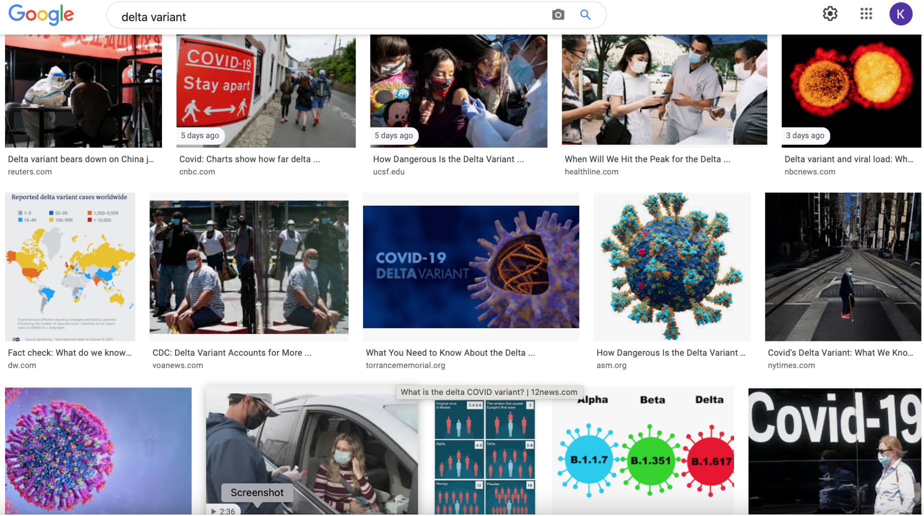
Tracking Recombination in the Evolution of SARS-CoV-2

Kristina Moen

UC Davis Math REU 2021

August 11, 2021

Motivation - A Deadly and Mutating Virus



delta variant

5 days ago

Delta variant bears down on China j...
reuters.com

Covid: Charts show how far delta ...
cnbc.com

5 days ago

How Dangerous Is the Delta Variant ...
ucsf.edu

When Will We Hit the Peak for the Delta ...
healthline.com

3 days ago

Delta variant and viral load: Wh...
nbcnews.com

Fact check: What do we know...
dw.com

CDC: Delta Variant Accounts for More ...
voanews.com

What You Need to Know About the Delta ...
torrancememorial.org

How Dangerous Is the Delta Variant ...
asm.org

Covid's Delta Variant: What We Kno...
nytimes.com

What is the delta COVID variant? | 12news.com

Screenshot

2:36

Outline

Introduction to Recombination and Topological Data Analysis

- Molecular Evolution and Recombination
- Topological Data Analysis and Generators

Applying Topological Data Analysis to SARS-CoV-2 RNA Sequences

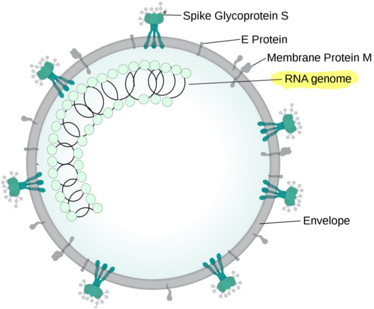
- Collecting and Preparing Data
- Results

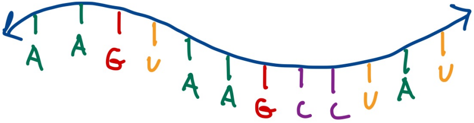
Simulations of Homologous Recombination

- Missing Recombination Events
- Identifying Recombinants from Generators

Future Research

SARS-CoV-2 is a Single-Stranded RNA Virus





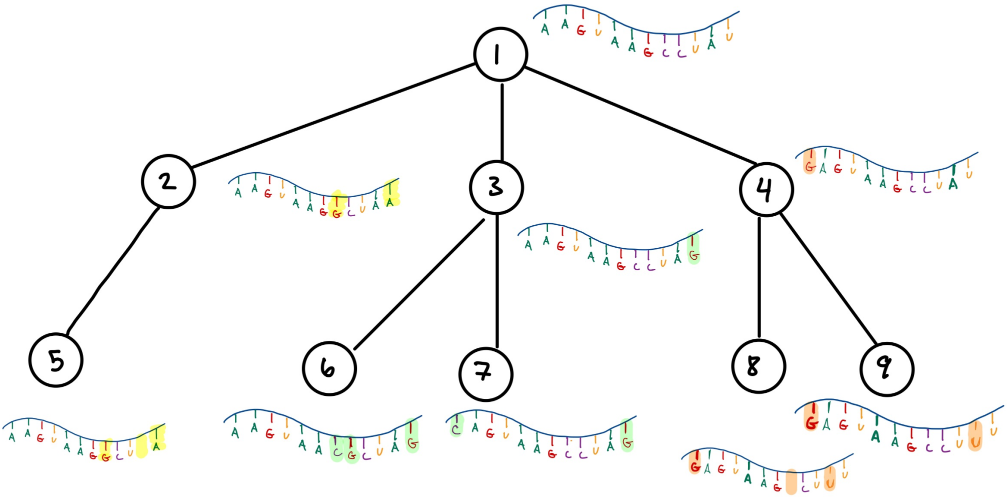
Bases:

Adenine
Cytosine
Uracil
Guanine

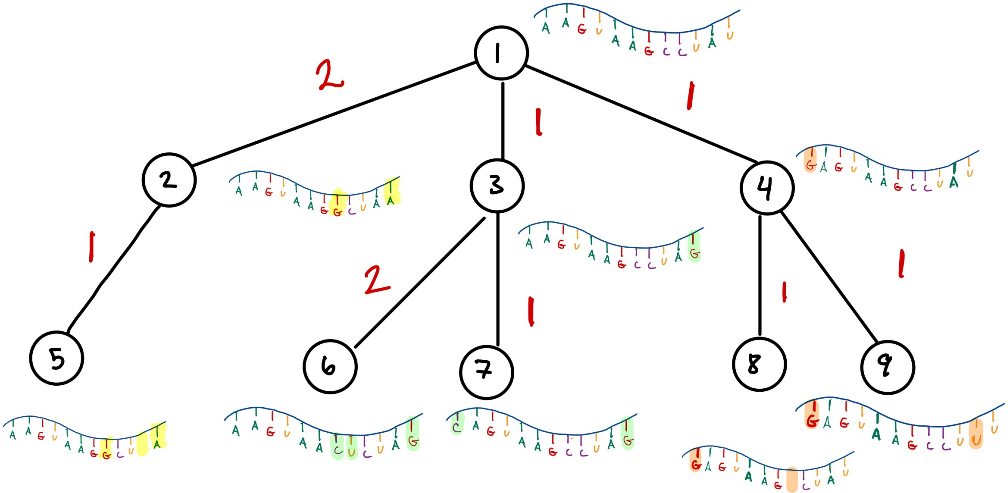
~30,000 nucleotides

{A, C, U, G}

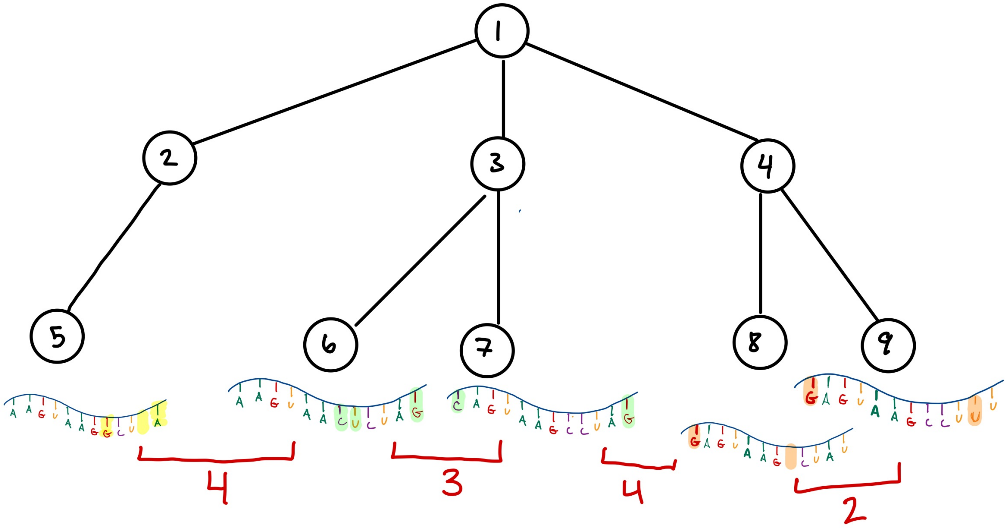
Point Mutations and Phylogenetic Tree



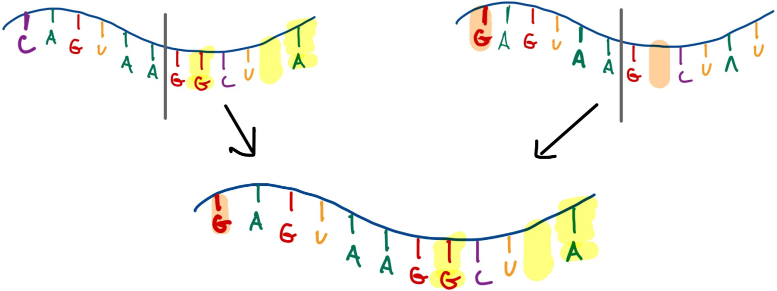
Hamming Distance Along Edges of Phylogenetic Tree - Point Mutations Add Up Through Vertical Evolution



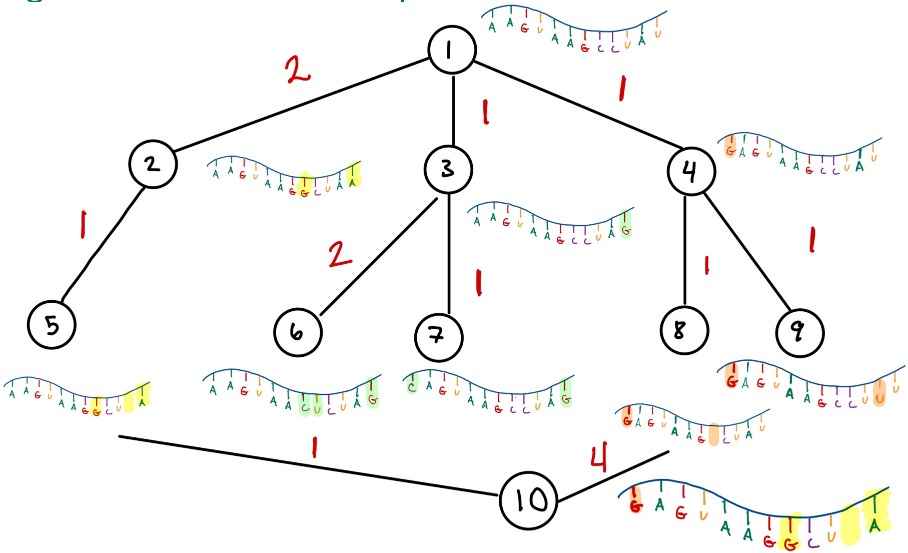
Basic Problem of Phylogenetics - Inferring Evolutionary History



Horizontal Evolution - Recombination is an Exchange of Genetic Information



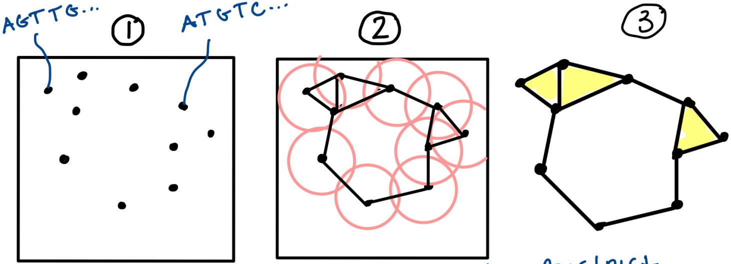
Phylogenetic Trees Fail to Capture Recombination



[1]

[1] Chan, J., et al., 2013. Topology of viral evolution. PNAS.

Topological Data Analysis as a Tool for Finding 1-Dimensional Holes



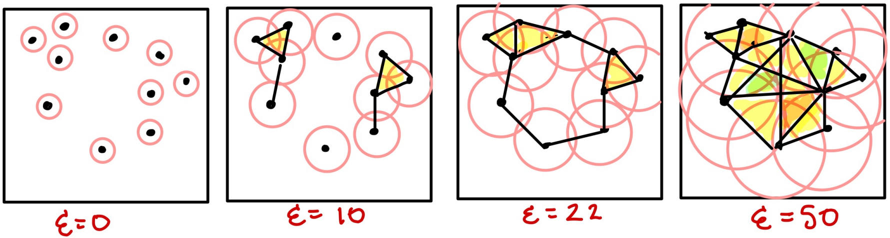
1. Embed sequences in a metric space

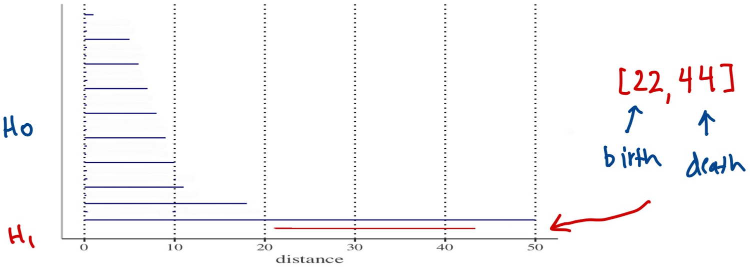
2. Draw balls around points, and connect points with intersecting balls

3. Construct simplicial complex

4. Compute homology groups to find connected components and 1-dimensional holes

Persistent Homology Can Be Visualized as a Barcode

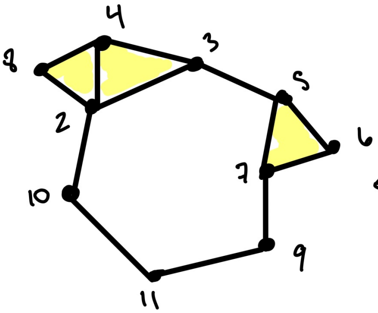




[2]

[2] Ghrist, R., 2007. Barcodes: The persistent topology of data. Bulletin of AMS.

Generators of a Homology Group



H_1 generator :

Nodes and edges that bound a 1-dimensional hole at birth

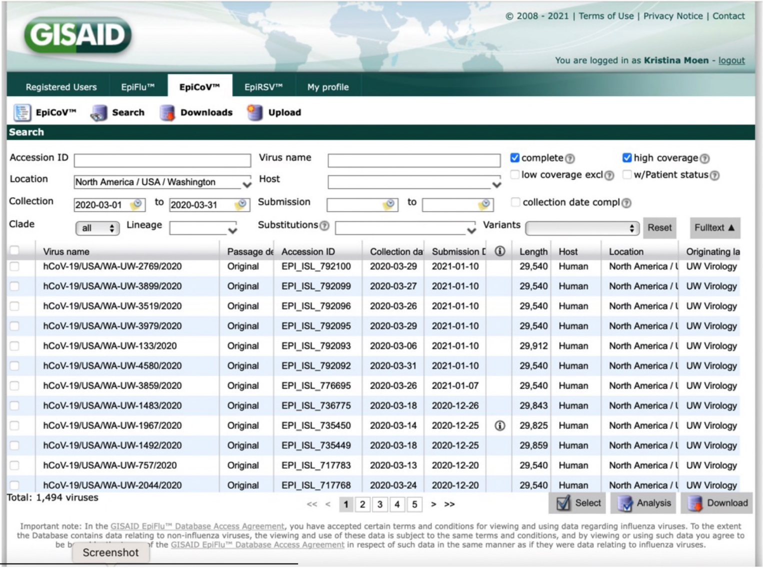
← Possible generator :

[3, 5]
[5, 7]
[7, 9]
[9, 11]
[11, 10]
[10, 2]
[2, 3]
[3, 5]

First Question

What can topological data analysis tell us about the evolution of SARS-CoV-2?

Collecting SARS-CoV-2 Sequences


GISAID
© 2008 - 2021 | Terms of Use | Privacy Notice | Contact
You are logged in as **Kristina Moen** - logout
Registered Users
EpiFlu™
EpiCoV™
EpiRSV™
My profile
EpiCoV™
Search
Downloads
Upload

Search

Accession ID:

Virus name:

☑ complete

☑ high coverage

☐ low coverage excl

☐ w/Patient status

Location: North America / USA / Washington

Host:

Collection: 2020-03-01 to 2020-03-31

Submission: to

☐ collection date compl

Clade: all

Lineage:

Substitutions:

Variants:

☐	Virus name	Passage de	Accession ID	Collection da	Submission D	ⓘ	Length	Host	Location	Originating la
☐	hCoV-19/USA/WA-UW-2769/2020	Original	EPI_ISL_792100	2020-03-29	2021-01-10		29,540	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-3899/2020	Original	EPI_ISL_792099	2020-03-27	2021-01-10		29,540	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-3519/2020	Original	EPI_ISL_792096	2020-03-26	2021-01-10		29,540	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-3979/2020	Original	EPI_ISL_792095	2020-03-29	2021-01-10		29,540	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-133/2020	Original	EPI_ISL_792093	2020-03-06	2021-01-10		29,912	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-4580/2020	Original	EPI_ISL_792092	2020-03-31	2021-01-10		29,540	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-3859/2020	Original	EPI_ISL_776695	2020-03-26	2021-01-07		29,540	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-1483/2020	Original	EPI_ISL_736775	2020-03-18	2020-12-26		29,843	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-1967/2020	Original	EPI_ISL_735450	2020-03-14	2020-12-25	ⓘ	29,825	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-1492/2020	Original	EPI_ISL_735449	2020-03-18	2020-12-25		29,859	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-757/2020	Original	EPI_ISL_717783	2020-03-13	2020-12-20		29,540	Human	North America / U	UW Virology
☐	hCoV-19/USA/WA-UW-2044/2020	Original	EPI_ISL_717768	2020-03-24	2020-12-20		29,540	Human	North America / U	UW Virology

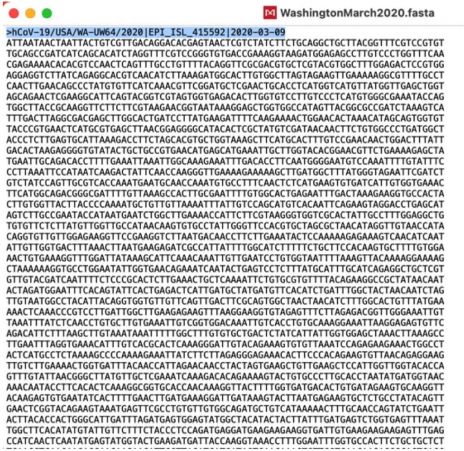
Total: 1,494 viruses

Important note: In the GISAID EpiFlu™ Database Access Agreement, you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be b f the GISAID EpiFlu™ Database Access Agreement in respect of such data in the same manner as if they were data relating to influenza viruses.

[3]

[3] www.gisaid.org

Data is Downloaded as Text File

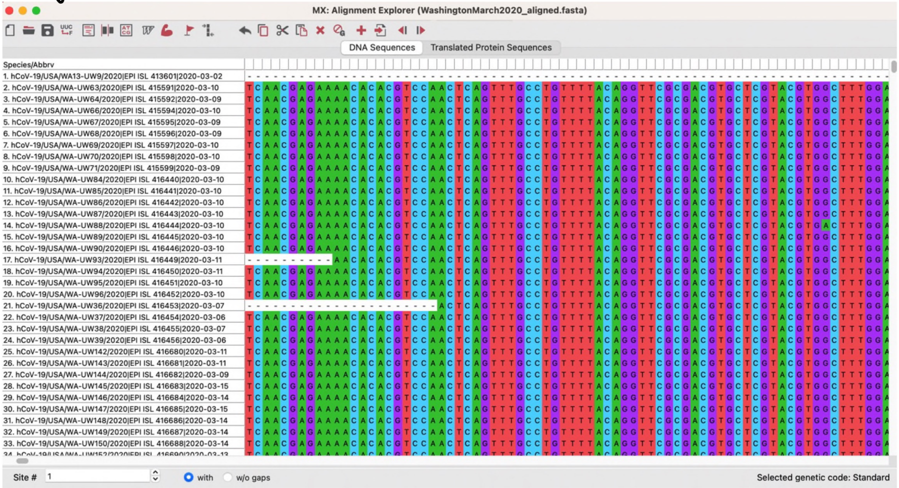


```
>hCoV-19/USA/WA-UW64/2020|EPI_ISL_415592|2020-03-09
ATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGT
TGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAA
CGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGG
AGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCT
CAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGT
AGCAGAACTCGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAG
TGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCA
TTTGACTTAGGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGT
TACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCT
ACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTGTCCGAACAACTGGACTTTATT
GACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTA
TGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTC
CCTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCT
GTCTATCCAGTTGCGTCACCAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAAC
TTCATGGCAGACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCACTA
CTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGCATGTCACAATTCAGAAGTAGGACCTGAGCAT
AGTCTTGCCGAATACCATAATGAATCTGGCTTGAAAACCATTCTTCGTAAGGGTGGTCGCACTATTGCCTTTGGAGGCTG
TGTGTTCTCTTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCACGTGCTAGCGCTAACATAGGTTGTAACCATA
CAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACAACCTTCTTGAAATACTCCAAAAAGAGAAAGTCAACATCAAT
ATTGTTGGTGACTTTAAACTTAATGAAGAGATCGCCATTATTTTGGCATCTTTTTCTGCTTCCACAAGTGCTTTTGTGGA
AACTGTGAAAGGTTTGGATTATAAAGCATTCAAACAAATTGTTGAATCCTGTGGTAATTTTAAAGTTACAAAAGGAAAAG
CTAAAAAAGGTGCCTGGAATATTGGTGAACAGAAATCAATACTGAGTCCTCTTTATGCATTTGCATCAGAGGCTGCTCGT
GTTGTACGATCAATTTTCTCCCGCACTCTTGAAACTGCTCAAAATTCTGTGCGTGTTTTACAGAAGGCCGCTATAACAAT
ACTAGATGGAATTTCACAGTATTCACTGAGACTCATTGATGCTATGATGTTCACATCTGATTTGGCTACTAACAATCTAG
TTGTAATGGCCTACATTACAGGTGGTGTTGTTCAGTTGACTTCGCAGTGGCTAACTAACATCTTTGGCACTGTTTATGAA
AAACTCAAACCCGTCCTTGATTGGCTTGAAGAGAAGTTTAAGGAAGGTGTAGAGTTTCTTAGAGACGGTTGGGAAATTGT
TAAATTTATCTCAACCTGTGCTTGTGAAATTGTCGGTGGACAAATTGTCACCTGTGCAAAGGAAATTAAGGAGAGTGTTC
AGACATTCTTTAAGCTTGTAAATAAATTTTTGGCTTTGTGTGCTGACTCTATCATTATTGGTGGAGCTAAACTTAAAGCC
TTGAATTTAGGTGAAACATTTGTCACGCACTCAAAGGGATTGTACAGAAAGTGTGTTAAATCCAGAGAAGAAACTGGCCT
ACTCATGCCTCTAAAAGCCCCAAAAGAAATTATCTTCTTAGAGGGAGAAACACTTCCCACAGAAGTGTTAACAGAGGAAG
TTGTCTTGAAAACTGGTGATTTACAACCATTAGAACAACCTACTAGTGAAGCTGTTGAAGCTCCATTGGTTGGTACACCA
GTTTGTATTAACGGGCTTATGTTGCTCGAAATCAAAGACACAGAAAAGTACTGTGCCCTTGCACCTAATATGATGGTAAC
AAACAATACCTTCACACTCAAAGGCGGTGCACCAACAAAGGTTACTTTTGGTGATGACACTGTGATAGAAGTGCAAGGTT
ACAAGAGTGTGAATATCACTTTTGAACTTGATGAAAGGATTGATAAAGTACTTAATGAGAAGTGCTCTGCCTATACAGTT
GAACTCGGTACAGAAGTAAATGAGTTCGCCTGTGTTGTGGCAGATGCTGTCATAAAAACTTTGCAACCAGTATCTGAATT
ACTTACACCACTGGGCATTGATTTAGATGAGTGGAGTATGGCTACATACTACTTATTTGATGAGTCTGGTGAGTTTAAAT
TGGCTTCACATATGTATTGTTCTTTCTACCCTCCAGATGAGGATGAAGAAGAAGGTGATTGTGAAGAAGAAGAGTTTGAG
CCATCAACTCAATATGAGTATGGTACTGAAGATGATTACCAAGGTAAACCTTTGGAATTTGGTGCCACTTCTGCTGCTCT
```

FASTA
file

Align Sequences Before Computing Distance

Mega X: Molecular Evolutionary Genetic Analysis



MX: Alignment Explorer (WashingtonMarch2020_aligned.fasta)

Species/Abbrv	
1. hCoV-19/USA/WA13-UW9/2020|EPI ISL 413601|2020-03-02	--
2. hCoV-19/USA/WA-UW63/2020|EPI ISL 415591|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
3. hCoV-19/USA/WA-UW64/2020|EPI ISL 415592|2020-03-09	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
4. hCoV-19/USA/WA-UW66/2020|EPI ISL 415594|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
5. hCoV-19/USA/WA-UW67/2020|EPI ISL 415595|2020-03-09	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
6. hCoV-19/USA/WA-UW68/2020|EPI ISL 415596|2020-03-09	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
7. hCoV-19/USA/WA-UW69/2020|EPI ISL 415597|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
8. hCoV-19/USA/WA-UW70/2020|EPI ISL 415598|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
9. hCoV-19/USA/WA-UW71/2020|EPI ISL 415599|2020-03-09	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
10. hCoV-19/USA/WA-UW84/2020|EPI ISL 416440|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
11. hCoV-19/USA/WA-UW85/2020|EPI ISL 416441|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
12. hCoV-19/USA/WA-UW86/2020|EPI ISL 416442|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
13. hCoV-19/USA/WA-UW87/2020|EPI ISL 416443|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
14. hCoV-19/USA/WA-UW88/2020|EPI ISL 416444|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGACTTTGGA
15. hCoV-19/USA/WA-UW89/2020|EPI ISL 416445|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
16. hCoV-19/USA/WA-UW90/2020|EPI ISL 416446|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
17. hCoV-19/USA/WA-UW93/2020|EPI ISL 416449|2020-03-11	----------AACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
18. hCoV-19/USA/WA-UW94/2020|EPI ISL 416450|2020-03-11	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
19. hCoV-19/USA/WA-UW95/2020|EPI ISL 416451|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
20. hCoV-19/USA/WA-UW96/2020|EPI ISL 416452|2020-03-10	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
21. hCoV-19/USA/WA-UW36/2020|EPI ISL 416453|2020-03-07	----------------------ACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
22. hCoV-19/USA/WA-UW37/2020|EPI ISL 416454|2020-03-06	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
23. hCoV-19/USA/WA-UW38/2020|EPI ISL 416455|2020-03-07	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
24. hCoV-19/USA/WA-UW39/2020|EPI ISL 416456|2020-03-06	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
25. hCoV-19/USA/WA-UW142/2020|EPI ISL 416680|2020-03-11	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
26. hCoV-19/USA/WA-UW143/2020|EPI ISL 416681|2020-03-11	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
27. hCoV-19/USA/WA-UW144/2020|EPI ISL 416682|2020-03-09	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
28. hCoV-19/USA/WA-UW145/2020|EPI ISL 416683|2020-03-15	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
29. hCoV-19/USA/WA-UW146/2020|EPI ISL 416684|2020-03-14	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
30. hCoV-19/USA/WA-UW147/2020|EPI ISL 416685|2020-03-15	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
31. hCoV-19/USA/WA-UW148/2020|EPI ISL 416686|2020-03-14	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
32. hCoV-19/USA/WA-UW149/2020|EPI ISL 416687|2020-03-14	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
33. hCoV-19/USA/WA-UW150/2020|EPI ISL 416688|2020-03-14	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA
34. hCoV-19/USA/WA-UW152/2020|EPI ISL 416690|2020-03-13	TCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGA

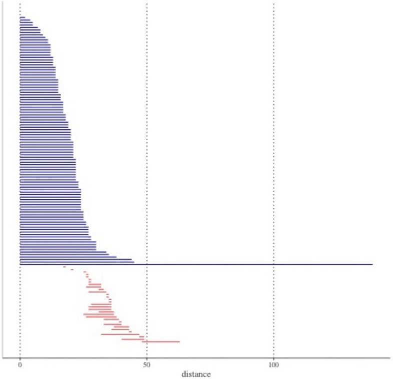
Site # 1 ● with ○ w/o gaps

Selected genetic code: Standard

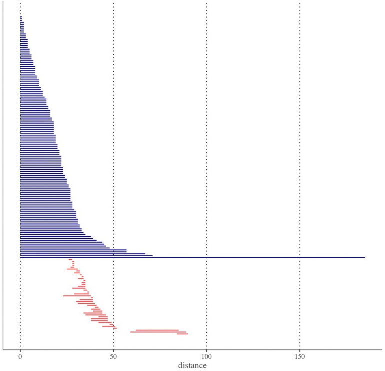
[4]

[4] Alignment support by Sofia Jacovcevic

TDA Results Were Inconclusive



Contra Costa -
100 sequences



San Francisco - 129
sequences

5

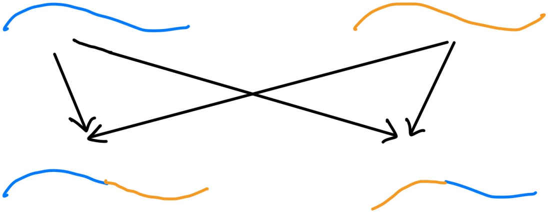
[5] TDA code (R) by Teresa Díaz Jordá, thesis on resources slide

Second Question

What do TDA barcode lengths and generators reveal about the underlying data?

Simulation Method

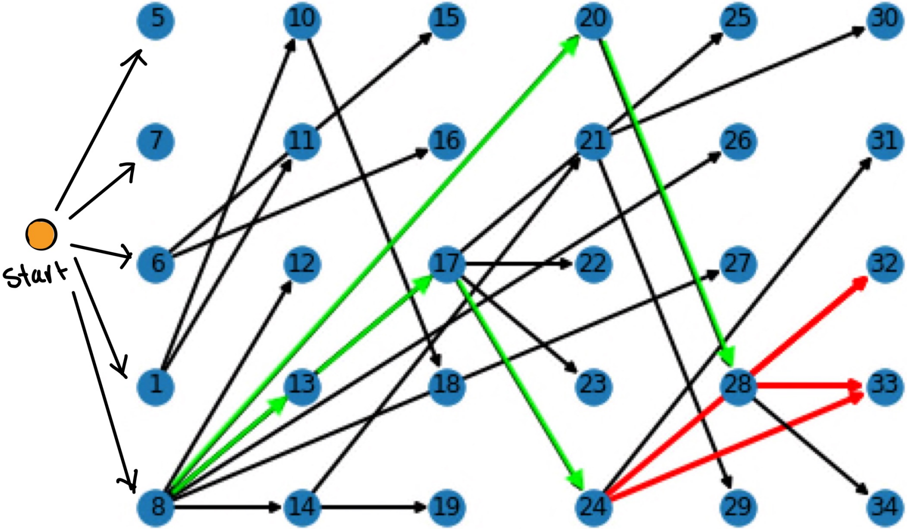
- Start with randomly generated 1,000-length sequence
- Choose population size, number of generations, mutation rate, survivors per generation
- 1 homologous recombination event per simulation cut at 50%



[6]

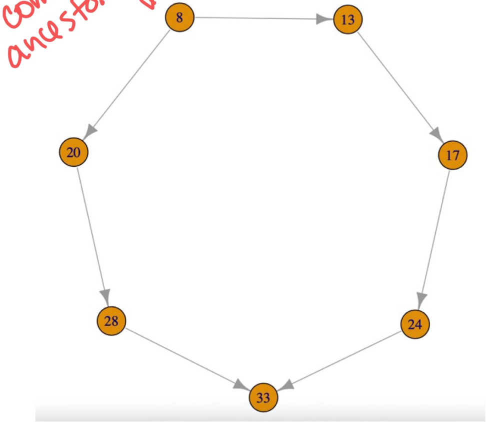
[6] Simulation code (Python) and support by Nathan Solomon

Example of Simulation

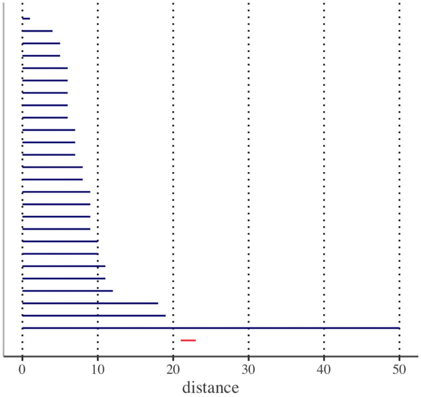


Recombination Cycle and TDA Results

common ancestor

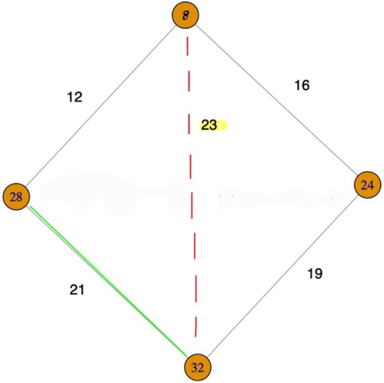


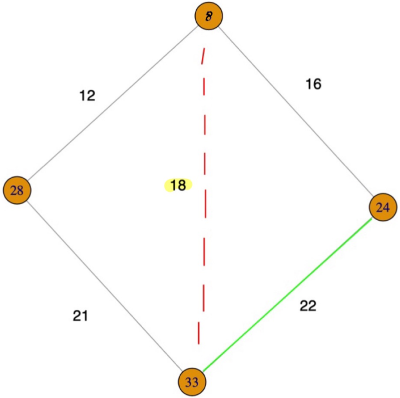
recombinant



Explaining the Missing Recombination and Hole

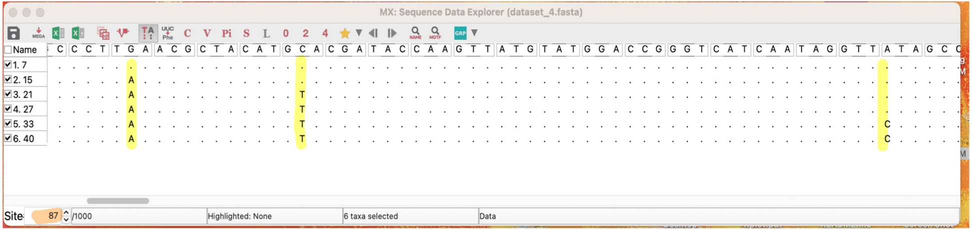
H_1 Birth: 21 Generator: [28:32], [24:32]
Death: 23 [8 : 24], [8: 28]





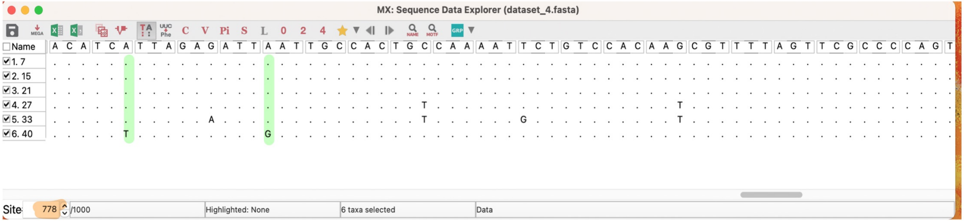
No hole formed

An Aside: How a Recombinant Can Be Genetically Closer to a Common Ancestor than Its Direct Ancestor



```
□Name  C CCT TGA ACG CTA CAT GCA CGA TAC CAA GTT ATG TAT GGA CCG GGT CAT CAA TAG GTT ATA GCC
☑1. 7  . ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
☑2. 15 . ... .A. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
☑3. 21 . ... .A. ... ... ... .T. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
☑4. 27 . ... .A. ... ... ... .T. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
☑5. 33 . ... .A. ... ... ... .T. ... ... ... ... ... ... ... ... ... ... ... ... ... C.. ...
☑6. 40 . ... .A. ... ... ... .T. ... ... ... ... ... ... ... ... ... ... ... ... ... C.. ...
```

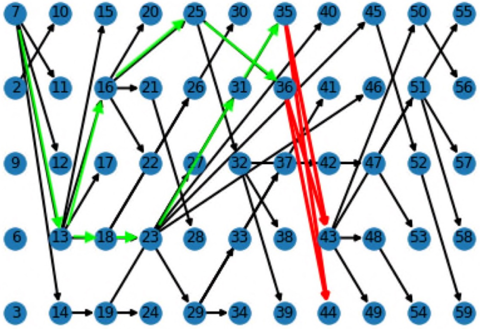
Site: 87 /1000 | Highlighted: None | 6 taxa selected | Data

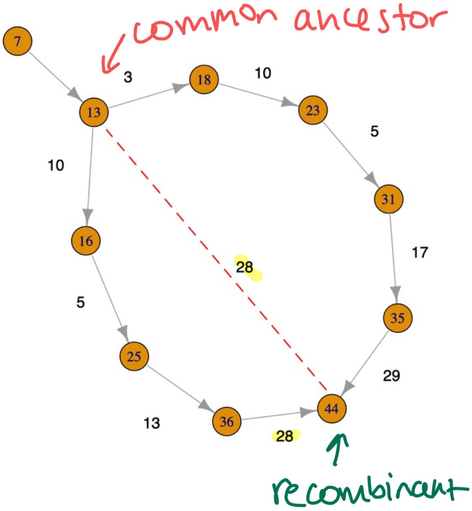


```
□Name  ACA TCA TTA GAG ATT AAT TGC CAC TGC CAA AAT TCT GTC CAC AAG CGT TTT AGT TCG CCC CAG T
☑1. 7  ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .
☑2. 15 ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .
☑3. 21 ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .
☑4. 27 ... ... ... ... ... ... ... ... ..T ... ... ... ... ... ..T ... ... ... ... ... ... .
☑5. 33 ... ... ... ..A ... ... ... ... ..T ... ... G.. ... ... ..T ... ... ... ... ... ... .
☑6. 40 ... ..T ... ... ... G.. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .
```

Site: 778 /1000 | Highlighted: None | 6 taxa selected | Data

Another Example of Missed Recombination

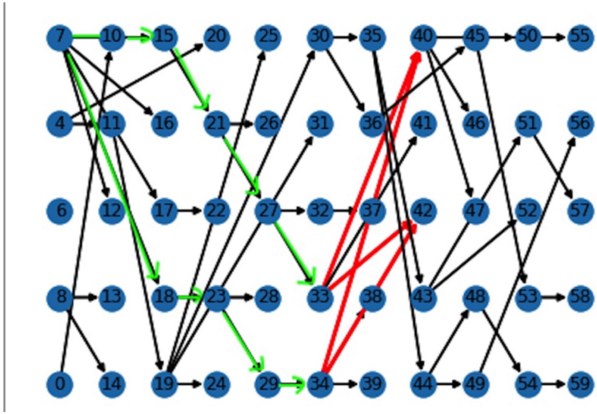


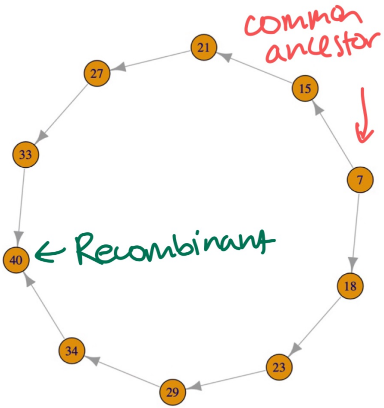


Hypothesis to Explain Missed Homologous Recombinations

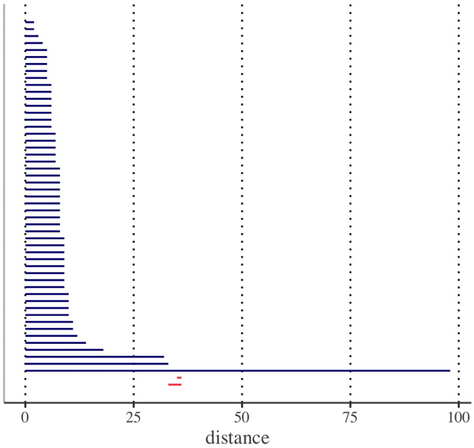
Homologous recombination is missed when the recombinant sequence has closer genetic distance to a common ancestor than to one or more of its direct ancestors.

Another Simulation Example to Develop Understanding of Generators

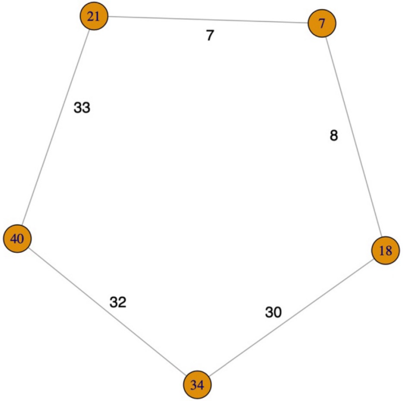




TDA Results of Simulation



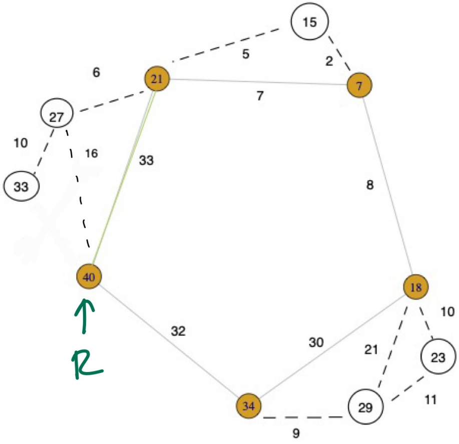
Generator



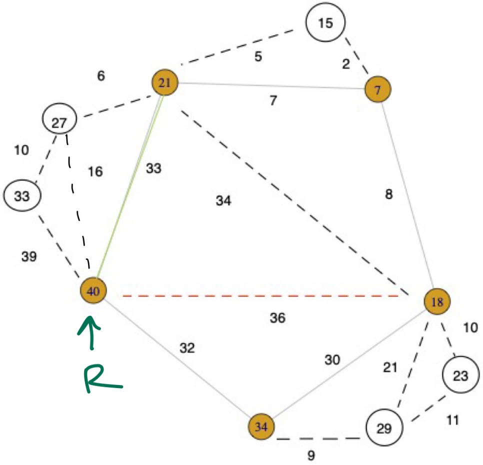
Birth: 33

Death: 36

Where Are the Missing Ancestors In the Generator?

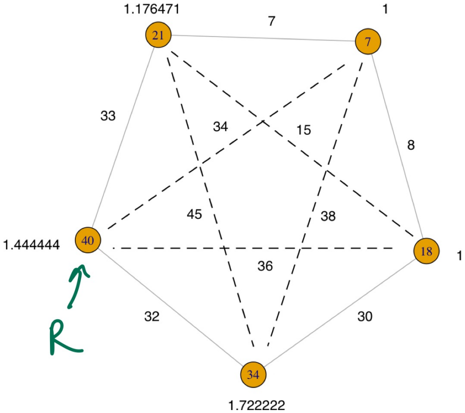


Birth: 33



Death: 36

Conjecture of Relationship Between Recombinant and Other Sequences in Generator



$$i \in \mathbb{Z}/n\mathbb{Z}$$

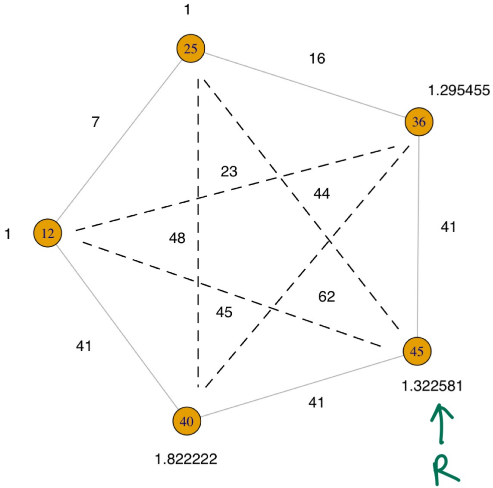
$$R_i = \frac{d(V_i, V_{i-1}) + d(V_i, V_{i+1})}{d(V_{i-1}, V_{i+1})}$$

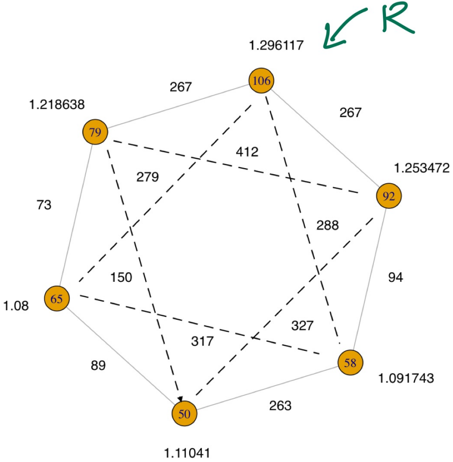
R-score of V_i

$$\frac{R_i + R_{i-1} + R_{i+1}}{3}$$

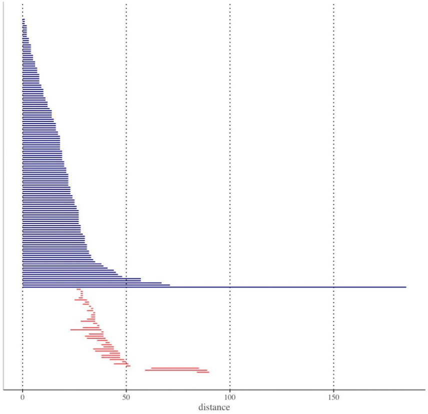
Greatest R-score: likely recombinant

Further Examples of Generator Conjecture



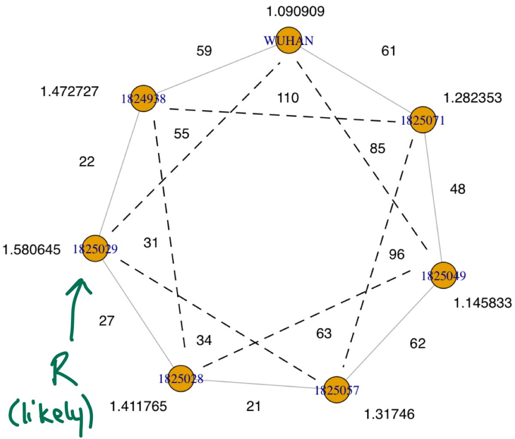


Cycling Back - Looking at the California Data



San Francisco
Feb - April 2021

San Francisco - H1 Cycle - Birth: 62, Death: 85, Length: 23



R
(likely)

Future Research

- Develop Generator Conjecture - Run simulations with different types of recombinations, metrics, complexes.
- Analyze SARS-CoV-2 data with respect to H1 generators - Compare suspected recombinants point-by-point with other sequences in the generator.

THANK YOU!

Dr. Javier Arsuaga, Dr. Mariel Vazquez, Sofia Jakovcevic, Nathan Solomon, Michael Keith, Emil Geisler, Georgina Gonzalez, Arsuaga Vazquez Lab, Greg Kuperberg and the UC Davis REU

References

Allman, E., Rhodes, J., 2016. Lecture notes: the mathematics of phylogenetics, IAS/Park City Mathematics Institute and University of Alaska Fairbanks.

Chan, J., Carlsson, G. and Rabadan, R., 2013. Topology of viral evolution. Proceedings of the National Academy of Sciences, 110(46), pp.18566-18571.

Ghrist, R., 2007. Barcodes: The persistent topology of data. Bulletin of the American Mathematical Society, 45(01), pp.61-76.

Jordá, Teresa Díaz, 2020. 'Characterization of Horizontal Evolution of RNA Viruses Using Topological Data Analysis', Bachelor's Thesis, University of California, Davis and Universitat Politècnica de València.