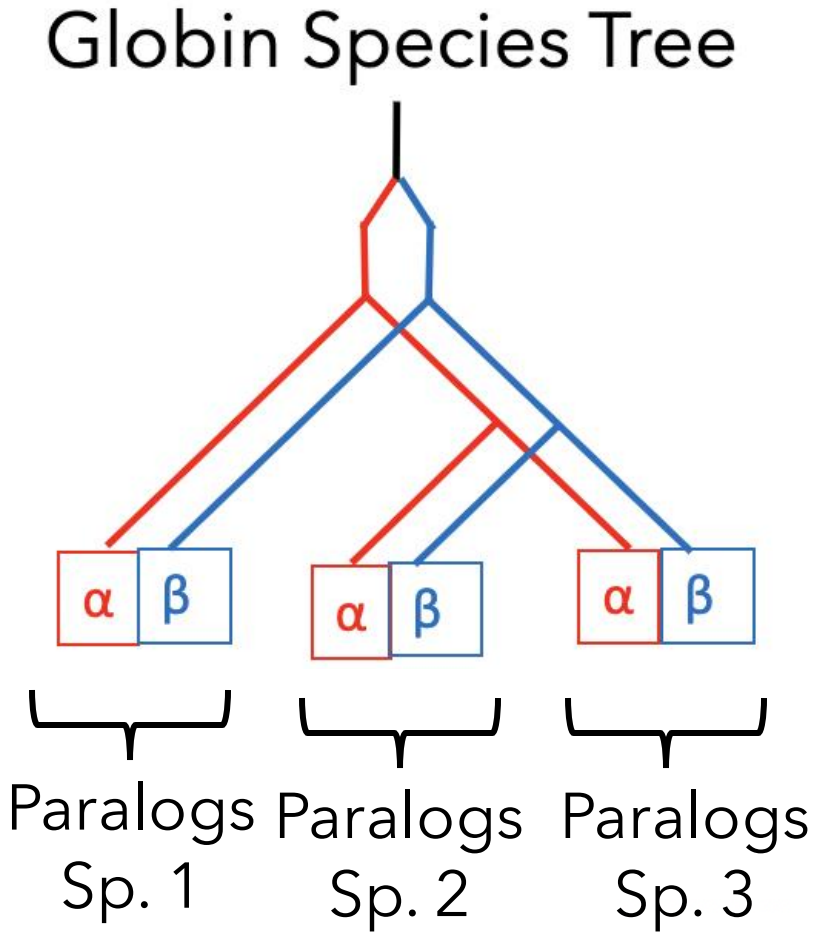


Rooting:



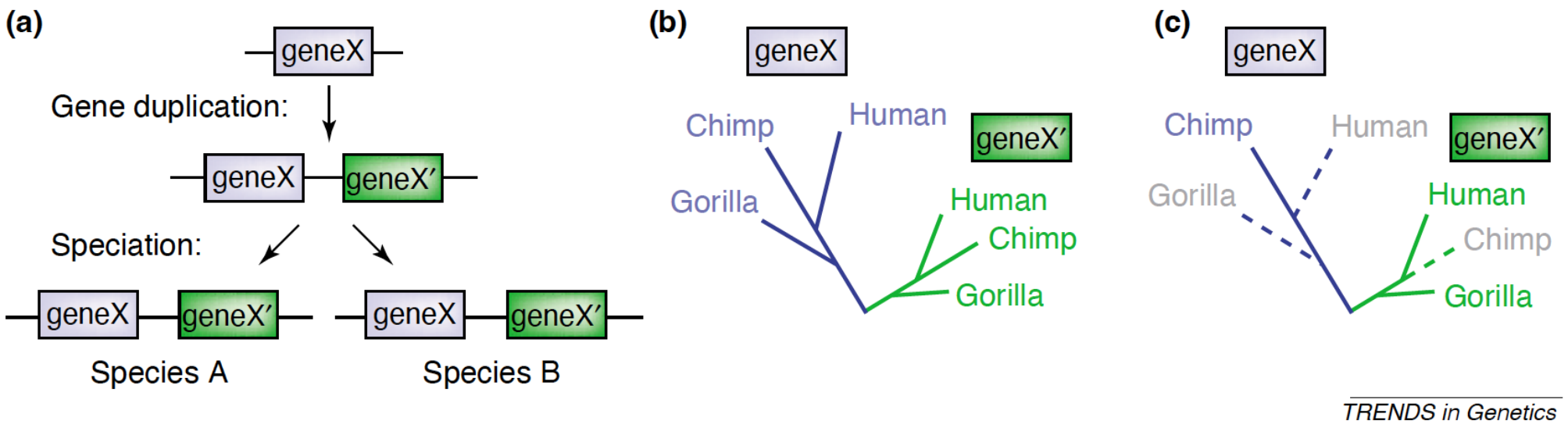
How does gene homology influence the structure of a tree?





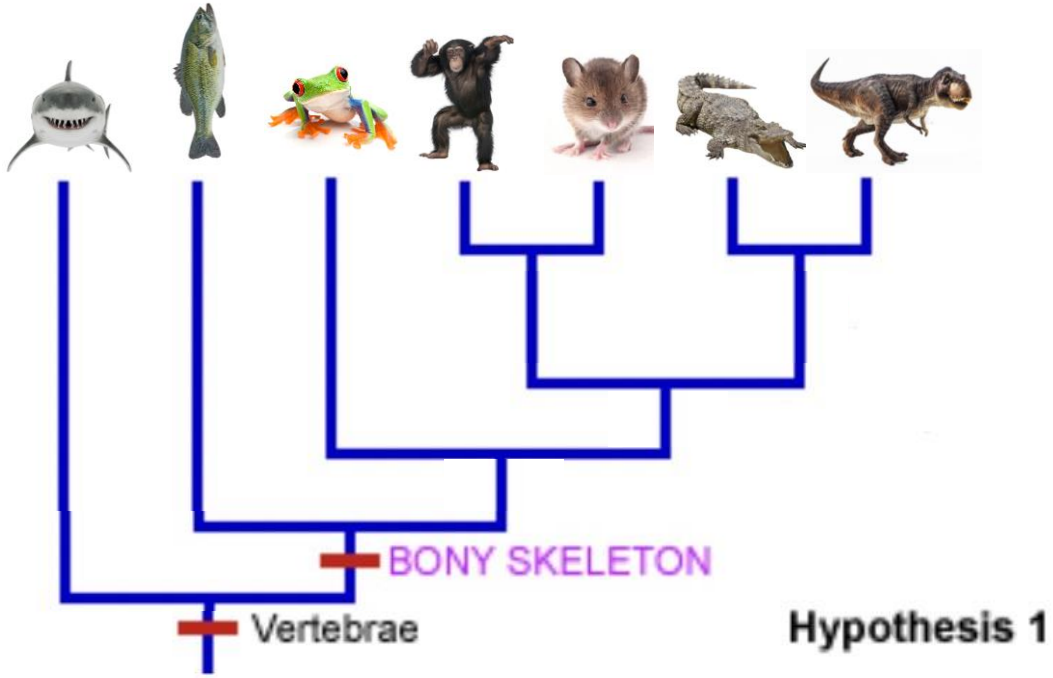
*Paralogous relationships can also occur between species

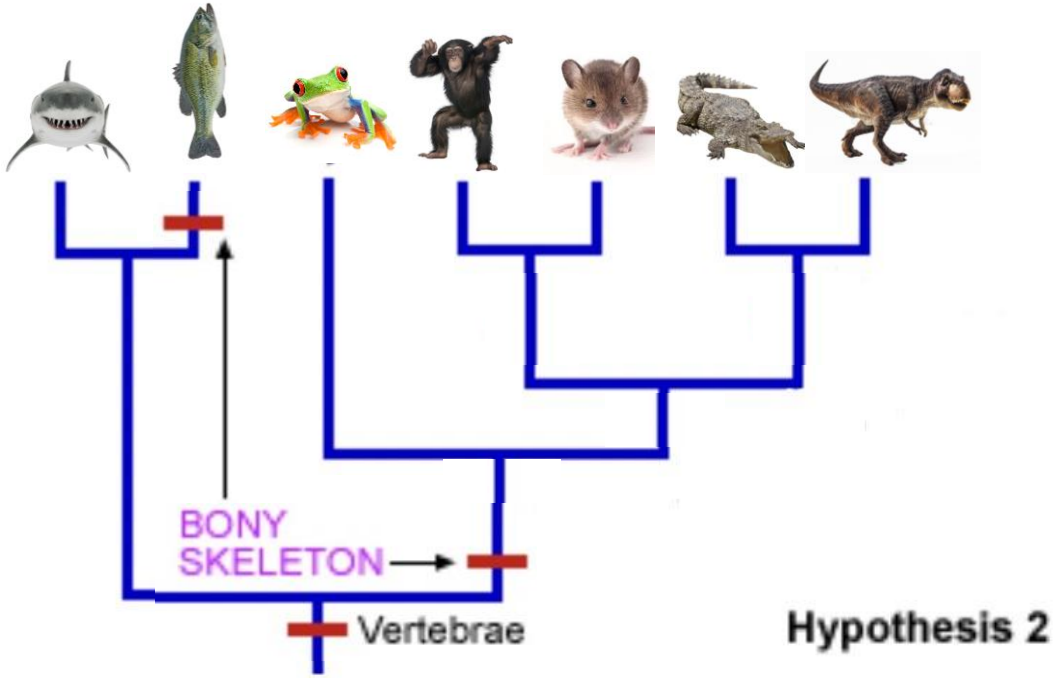
What if a duplicate paralog is deleted?



Missing genetic information can easily misrepresent true lineages

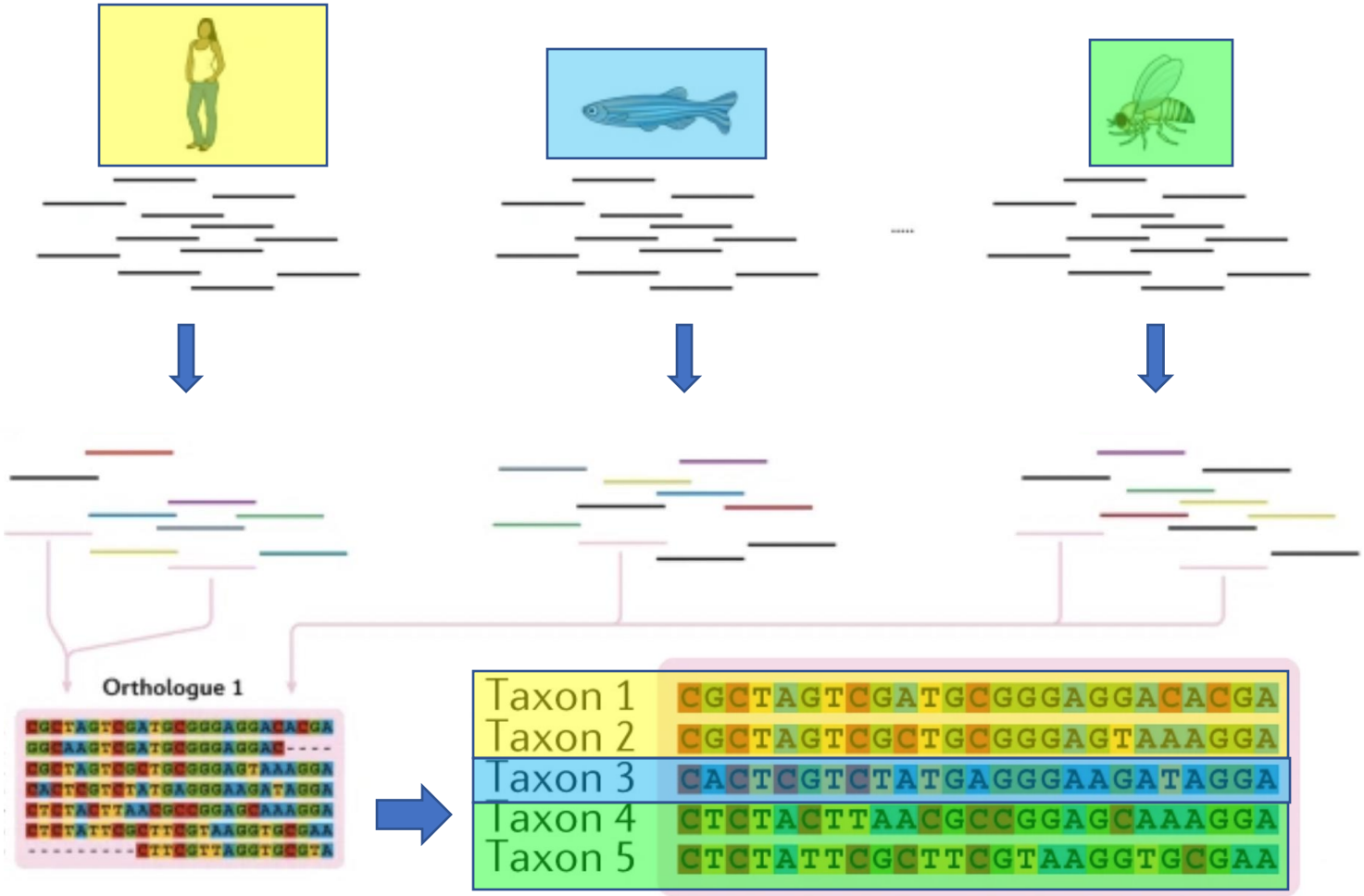
How can accuracy of a tree be tested?



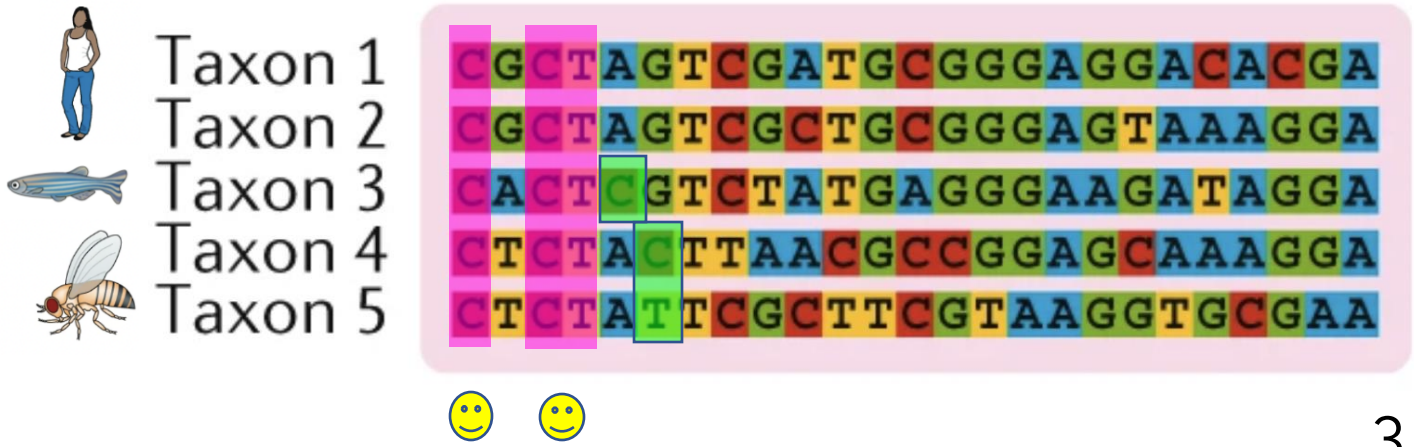


Parsimony: A preference for the least complicated hypothesis

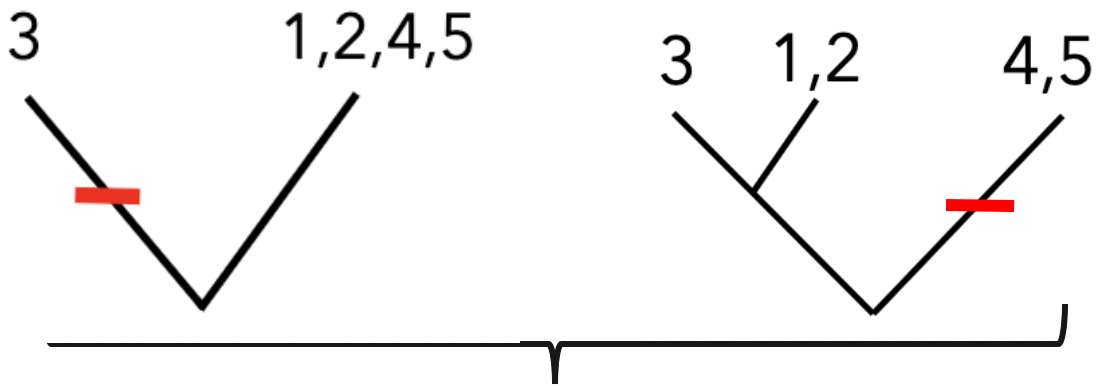
How is DNA aligned between species?



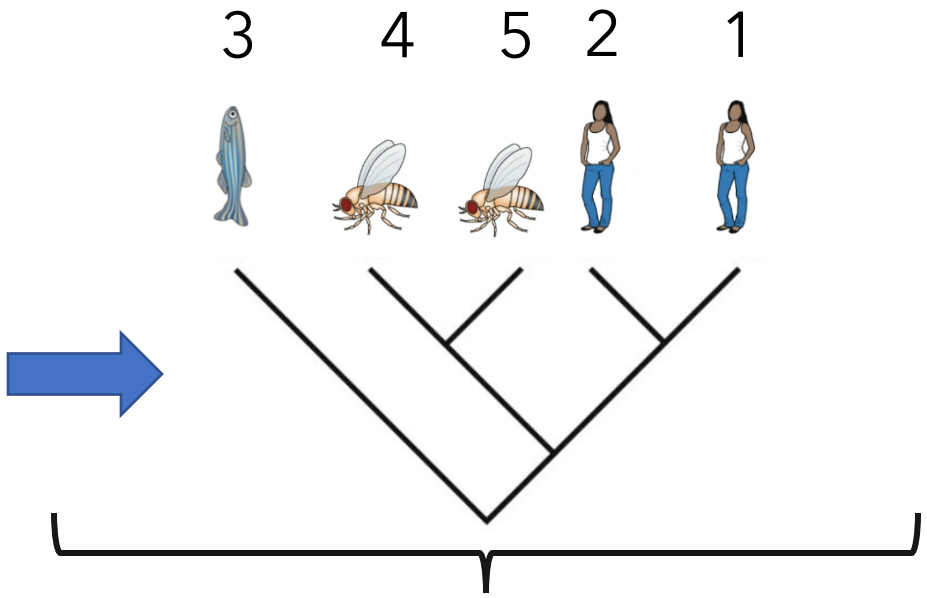
How do DNA deviations become a tree?



Progressive Sequence Alignment



Discrete Data Method



Distance-Matrix Method

Where do we find the genomic or proteomic data for our tree?

GenBank
NCBI

DDBJ
DNA Data Bank of Japan

EMBL
European Molecular Biology Laboratory

Protein BLAST
protein ▸ protein

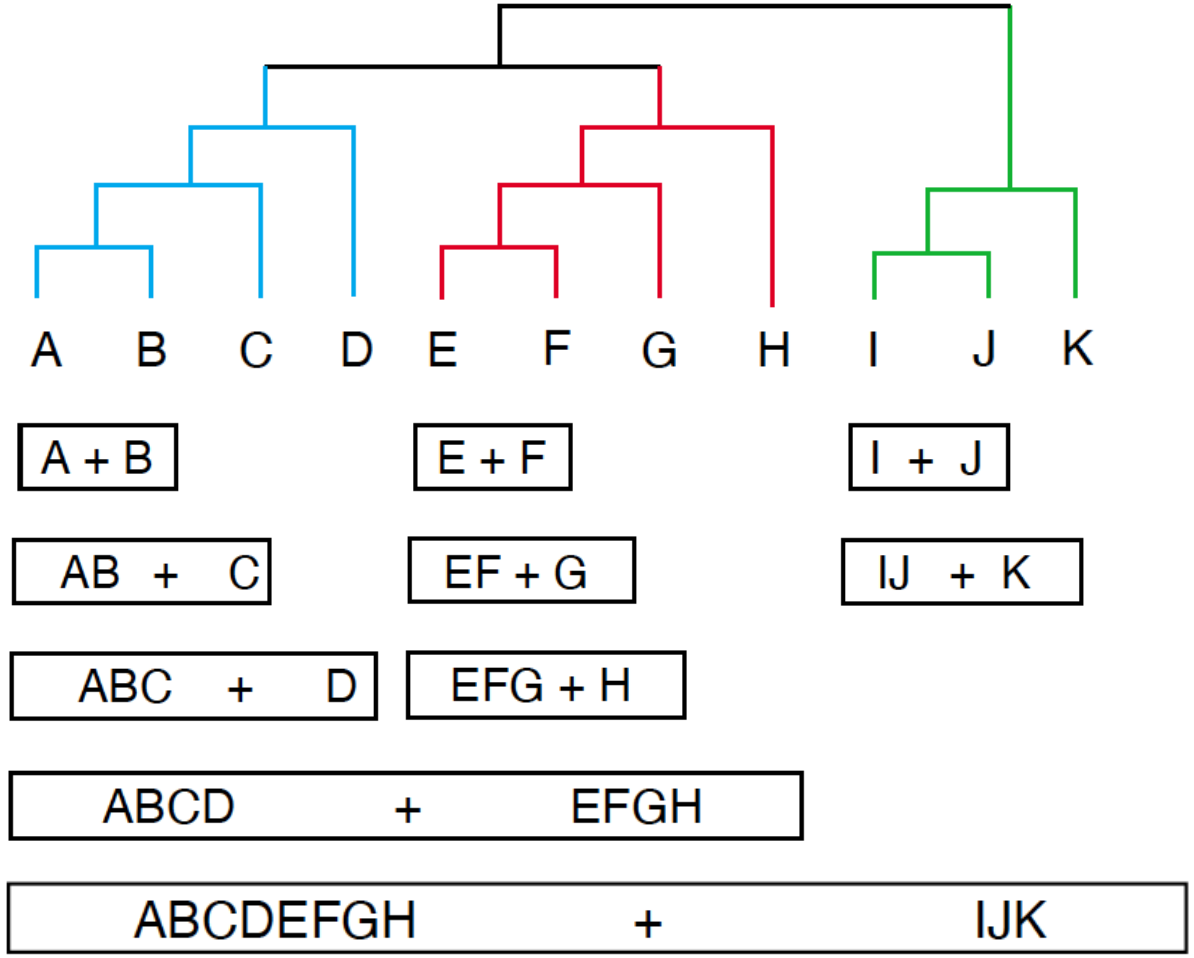
e!Ensembl

UniProt

Entrez

Public database similarity and keyword searches

How do we align multiple sequences?



Progressive Sequence Alignment

What if there are gaps in sequences during alignment?

```
taxon                    ....|....10...|....20...|....30...|....40...|....50
 Fu  Nosema.40928        QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAGHFGHIELVK
 Fu  Aspergillus.        QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
 Ap  Plasmodium.3        ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
 An  Cricetulus.2        QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An  Homo.7434727        QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An  Drosophila.9        QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
 An  Celegans.133        QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
 Fu  Spombe.54881        QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
 Pl  Athaliana.40        QFGILSPDEIRQMSVIH----VEHSETTEKGKPKVGGLECPGHFGYLELAK
 My  Ddiscoideum.        ----------------------------------------ECPGHFGHIELAK
 Rh  Porphyra.316        ----------------------------------------ECPGHFGFIELAK
 Kt  Tbrucei.1021        QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
 Kt  Leishmania.7        QFEVFKEAQIKAYAKCIIEHAKSYEHG----QPVRGGIECPGHFGYVELAE

taxon                    ....|....10...|....20...|....30...|....40...|....50
 Fu  Nosema.40928        QFGLFSPEEIRASSVAL--IRYPETLE--NGVPKESGLVCAGHFGHIELVK
 Fu  Aspergillus.        QFGLFSPEEIKRMSVVH--VEYPETMDEQRQRPRTKGLECPGHFGHIELAT
 Fu  Spombe.54881        QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
 Ap  Plasmodium.3        ELGVLDPEIIKKISVCE--IVNVDIYK--DGFPREGGLYCPGHFGHIELAK
 An  Cricetulus.2        QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An  Homo.7434727        QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An  Drosophila.9        QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
 An  Celegans.133        QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
 Pl  Athaliana.40        QFGILSPDEIRQMSVIH--VEHSETTE--KGKPKVGGLECPGHFGYLELAK
 My  Ddiscoideum.        ------------------------------------ECPGHFGHIELAK
 Rh  Porphyra.316        ------------------------------------ECPGHFGFIELAK
 Kt  Tbrucei.1021        QFEIFKERQIKSYAVCL--VEHAKSYA--NAADQSGEAECPGHFGYIELAE
 Kt  Leishmania.7        QFEVFKEAQIKAYAKCI--IEHAKSY--EHGQPVRGGIECPGHFGYVELAE
```

Gaps can be edited by eye to minimize gap events

In progressive sequence alignment, gaps cannot be moved or removed

How do we piece together a tree?

Sequence alignment

X: AAT TTG TCC

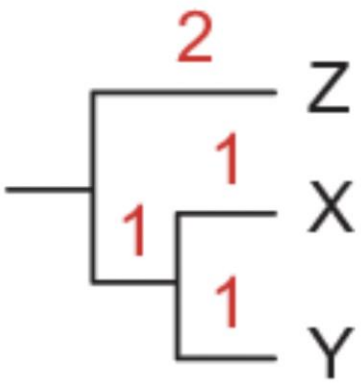
Y: AAT TTG TAA

Z: AAT TAA GGG

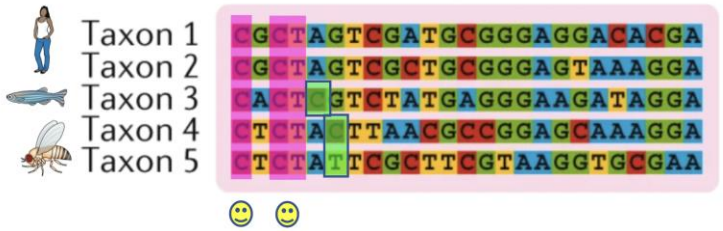
Pairwise genetic distance matrix

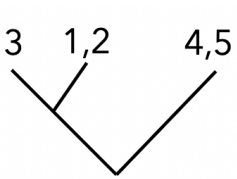
	X	Y	Z
X	–		
Y	2	–	
Z	4	4	–

Distance tree



Distance-Matrix Method





Discrete Data Method

Can statistical analysis validate our trees?

Dataset

```
        0123456789
seqA    ACCGTTCGGT
seqB    ATGGTTCAGA
seqC    ATCGATCGGA
```

Replicate 1

```
        1562314951
seqA    CTCCGCTTTC
seqB    TTCGGTTATT
seqC    TTCCGTAATT
```

Replicate 2

```
        5234924418
seqA    TCGTTCTTCG
seqB    TGGTAGTTTG
seqC    TCGAACAATG
```

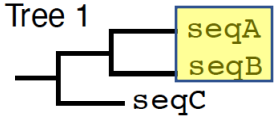
Replicate 3

```
        5607718907
seqA    TCAGGCGTAG
seqB    TCAAATGAAA
seqC    TCAGGTGAAG
```

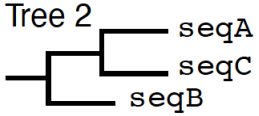
etc

x1000

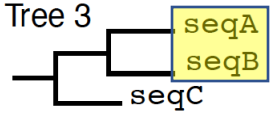
Tree 1



Tree 2

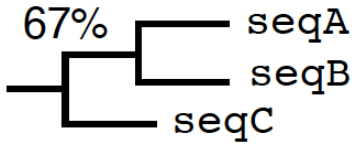


Tree 3



Bootstrap values greater than 70% are considered reliable

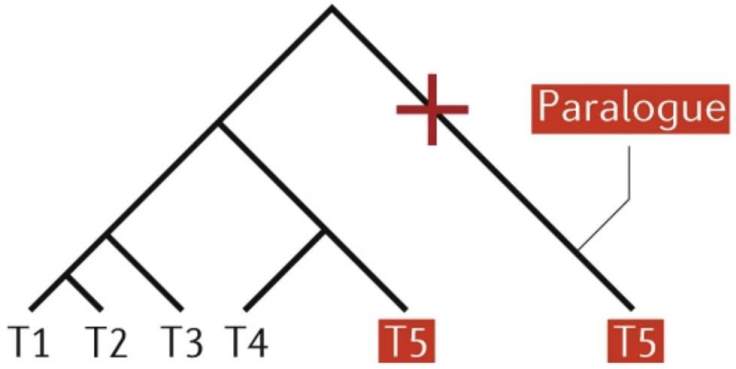
Any values lower than 50% should be removed from the tree

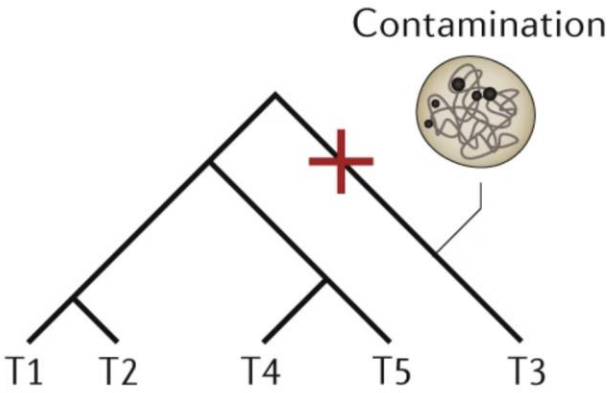


Bootstrap consensus tree

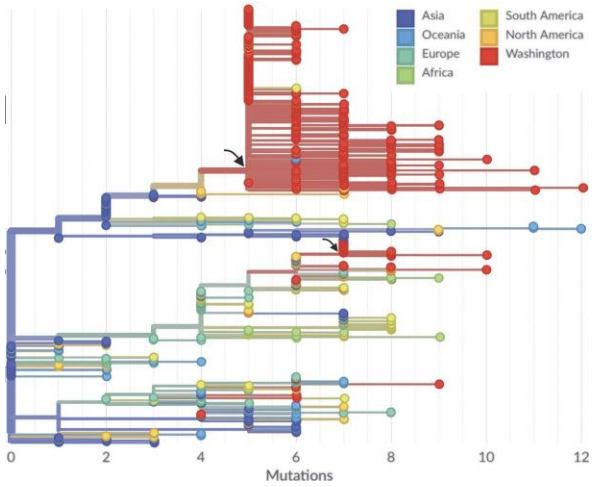
Bootstrapping!

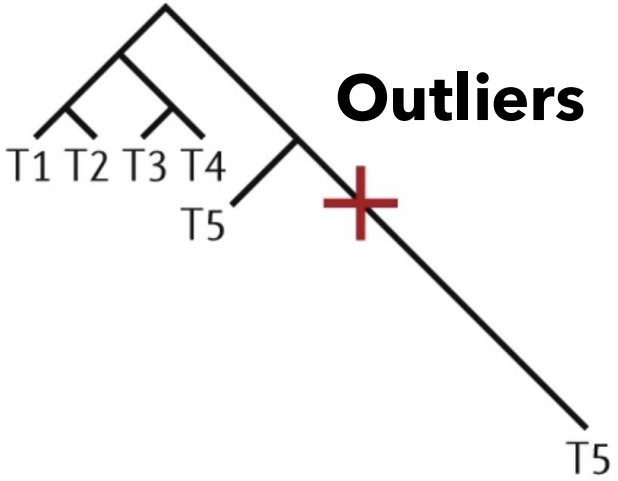
What uncertainties are there in phylogeny?





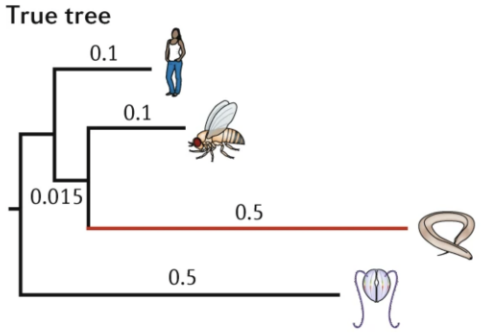
Mutations

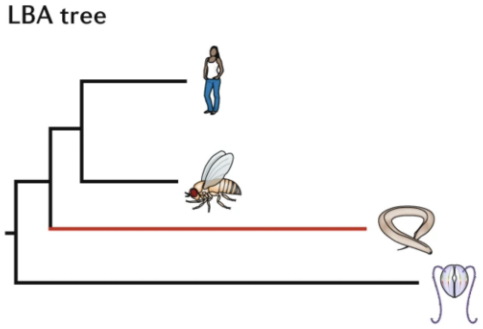




Outliers

Long Branch Attraction





How can phylogeny be applied to infectious disease?

GISAID

Obtain data

Nextstrain

Real-time tracking of pathogen evolution

Use Augur, a bioinformatics resource to process and align data

IQTREE to create maximum likelihood phylogenetic trees

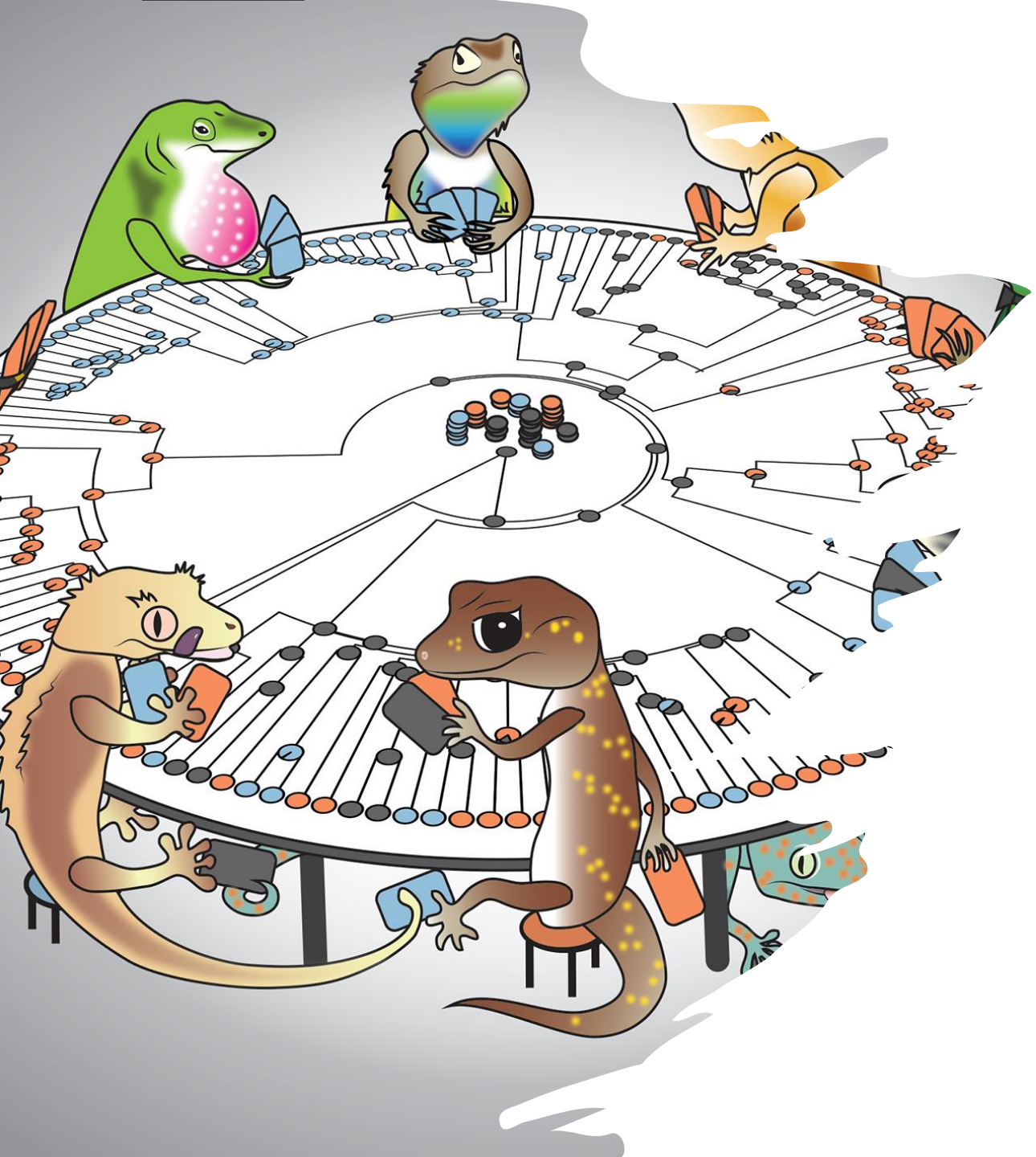
Nextstrain

Real-time tracking of pathogen evolution

Auspice to view result

The approach used by Trevor Bedford in this week's paper.

Images from Nextstrain, iqtree.org, and GISAID



Summary

Phylogenetics allows us to map the evolutionary past and better understand infectious disease

Phylogenetic trees are made by gathering sequences, aligning them, and finding the most probable pattern

Trees can be verified statistically, however, sources of error are always possible

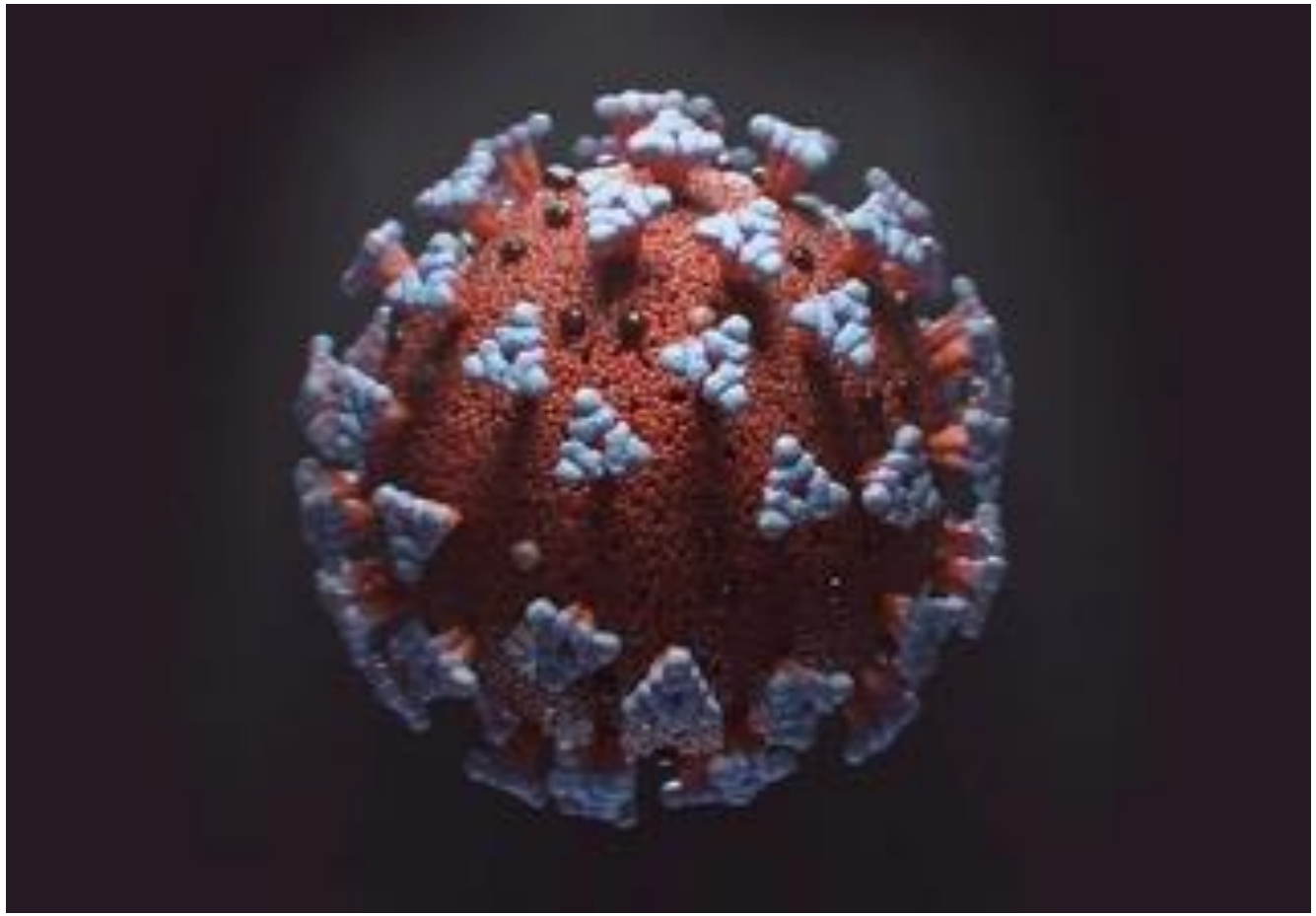
REPORT

CORONAVIRUS

Cryptic transmission of SARS-CoV-2 in Washington state

Trevor Bedford[1,2,3]*†, Alexander L. Greninger[1,4]†, Pavitra Roychoudhury[1,4]†, Lea M. Starita[2,3]†, Michael Famulare[5]†, Meei-Li Huang[1,4], Arun Nalla[4], Gregory Pepper[4], Adam Reinhardt[4], Hong Xie[4], Lasata Shrestha[4], Truong N. Nguyen[4], Amanda Adler[6], Elisabeth Brandstetter[7], Shari Cho[2,3], Danielle Giroux[3], Peter D. Han[2,3], Kairsten Fay[1], Chris D. Frazar[3], Misja Ilcisin[1], Kirsten Lacombe[6], Jover Lee[1], Anahita Kiavand[2,3], Matthew Richardson[3], Thomas R. Sibley[1], Melissa Truong[2,3], Caitlin R. Wolf[7], Deborah A. Nickerson[2,3], Mark J. Rieder[2,3], Janet A. Englund[2,6,8], The Seattle Flu Study Investigators‡, James Hadfield[1], Emma B. Hodcroft[9,10], John Huddleston[1,11], Louise H. Moncla[1], Nicola F. Müller[1], Richard A. Neher[9,10], Xianding Deng[12], Wei Gu[12], Scot Federman[12], Charles Chiu[12], Jeffrey S. Duchin[7,13], Romesh Gautom[14], Geoff Melly[14], Brian Hiatt[14], Philip Dykema[14], Scott Lindquist[14], Krista Queen[15], Ying Tao[15], Anna Uehara[15], Suxiang Tong[15], Duncan MacCannell[16], Gregory L. Armstrong[16], Geoffrey S. Baird[4], Helen Y. Chu[2,7]§, Jay Shendure[2,3,17]§, Keith R. Jerome[1,4]§

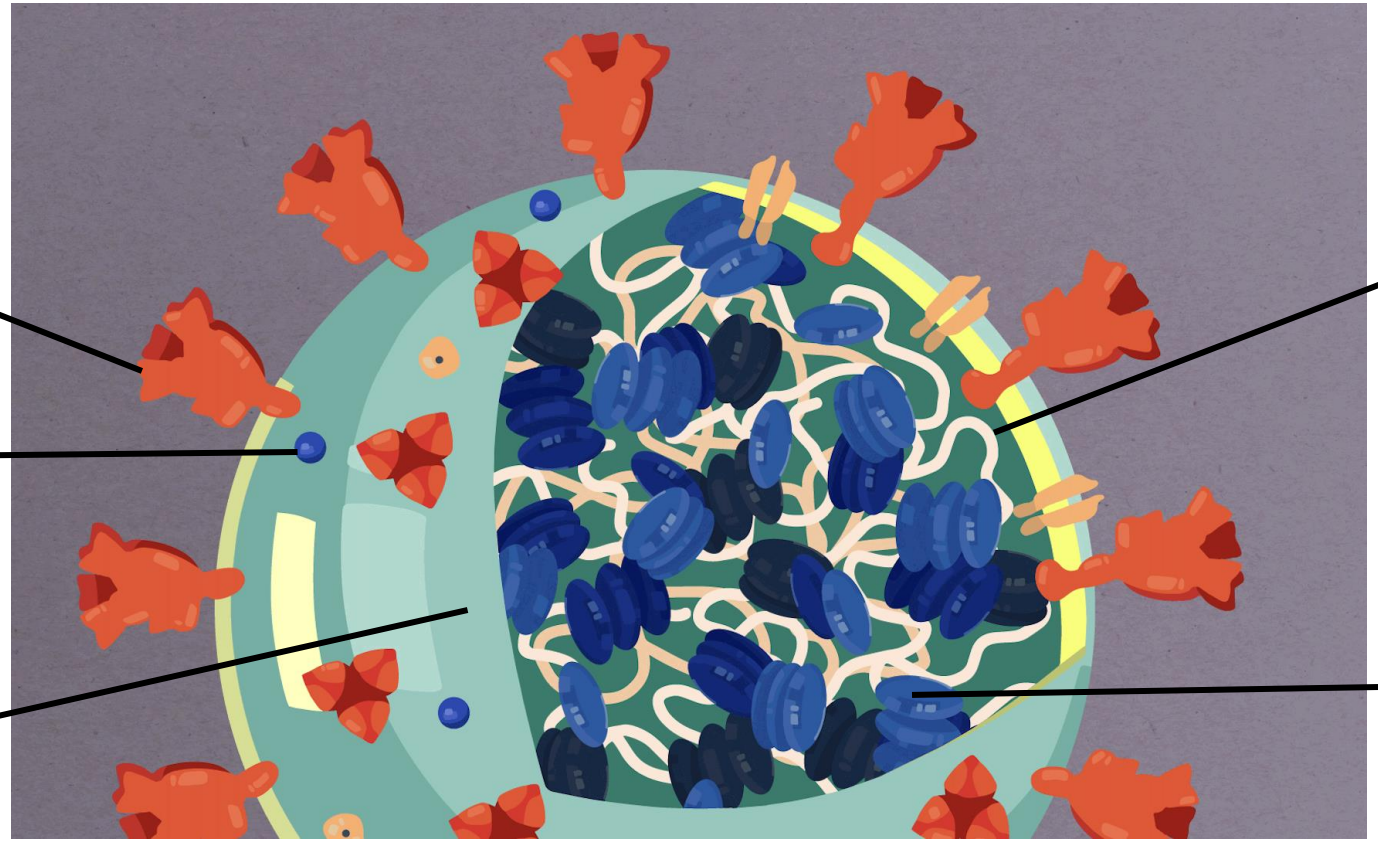
What is a coronavirus?



A family of zoonotic viruses that can lead to the development of respiratory illness.

(World Health Organization- Regional Office for Africa, 2021); Image from Cleveland Clinic

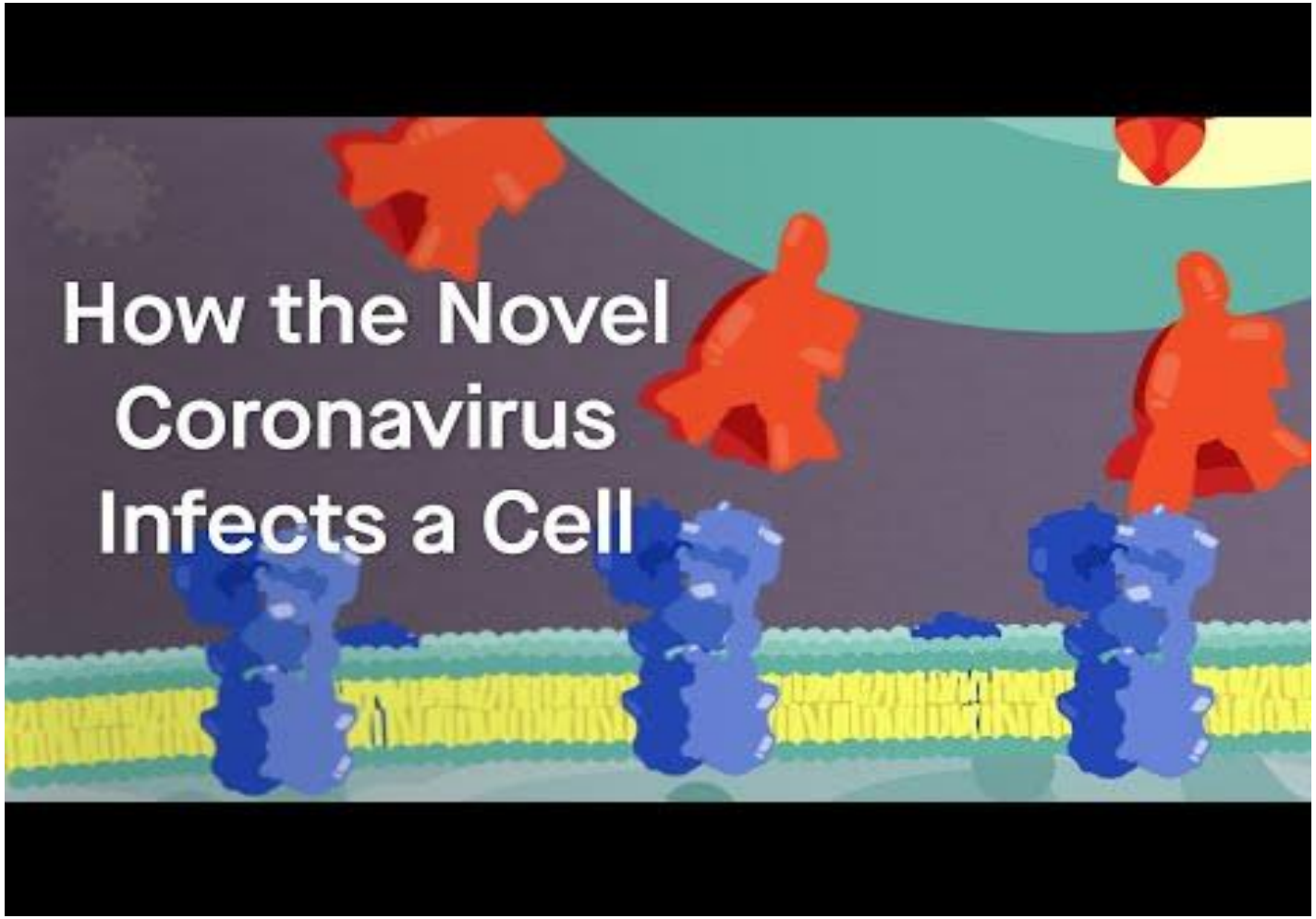
What does the structure of SARS-CoV-2 look like?



Severe Acute Respiratory Syndrome Coronavirus 2 is the virus that causes COVID-19.

(Scripps Research, 2020); (CDC, 2021)

How does SARS-CoV-2 enter a cell?



(Chowdhury et al., 2020); (Scripps Research, 2020)

What are the symptoms of COVID-19?

Symptoms can range from mild to severe illness, and appear 2–14 days after you are exposed to the virus that causes COVID-19.

Also:

- **Headache**
- **Sore throat**

Image from Georgia Cyber Academy

When did the COVID-19 outbreak start?

End of 2019 (late November or early December)

Cases of COVID-19 arise.

1/19/2020

First confirmed case of COVID-19 detected in the U.S.

2/27/2020

Centers for Disease Control changes testing recommendations.

2/28/2020

2nd case detected in Snohomish county, but not linked to travel.

2/28/2020

Sequencing starts following the Washington outbreak.

3/15/2020

Shutdowns begin in the United States.

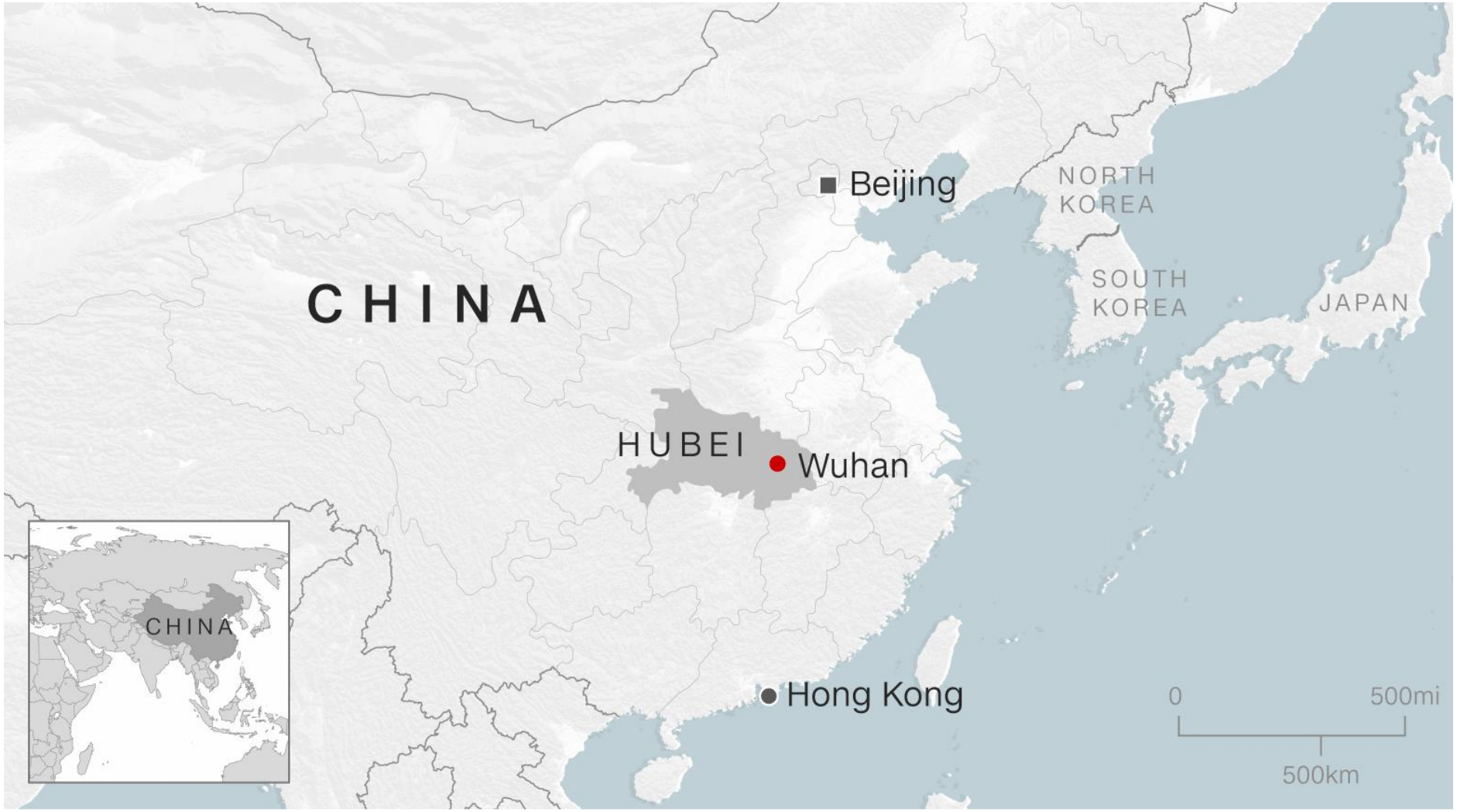
3/25/2020

The state of Washington has 2580 cases and 132 deaths due to COVID-19.

5/18/2020

More than four million cases of COVID-19 and more than 300,000 deaths resulted.

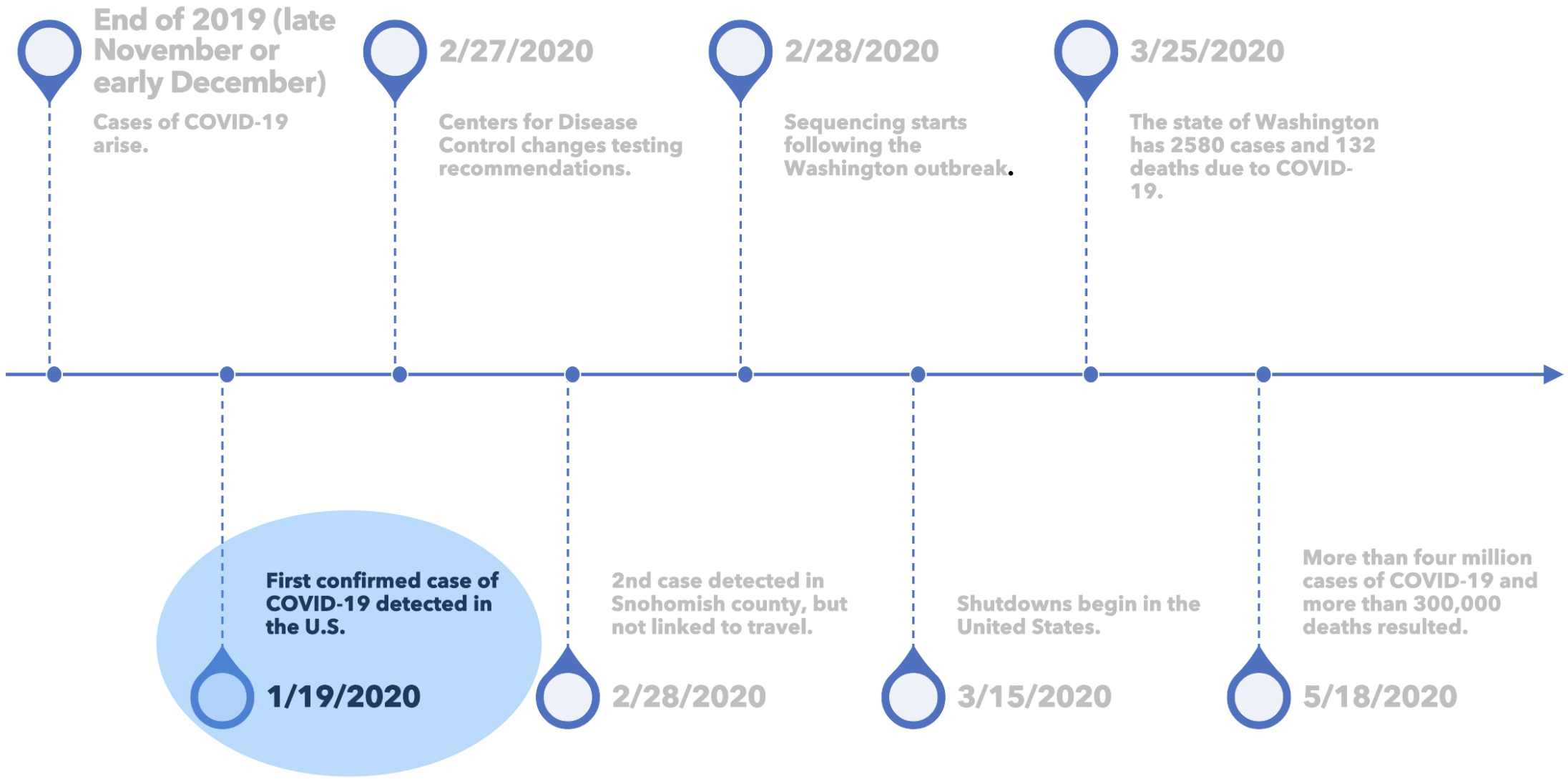
Where did the virus originate?



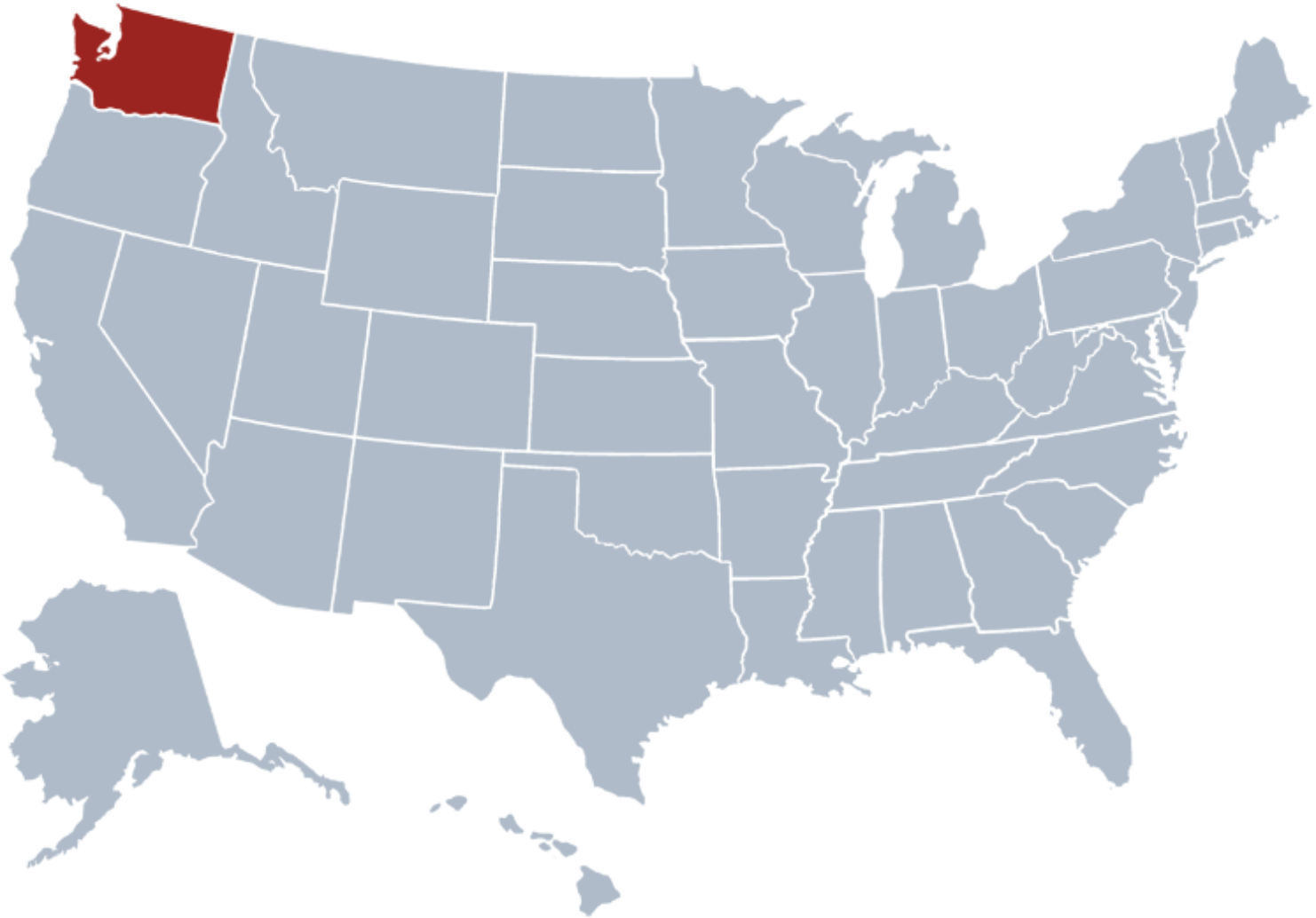
Source: Maps4news.com/©HERE

Image from CNN

When did COVID-19 begin to spread to the U.S.?

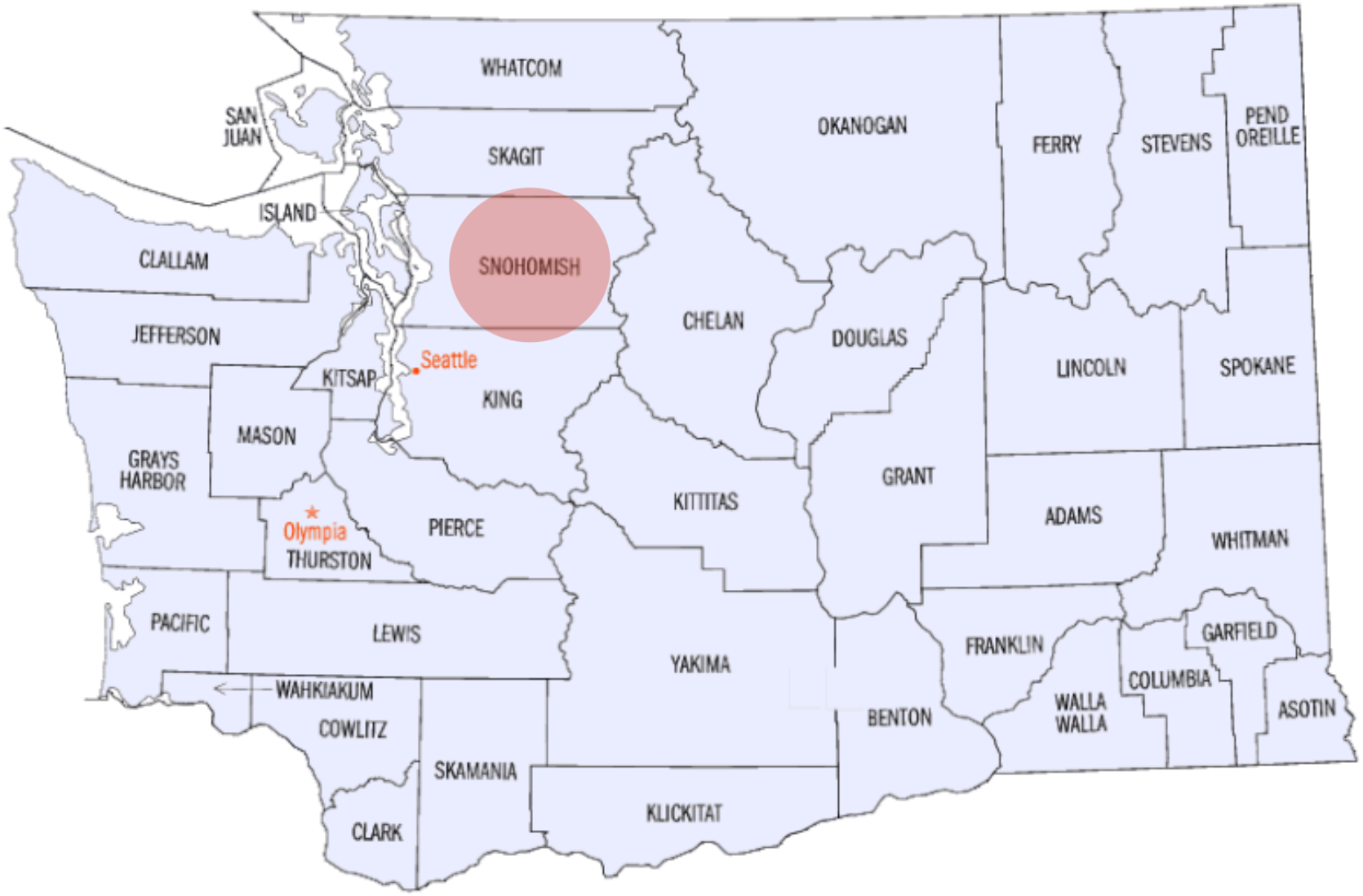


Where was the first confirmed case of COVID-19 located in the U.S.?



Snohomish County, Washington State

Image from 50states.com



Snohomish County borders King County, which is where Seattle is located.

Image from Wikipedia

How did the U.S. respond to the spread of COVID-19?

End of 2019 (late November or early December)

Cases of COVID-19 arise.

1/19/2020

First confirmed case of COVID-19 detected in the U.S.

2/27/2020

Centers for Disease Control changes testing recommendations.

2/28/2020

2nd case detected in Snohomish county, but not linked to travel.

2/28/2020

Sequencing starts following the Washington outbreak.

3/15/2020

Shutdowns begin in the United States.

3/25/2020

The state of Washington has 2580 cases and 132 deaths due to COVID-19.

5/18/2020

More than four million cases of COVID-19 and more than 300,000 deaths resulted.

How did the CDC change COVID testing recommendations?

Up to 2/27/2020, the CDC advised testing only for people who were exposed to a confirmed case of COVID-19 or with specific travel history.

Images from AAA Newsroom and MD Anderson Cancer Center

How do you test a sample for SARS-CoV-2?

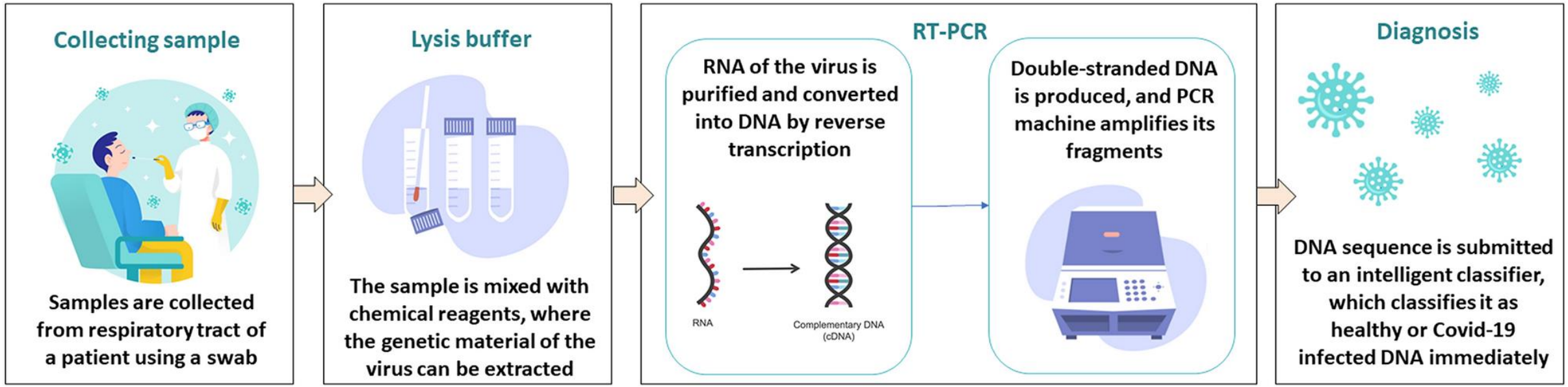
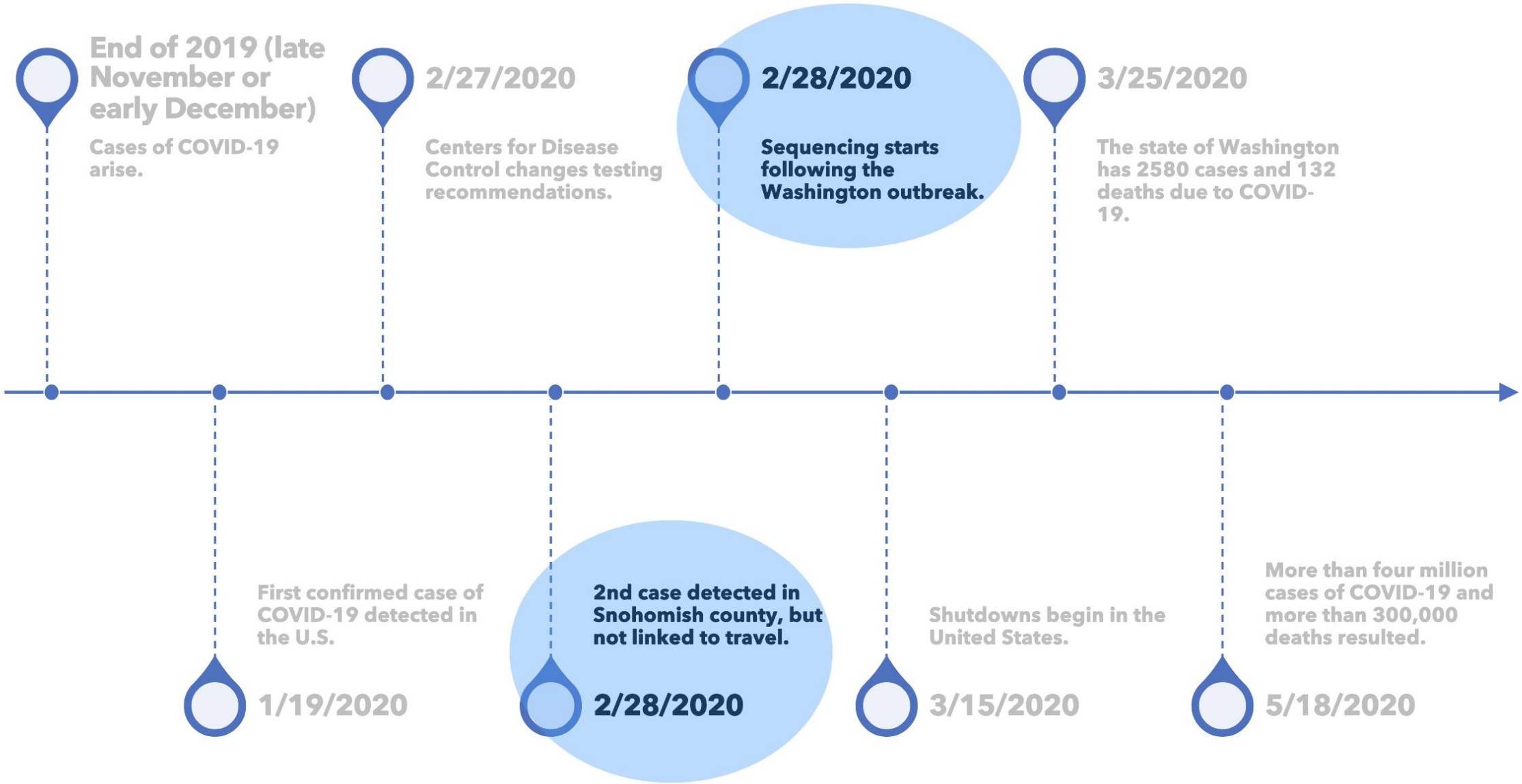
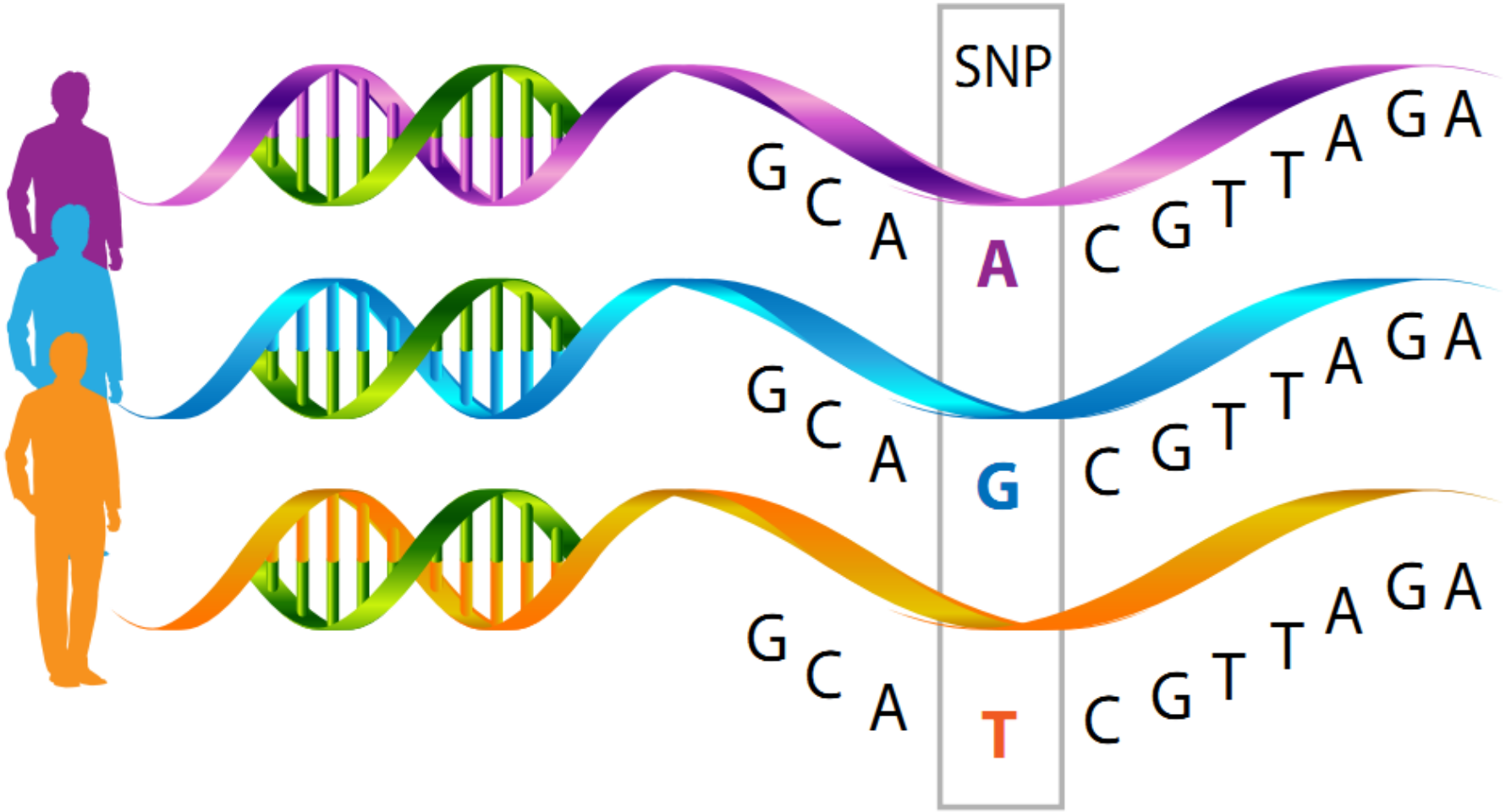


Image from nature.com

What happened after COVID-19 was detected in Washington?



What is a Single Nucleotide Polymorphism, or a SNP?



Change in a base (A, T, G, or C) at one position leading to genetic variation.

(Bean & Fischer, 2017); (Chowdhury et al., 2020)

What is the Washington State Outbreak Clade?

Clade comes from viruses spreading in China.

Samples from 2/20/2020-3/15/2020

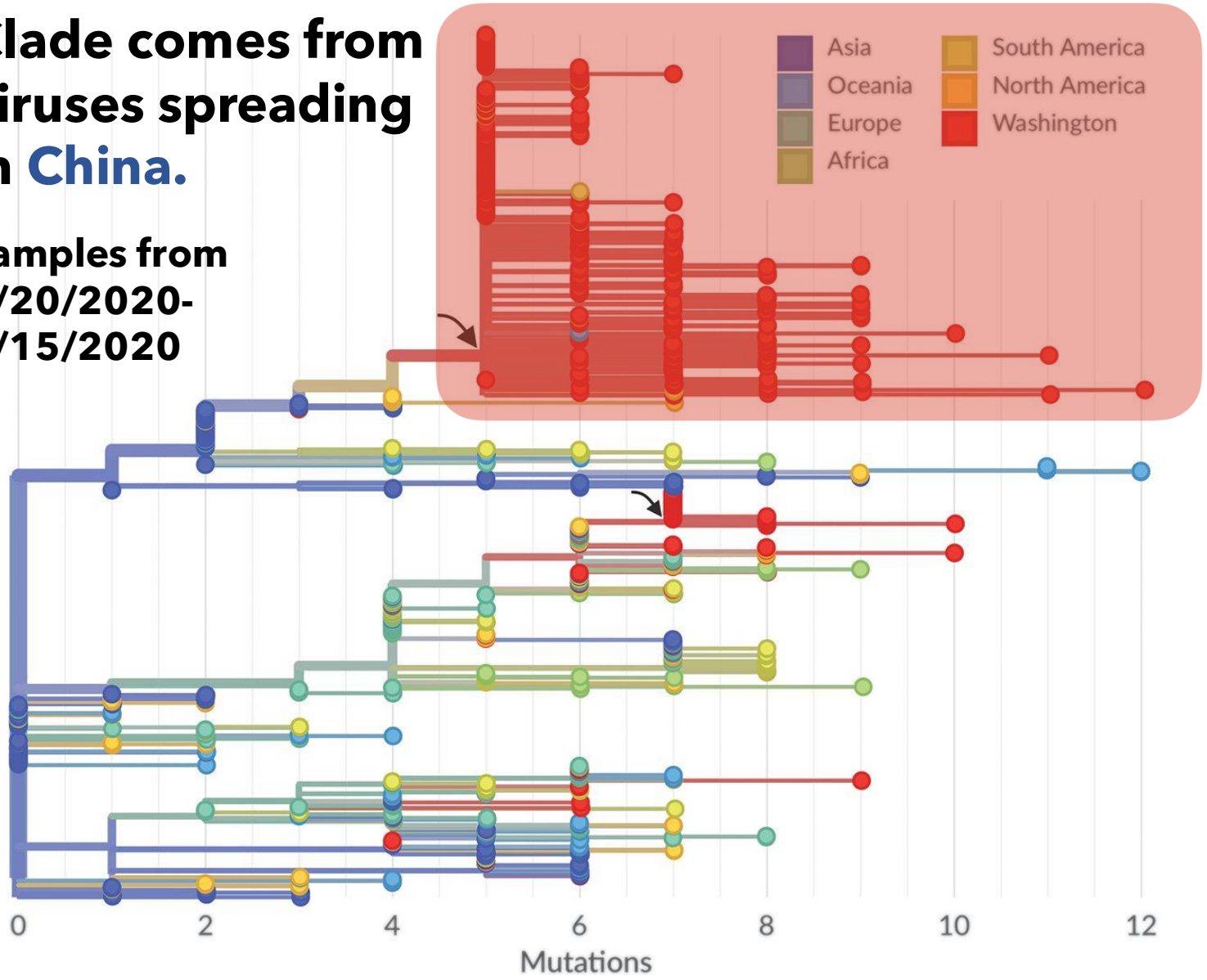
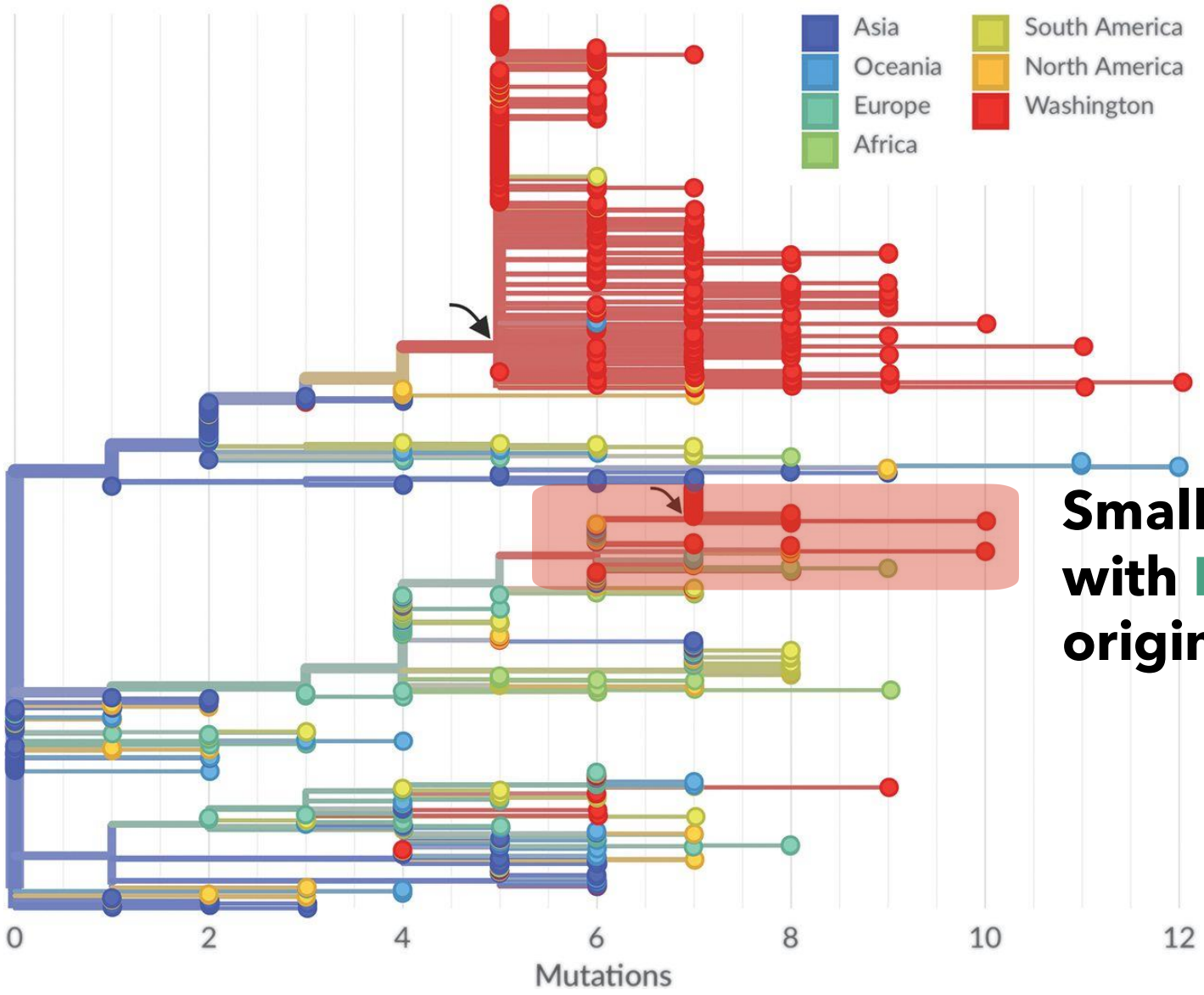


Figure 1

Are all early COVID-19 cases in Washington part of the same clade?



Small clade with European origins.

Figure 1

Are there early cases of COVID-19 in Washington that fall outside these two clades?

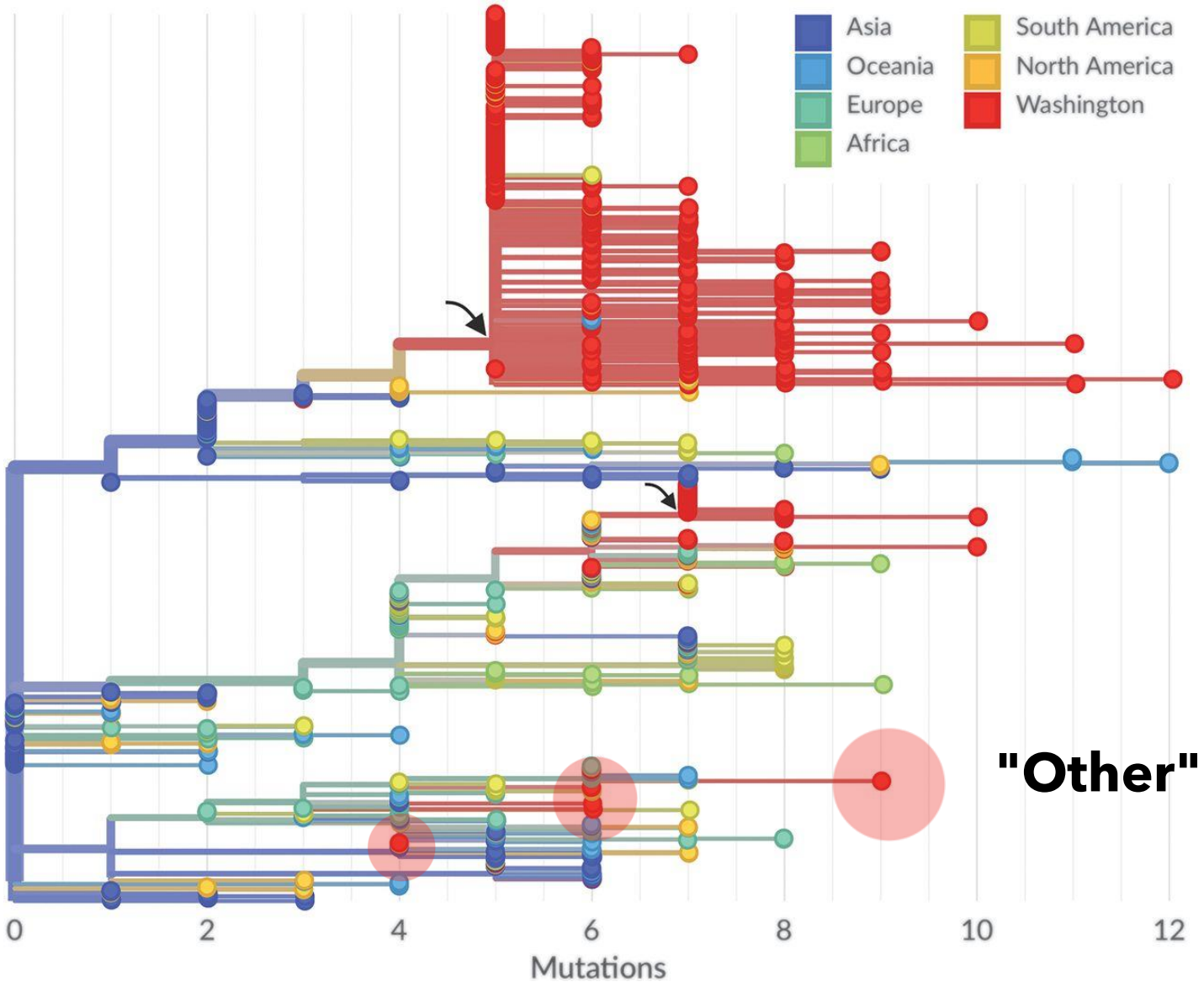
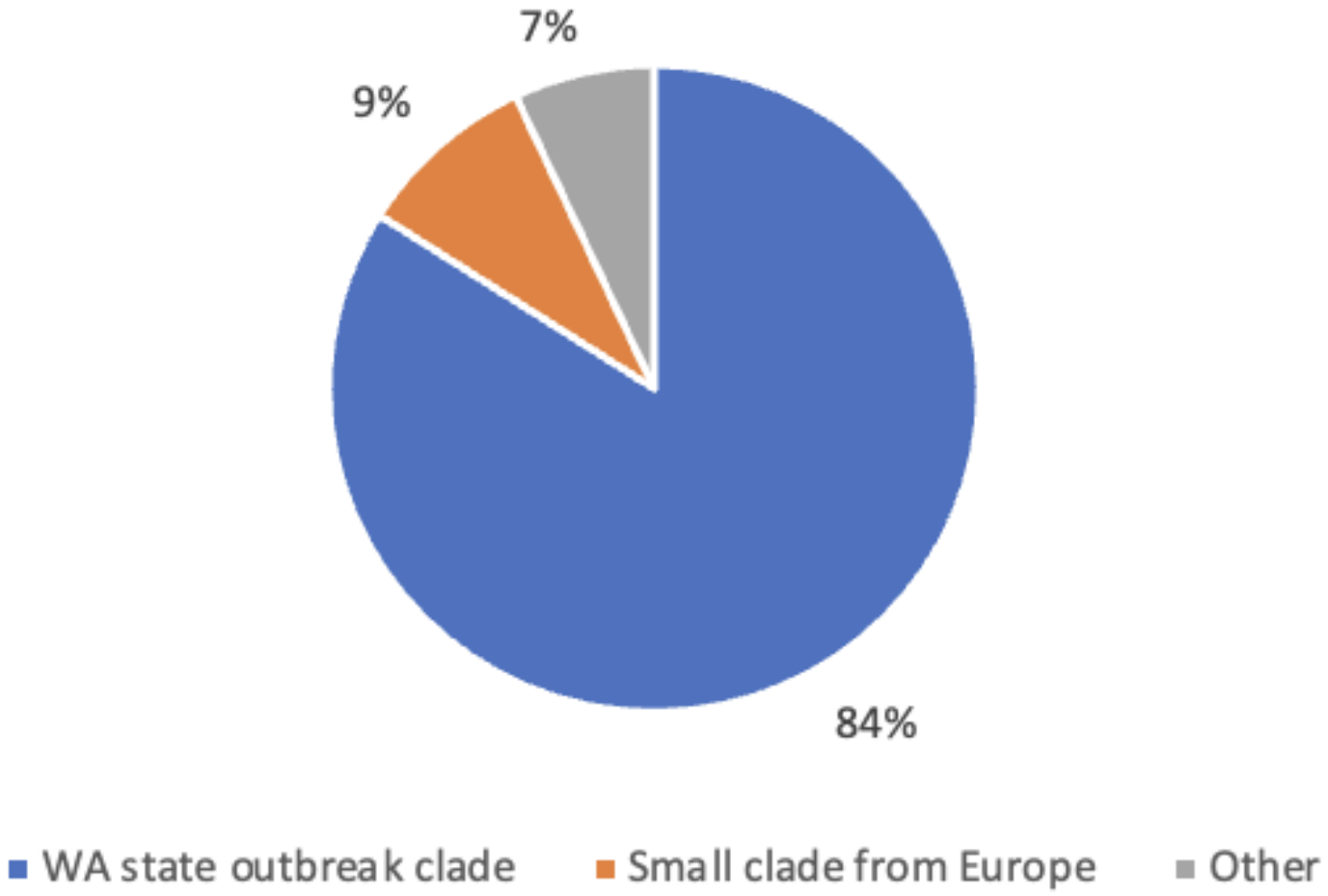


Figure 1

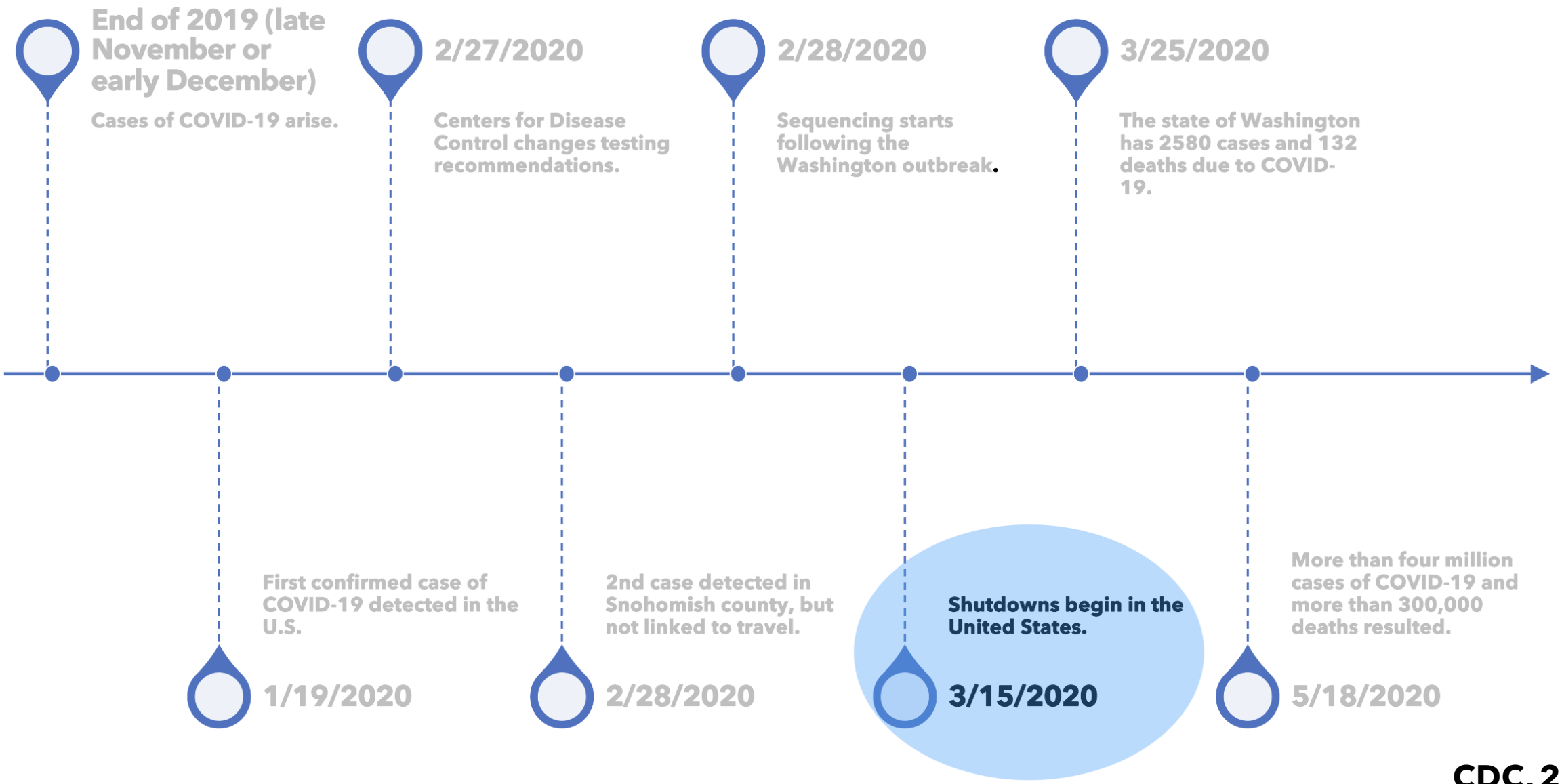
What does this data tell us about early COVID-19 transmission in Washington?



An introduction event took place. In turn, the virus spread locally.

Chart created on excel

What measures did the states take to prevent the spread of COVID-19?

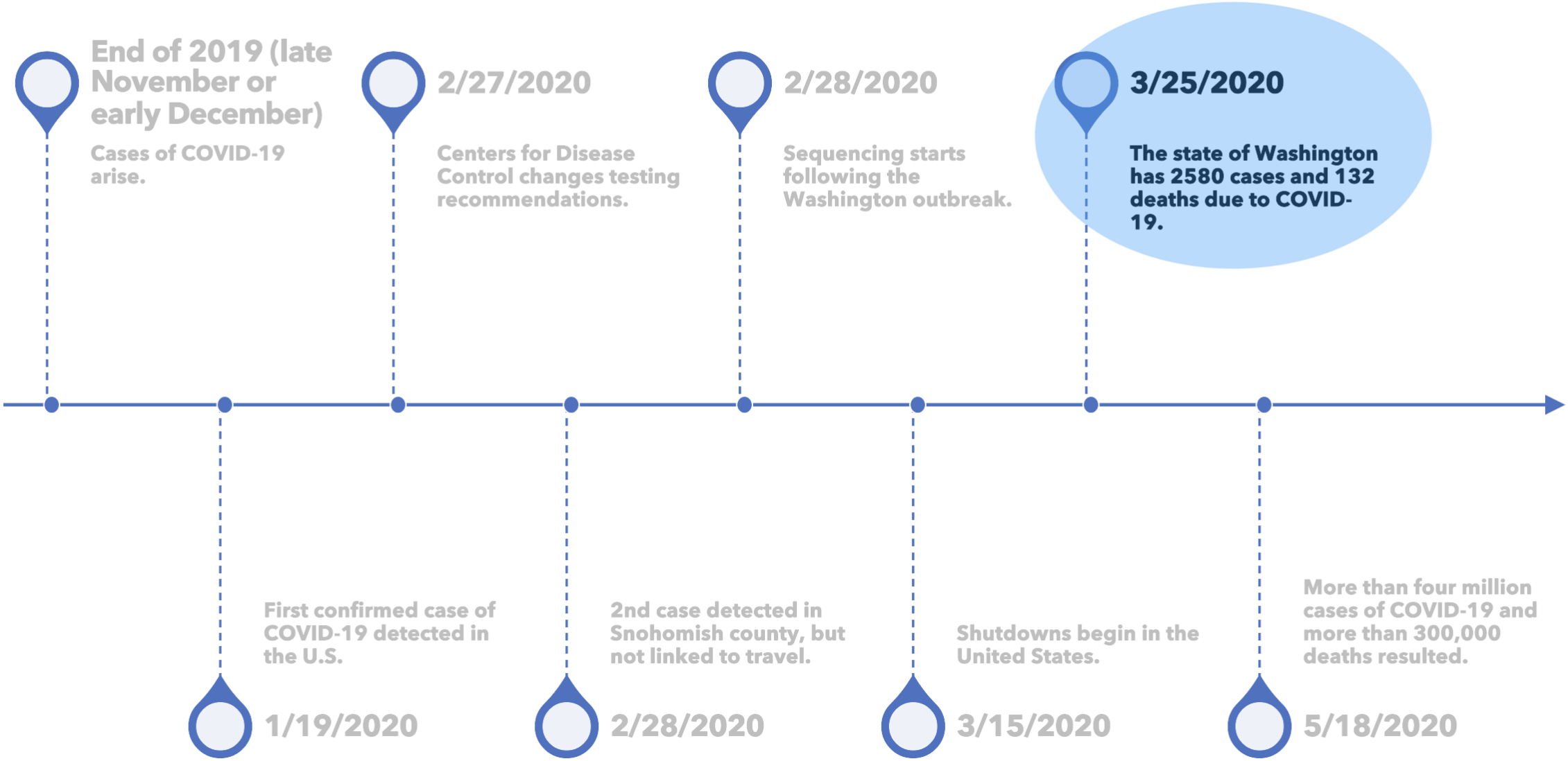


CDC, 2022

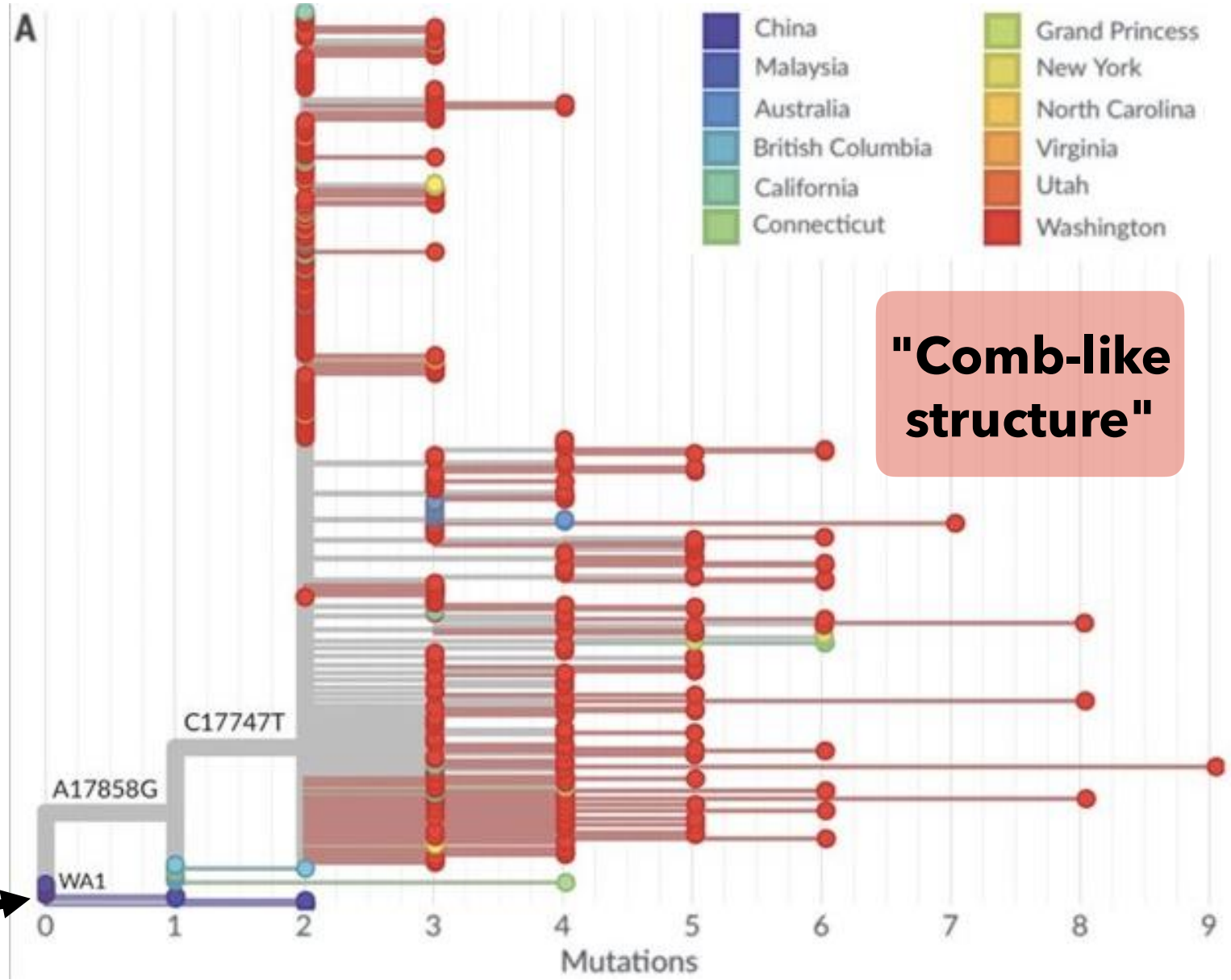
Non-essential businesses shut down across the nation.

Images from Berkley University of California Research, wbur.org, and wqxr.org

How did COVID-19 progress in Washington state?



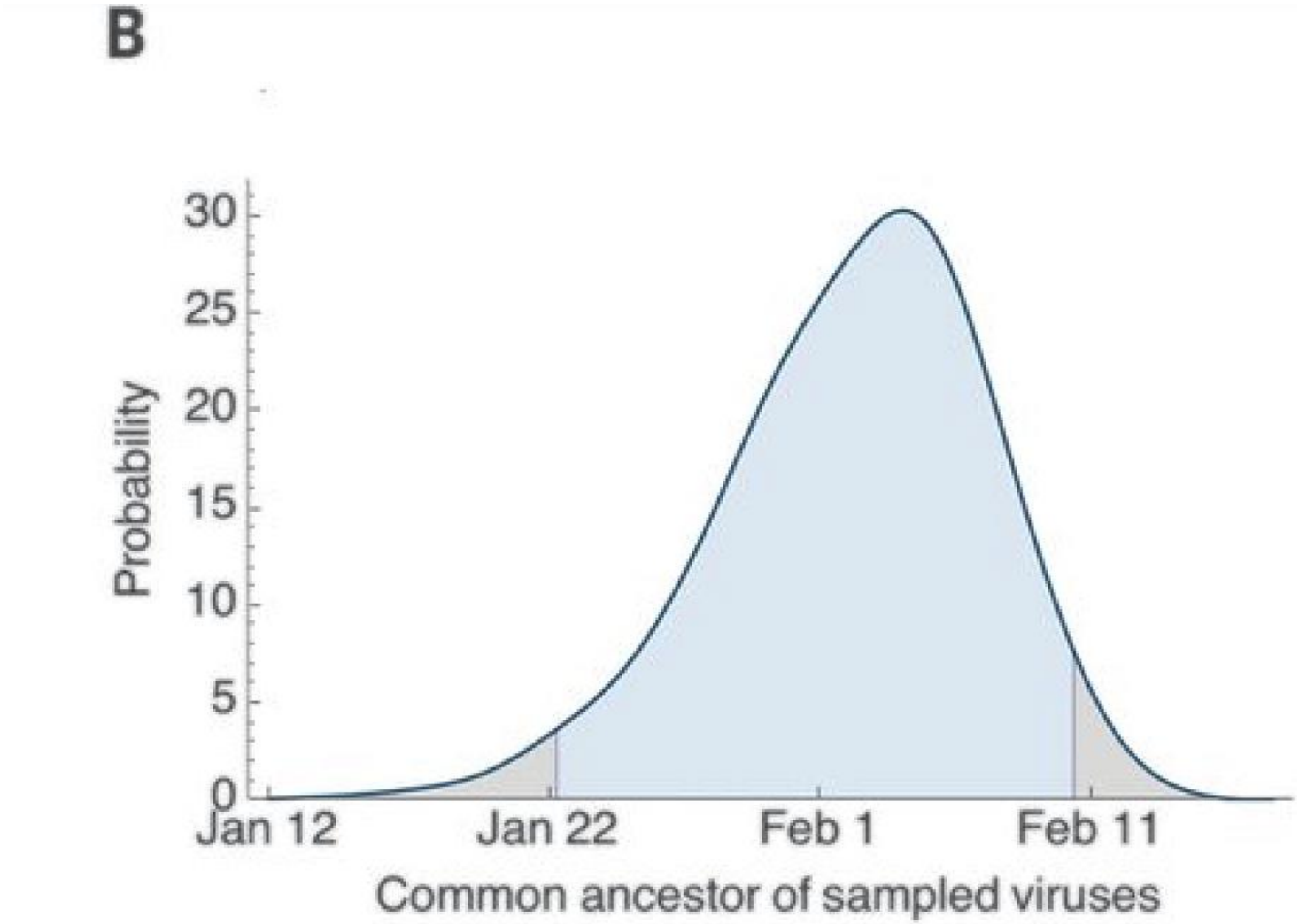
How did the COVID-19 cases increase so rapidly in Washington?



Strain from **Snohomish County** resident

Figure 2A

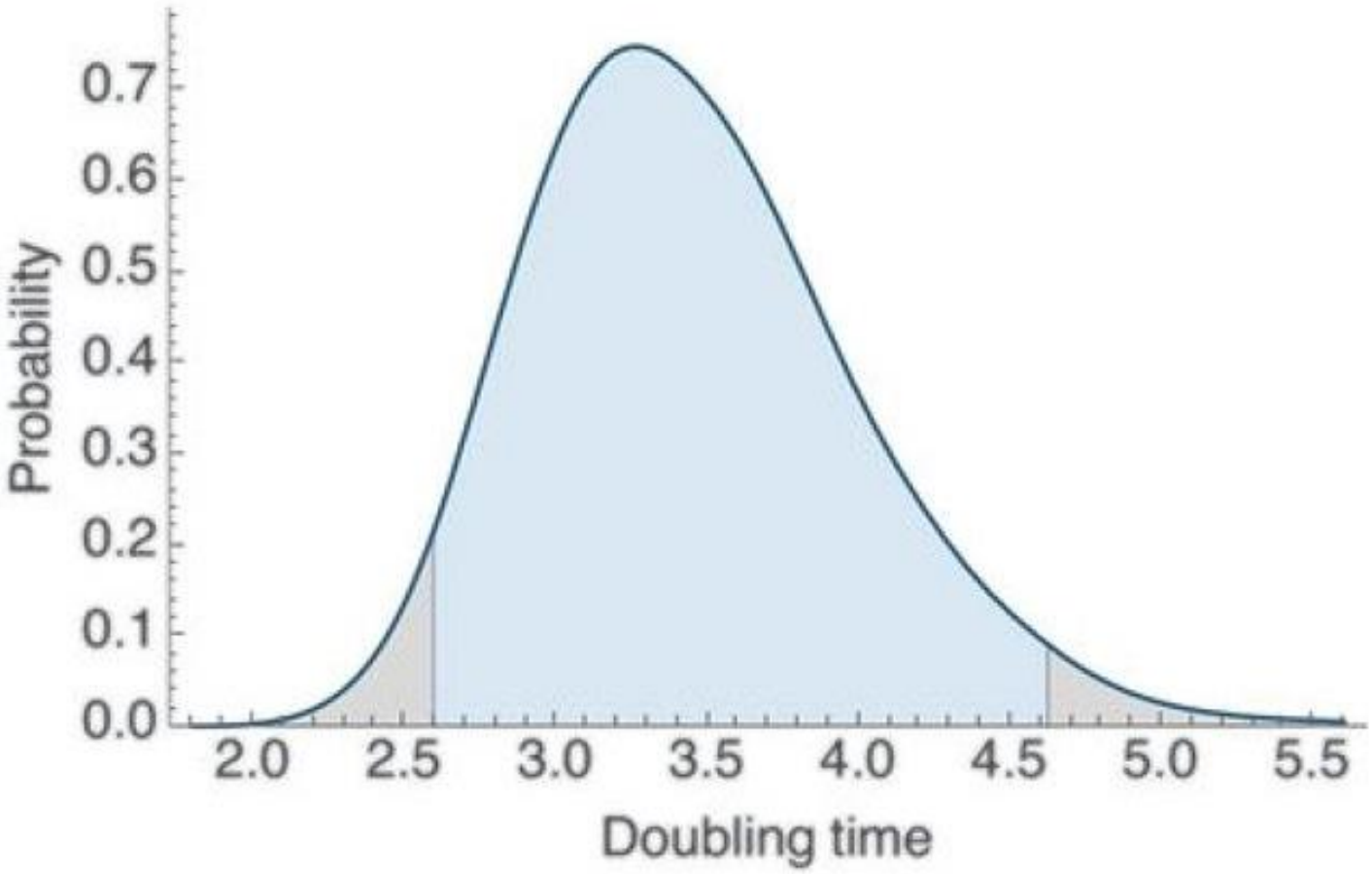
When did the common ancestor of the Washington state outbreak clade emerge?



Median = February 2nd, 2020

Figure 2B

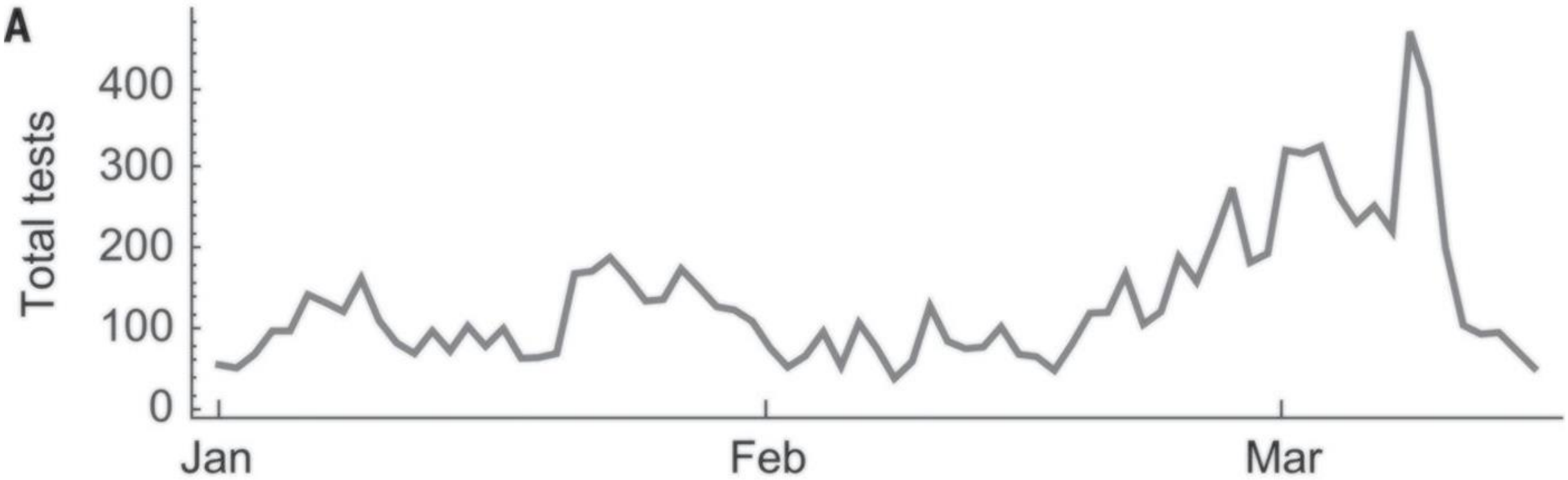
What was the growth rate of the Washington state clade?



Median doubling time = 3.4 days

Figure 2B

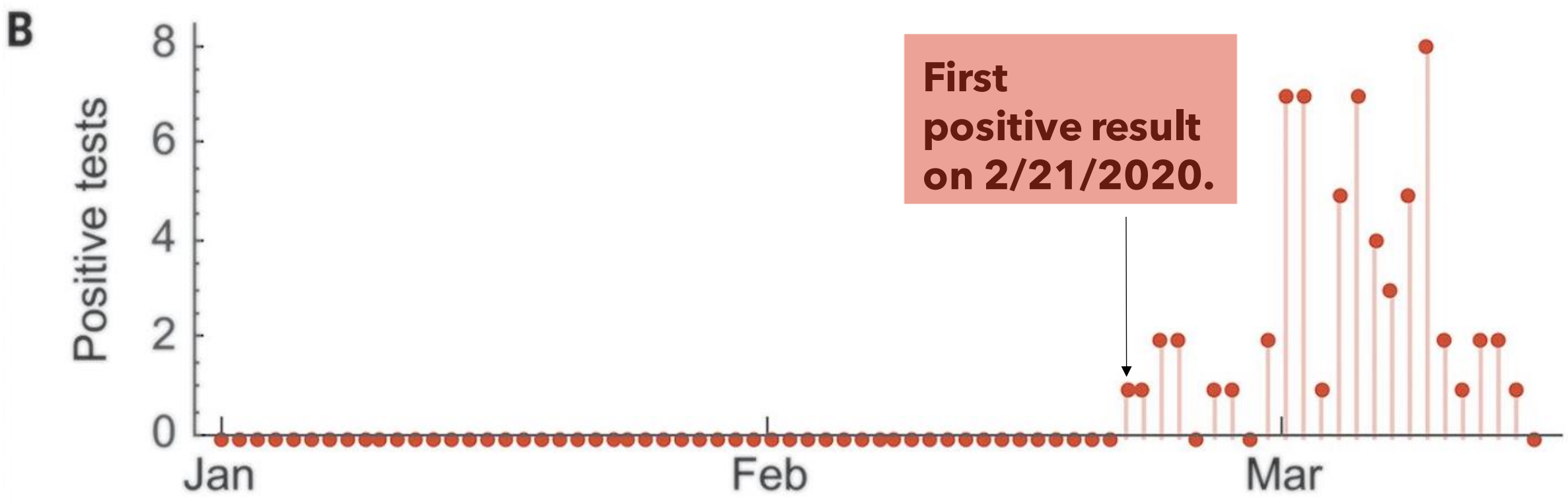
Can we determine much COVID-19 was circulating in WA in early 2020?



Between 1/1/2020 and 3/15/2020, there were 10,382 samples that were obtained and later used for analysis.

Figure 3 graph A

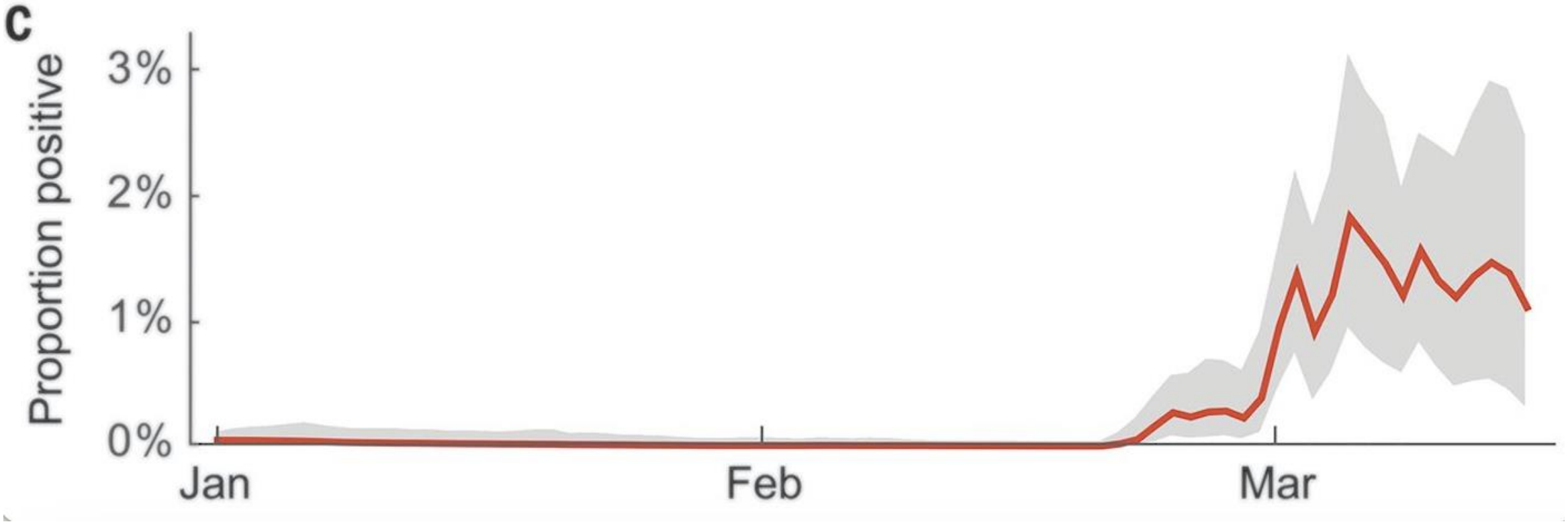
How many samples from the Seattle Flu Study were positive for COVID-19?



Subsequently, there were 65 positive samples.

Figure 3 graph B

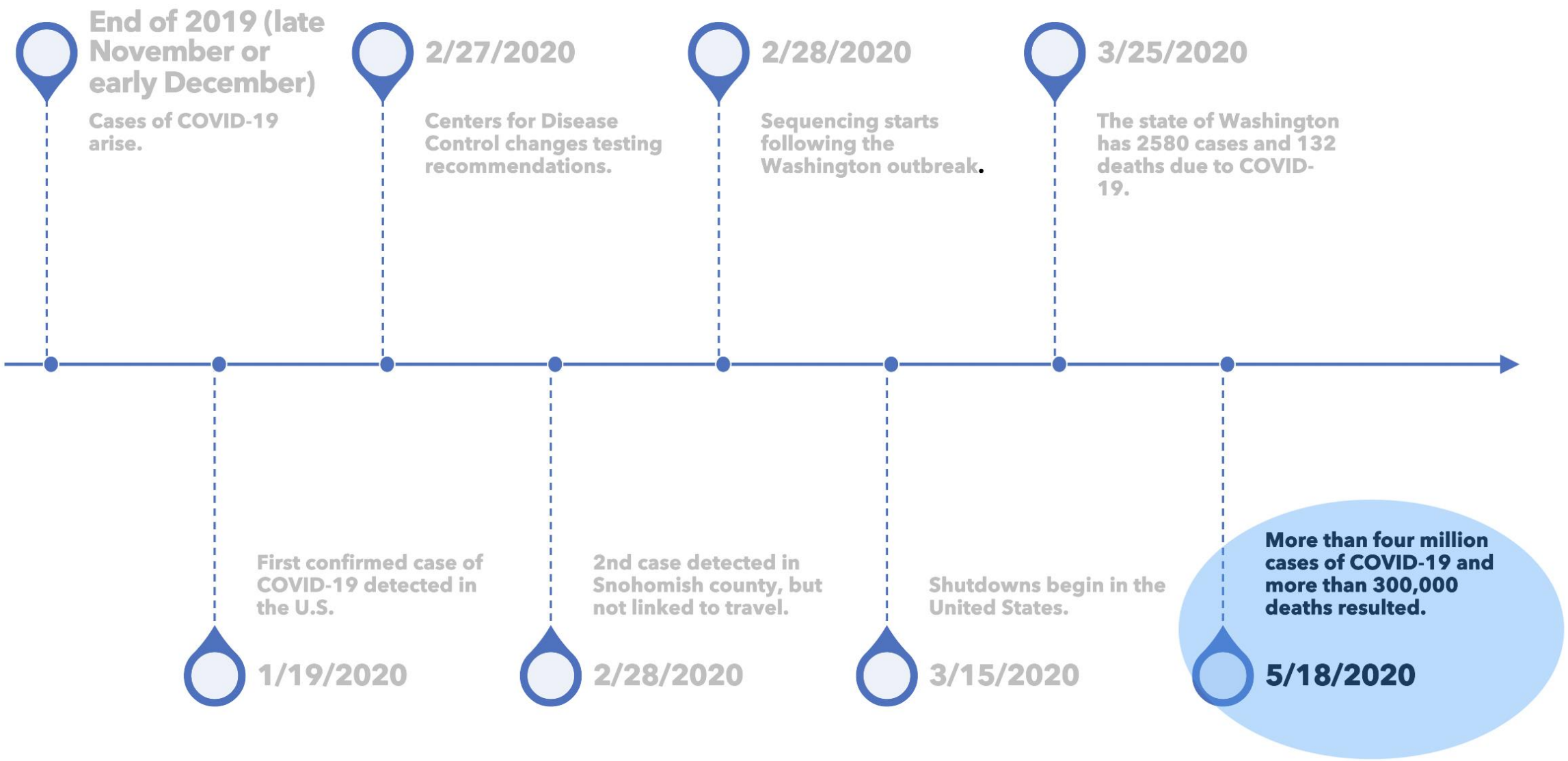
What proportion of samples were positive for COVID-19 in the Seattle Flu study?

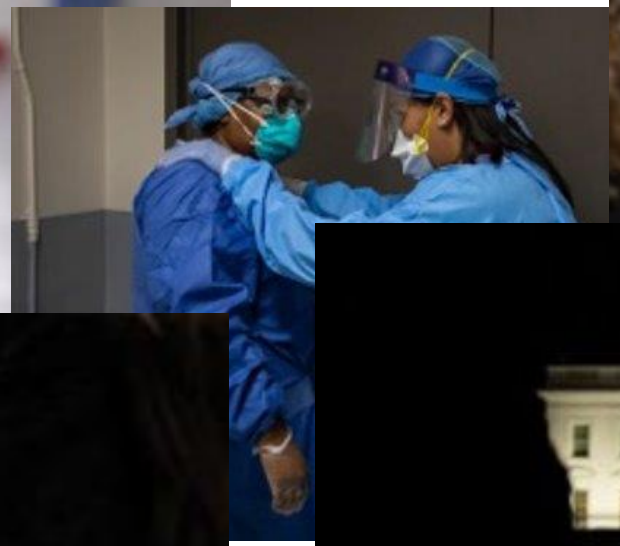


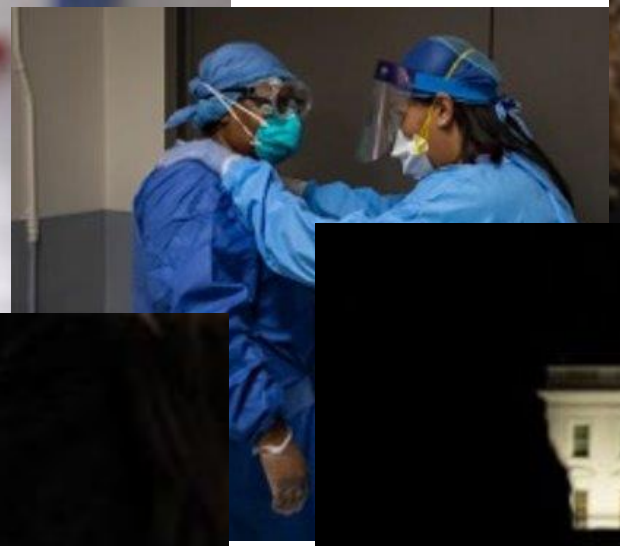
1.1% of specimens were positive for SARS-CoV-2 and there was a population prevalence of 0.05% on March 1st, 2020.

Figure 3 graph C

Two months into the pandemic, what was the burden?



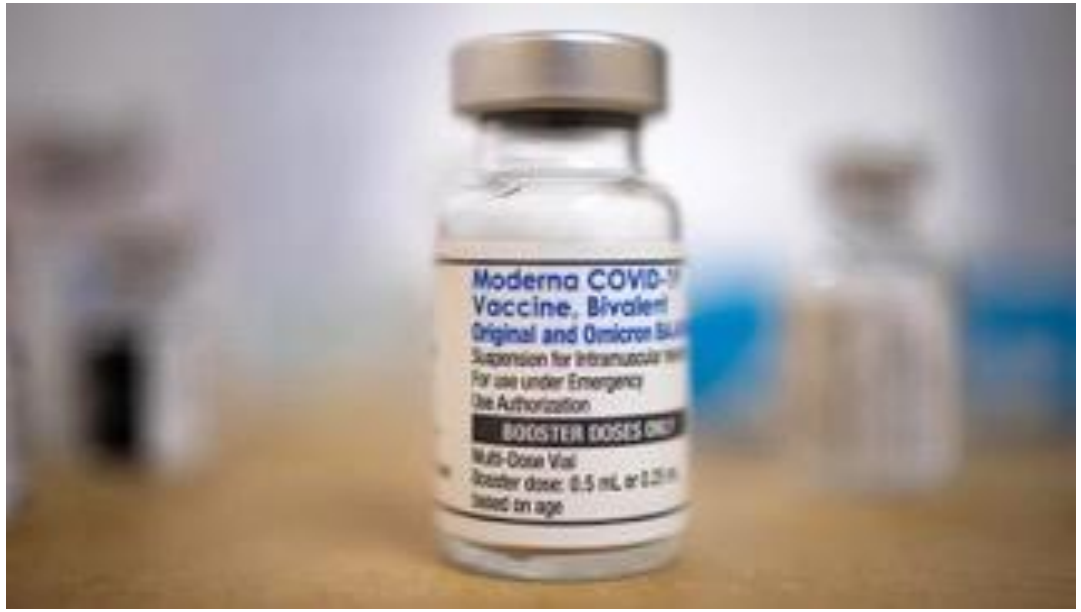




As of February 7th, 2023, COVID-19 has claimed 6,830,232 lives.

(World Health Organization); Images from cfr.org, npr.org, usatoday

Are there vaccines and treatments to help reduce the severity of COVID-19?



(Chowdhury et al., 2020); (MacMillan, 2023); (Service, 2022); Images from CNBC, medpagetoday.com, nbcnews, pharmaceutical-technology.com

Is COVID-19 still mutating?

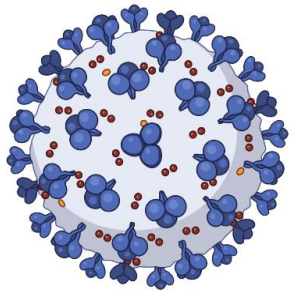
Nextstrain

Real-time tracking of pathogen evolution

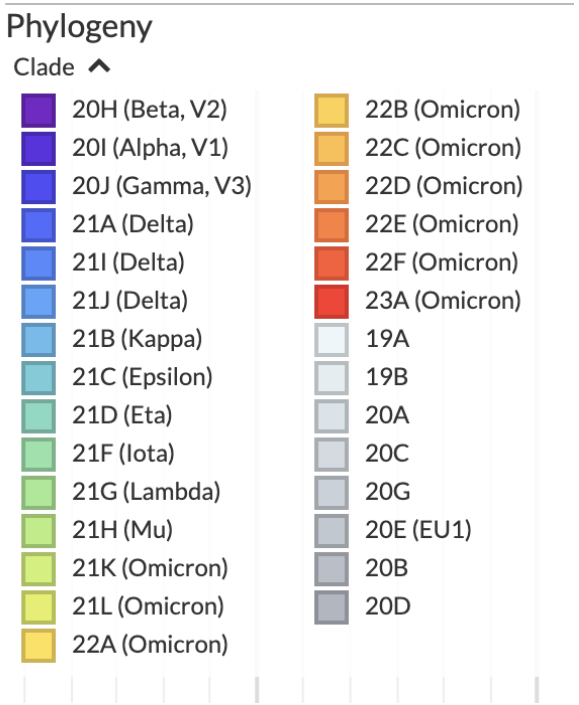
https://nextstrain.org/

Bedford

Summary



- **Over the past three years, COVID-19 has led to suffering worldwide.**
- **Information about the virus continues to evolve.**
- **Sequencing the virus using phylogenetic methods can allow experts to predict transmission trends and to create the best vaccines and treatments possible to reduce the virus's severity.**



Images from Biorender and Nextstrain

Photo and Video References

- https://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956/
- https://www.youtube.com/watch?v=TiKbMw_bKEk&ab_channel=BiologyLectures
- http://www.evolution-textbook.org/content/free/contents/ch27/ch27-f28.html
- https://www.geckoevolution.org/mbe_cover.html
- https://www.nicepng.com/maxp/u2q8o0o0w7u2r5q8/
- https://blogs.gwu.edu/himmelfarb/2021/01/18/ncbi-account-logins-changes/
- https://www.insdc.org/
- https://euraxess.ec.europa.eu/worldwide/japan/embl-interdisciplinary-postdocs-eipod4-programme-4-cofunding-marie-sk%C5%82odowska-0
- https://www.nature.com/articles/s41576-020-0233-0
- https://www.sanger.ac.uk/tool/ensembl-genome-browser/
- https://bioedit.software.informer.com/7.2/
- https://www.researchgate.net/figure/Distance-methods-of-phylogenetic-tree-construction-involve-converting-genetic-differences_fig2_43022487
- https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/long-branch-attraction

Photo and Video References

Slide 18: https://nextstrain.org/

Slide 18: http://www.iqtree.org/

Slide 18: https://gisaid.org/

Slide 21: https://my.clevelandclinic.org/health/diseases/21214-coronavirus-covid-19

Slide 22: https://www.scripps.edu/covid-19/science-simplified/parts-of-a-coronavirus/

Slide 23: https://www.youtube.com/watch?v=dA70ZdYhhCg

Slide 24: https://www.georgiacyber.org/resources/policies/health-safety/covid-19

Slide 26: https://edition.cnn.com/asia/live-news/coronavirus-outbreak-01-27-20-intl-hnk/h_c4da8563a5078f61e81957e5ba500aee

Slide 28: https://www.50states.com/washington.htm

Slide 29: https://en.wikipedia.org/wiki/List_of_counties_in_Washington

Slide 31: https://newsroom.aaa.com/2021/06/aaa-more-than-47m-americans-to-celebrate-with-an-independence-day-getaway/

Slide 31: https://www.mdanderson.org/cancerwise/what-counts-as-coronavirus-covid-19-exposure--how-does-contact-tracing-identify-who-has-been-exposed-to-covid-19.h00-159383523.html

Slide 32: nature.com/articles/s41598-021-90766-7

Slide 34: https://www.nutrigeneticsspecialists.com/single-post/2017/03/27/what-is-a-snp

Photo and Video References

Slide 40: https://www.wbur.org/news/2022/12/19/massachusetts-covid-business-shutdown-hindsight

Slide 40: https://vcresearch.berkeley.edu/news/covid-19-stimulus-good-start-more-needed-says-berkeley-ist

Slide 40: https://www.wqxr.org/story/how-sound-new-york-city-has-changed-during-covid-19-lockdown/

Slides 49, 50, 51, 52, 53, 54: https://www.cfr.org/councilofcouncils/global-memos/covid-19-after-two-years-failure-pandemic-governance

Slides 50, 51, 52, 53, 54: https://www.usatoday.com/story/news/nation/2020/05/03/coronavirus-death-count-has-doctors-struggling-mental-health/3063081001/

Slides 51, 52, 53, 54: https://www.npr.org/sections/health-shots/2022/05/17/1093651037/us-one-million-deaths

Slide 55: https://www.pharmaceutical-technology.com/features/molnupiravir-safety-questions-approvals-approach/

Slide 55: https://www.nbcnews.com/health/health-news/cdc-warns-covid-19-rebound-taking-paxlovid-antiviral-pills-rcna30311

Slide 55: https://www.medpagetoday.com/opinion/second-opinions/101432

Slide 55: https://www.cnbc.com/2022/12/06/pfizer-rejects-moderna-claim-it-copied-covid-vaccine-accuses-rival-of-rewriting-history.html

Slide 56 and 57: https://nextstrain.org/

Slide 57: https://biorender.com/

Citations of News Articles and Websites

Baldauf SL. (2003) Phylogeny for the faint of heart: a tutorial. Trends Genet. Retrieved February 7, 2023, from https://pubmed.ncbi.nlm.nih.gov/12801728/

Baum, D. (2008). *Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups*. Nature news. Retrieved February 7, 2023, from https://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956/

Bean, M., & Fischer, L. (2017, March 27). *What is a SNP?* Nutrigenetics. Retrieved February 7, 2023, from https://www.nutrigeneticsspecialists.com/single-post/2017/03/27/what-is-a-snp

Bedford et al., (2020). *Cryptic Transmission of SARS-CoV-2 in Washington state*. Retrieved February 1, 2023, from https://www.science.org/doi/10.1126/science.abc0523

Centers for Disease Control and Prevention. (2021). *Basics of Covid-19*. Centers for Disease Control and Prevention. Retrieved February 4, 2023, from https://www.cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19/basics-covid-19.html#:~:text=COVID%2D19%20is%20caused%20by,East%20respiratory%20syndrome%20(MERS).

Centers for Disease Control and Prevention. (2022, August 16). *CDC Museum Covid-19 Timeline*. Centers for Disease Control and Prevention. Retrieved February 2, 2023, from https://www.cdc.gov/museum/timeline/covid19.html

Chowdhury, J., Scarr, S., & Wardell, J. (2020, December 10). *How the novel coronavirus has evolved*. Reuters. Retrieved February 2, 2023, from https://www.reuters.com/graphics/HEALTH-CORONAVIRUS/EVOLUTION/yxmpjqkdzvr/

MacMillan, C. (2023, January 26). *The omicron booster: Your questions answered*. Yale Medicine. Retrieved February 7, 2023, from https://www.yalemedicine.org/news/omicron-booster-covid-19

Citations of News Articles and Websites (continued)

Service, R. F. (2022, June 29). *Bad news for Paxlovid? Coronavirus can find multiple ways to evade COVID-19 drug*. Retrieved February 7, 2023, from https://www.science.org/content/article/bad-news-paxlovid-coronavirus-can-find-multiple-ways-evade-covid-19-drug?cookieSet=1

Scripps Research. (2020). *What are the parts of a coronavirus?* What are the parts of a coronavirus? | Scripps Research. Retrieved February 5, 2023, from https://www.scripps.edu/covid-19/science-simplified/parts-of-a-coronavirus/

Scripps Research. (2020). *How does the novel coronavirus infect a cell? | Scripps Research*. Retrieved February 7, 2023, from https://www.scripps.edu/covid-19/science-simplified/how-the-novel-coronavirus-infects-a-cell/#:~:text=The%20microscopic%20virus%20enters%20through,on%20to%20the%20ACE2%20receptor.

World Health Organization- Regional Office for Africa. (2021). *Coronavirus*. World Health Organization. Retrieved February 5, 2023, from https://www.afro.who.int/publications/coronavirus

World Health Organization. (n.d.). *Who Coronavirus (COVID-19) dashboard*. World Health Organization. Retrieved February 5, 2023, from https://covid19.who.int/