Applications of Bioinformatics in the Real-Time Molecular Surveillance of Viral Pathogens

Niema Moshiri
Assistant Teaching Professor
Computer Science & Engineering
University of California, San Diego

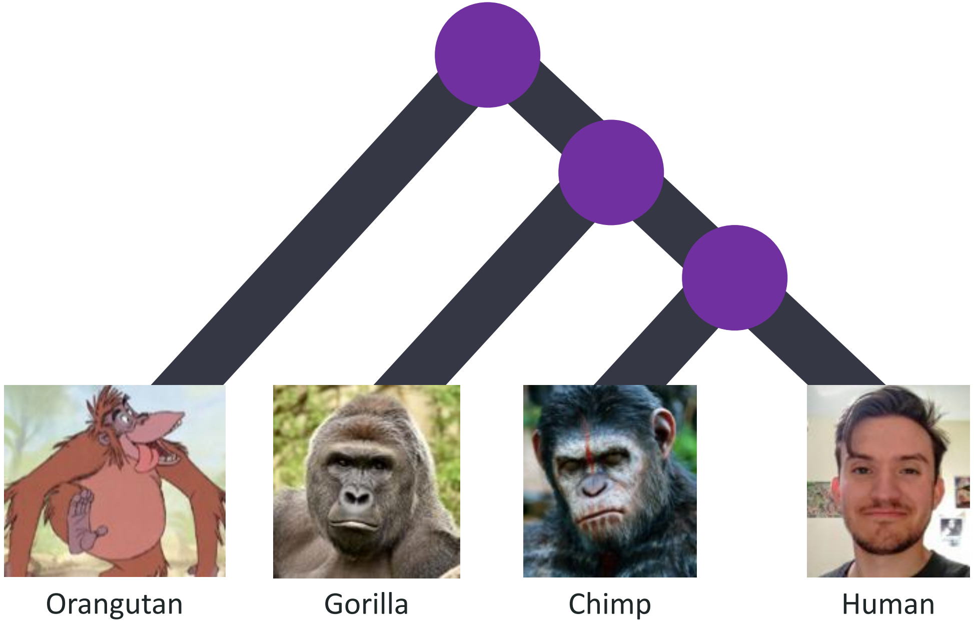
Outline

- Introduction to Viral Molecular Epidemiology
- Sequencing the First Viral Genome
- Annotating a Viral Genome
- Sequencing in the Midst of an Epidemic
- Aligning Viral Genome Sequences
- Phylogenetic Inference and Transmission Clustering

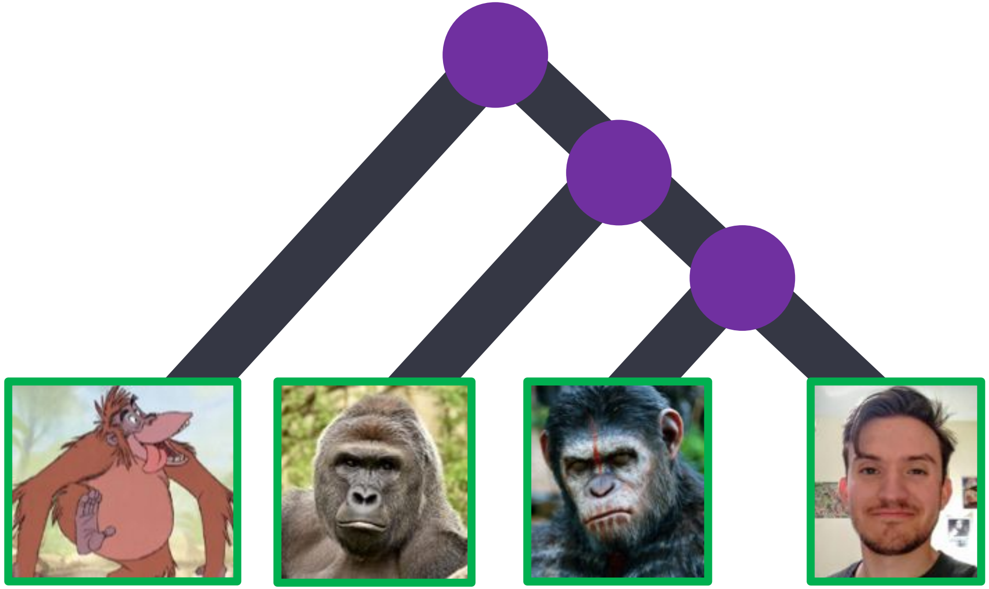
Outline

- **Introduction to Viral Molecular Epidemiology**
- Sequencing the First Viral Genome
- Annotating a Viral Genome
- Sequencing in the Midst of an Epidemic
- Aligning Viral Genome Sequences
- Phylogenetic Inference and Transmission Clustering

Phylogeny



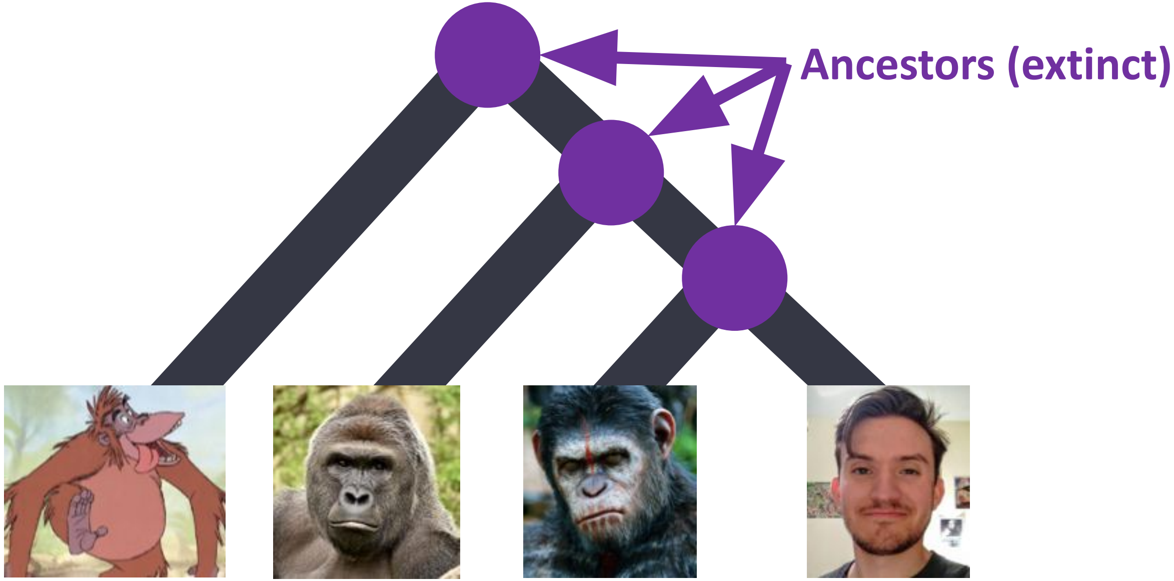
Phylogeny



Orangutan Gorilla Chimp Human

Present-Day Species

Phylogeny



Ancestors (extinct)

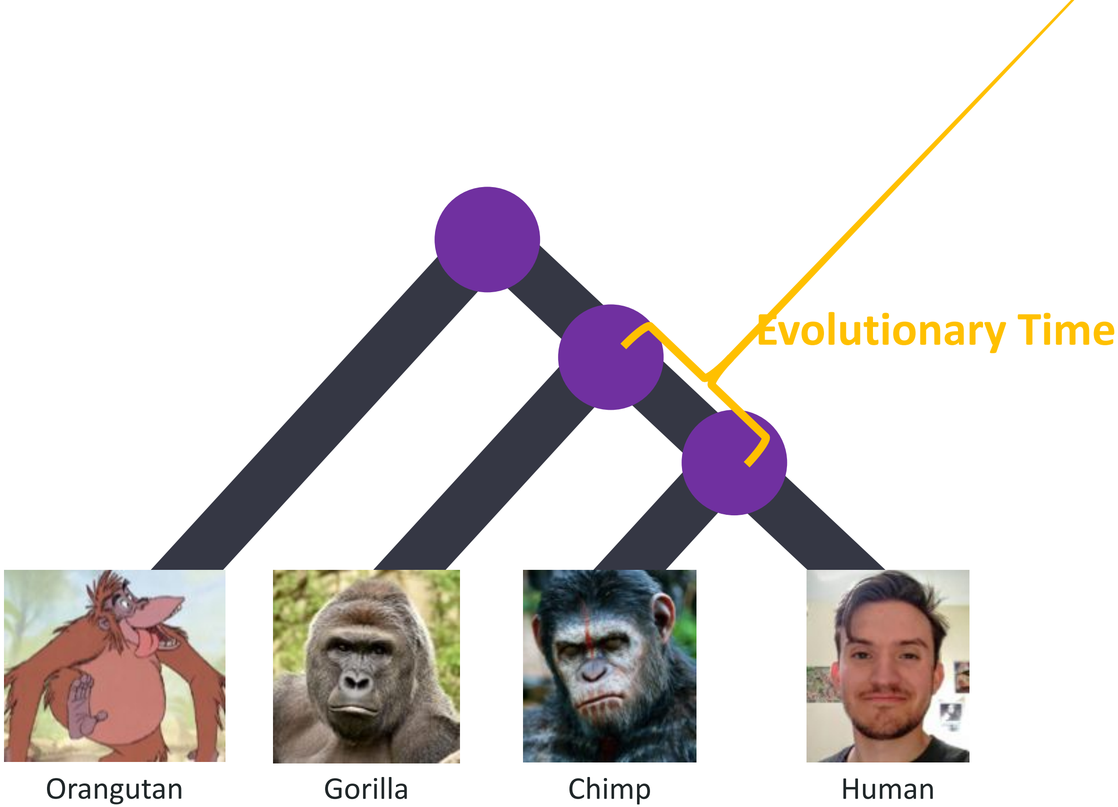
Orangutan

Gorilla

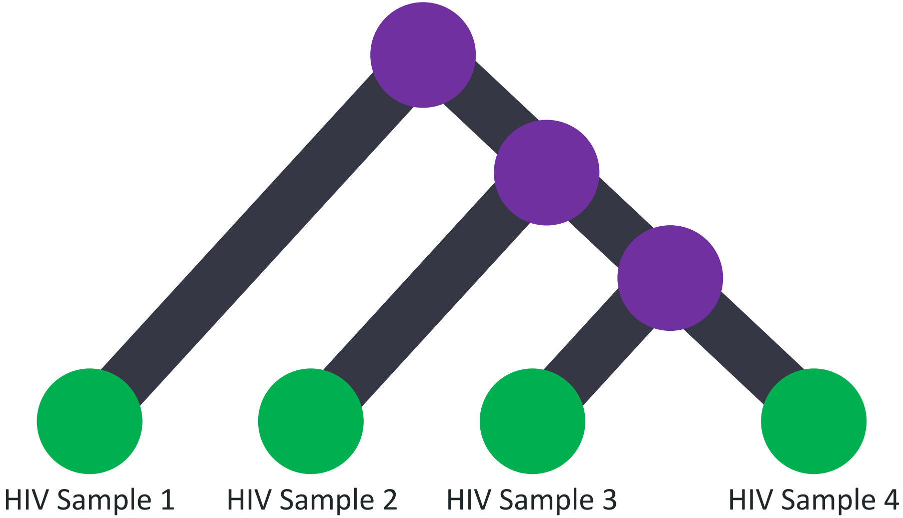
Chimp

Human

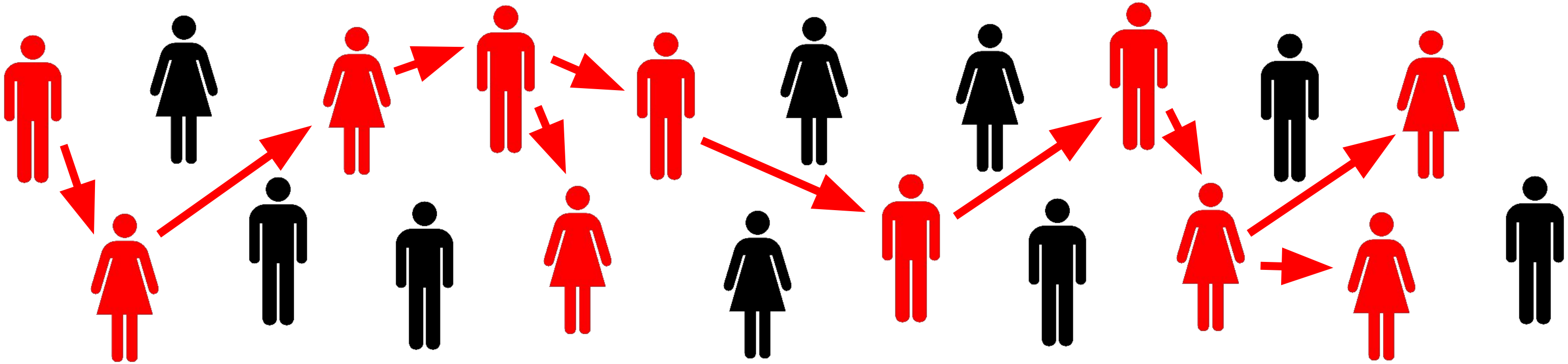
Phylogeny



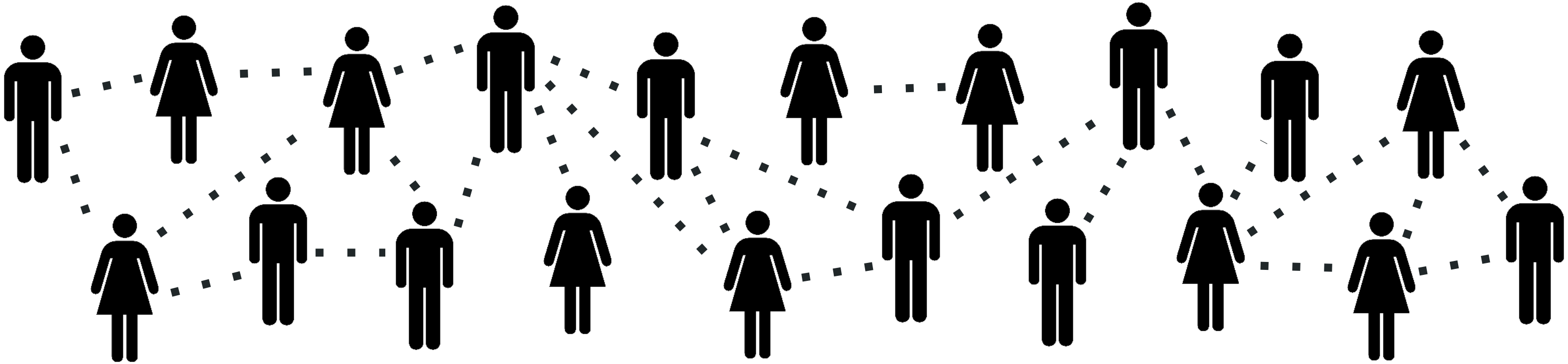
Phylogeny



Epidemic



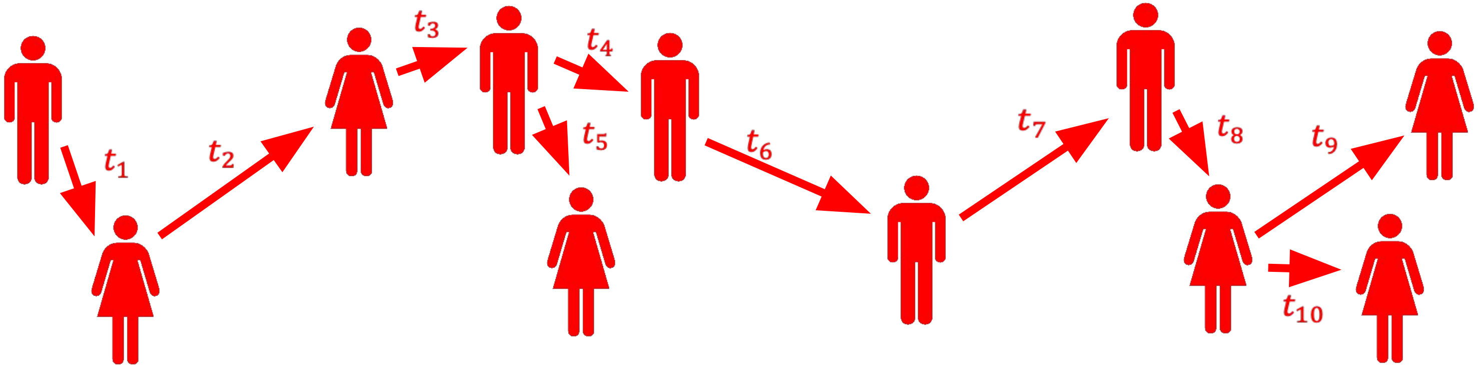
Contact Network



Nodes: Individuals

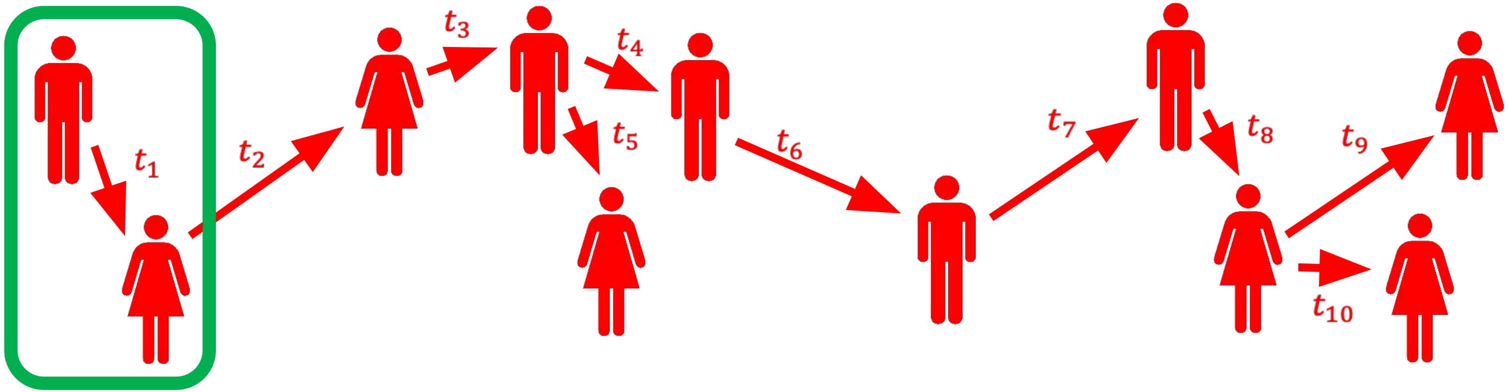
Edges: Risky contacts

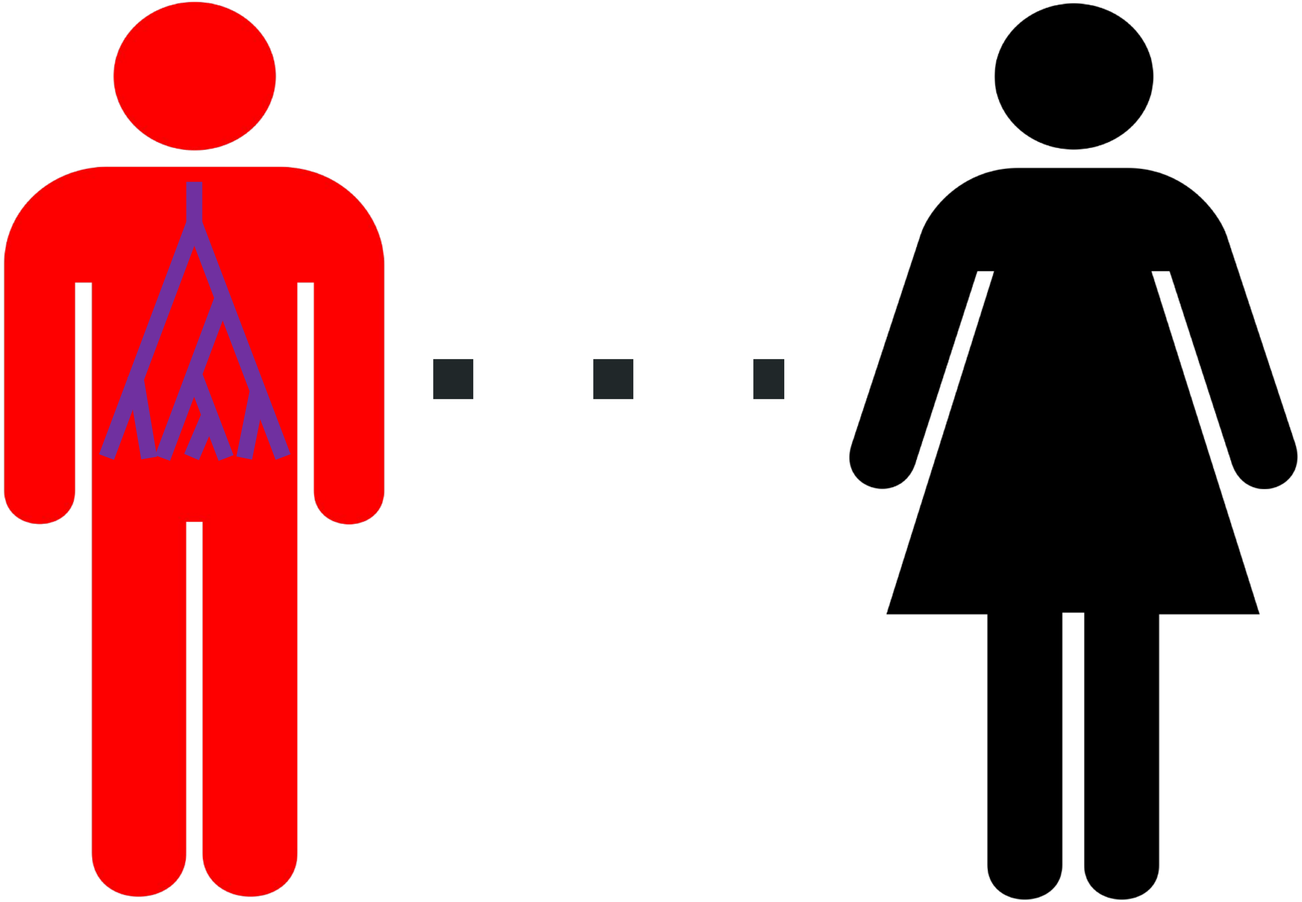
Transmission Network

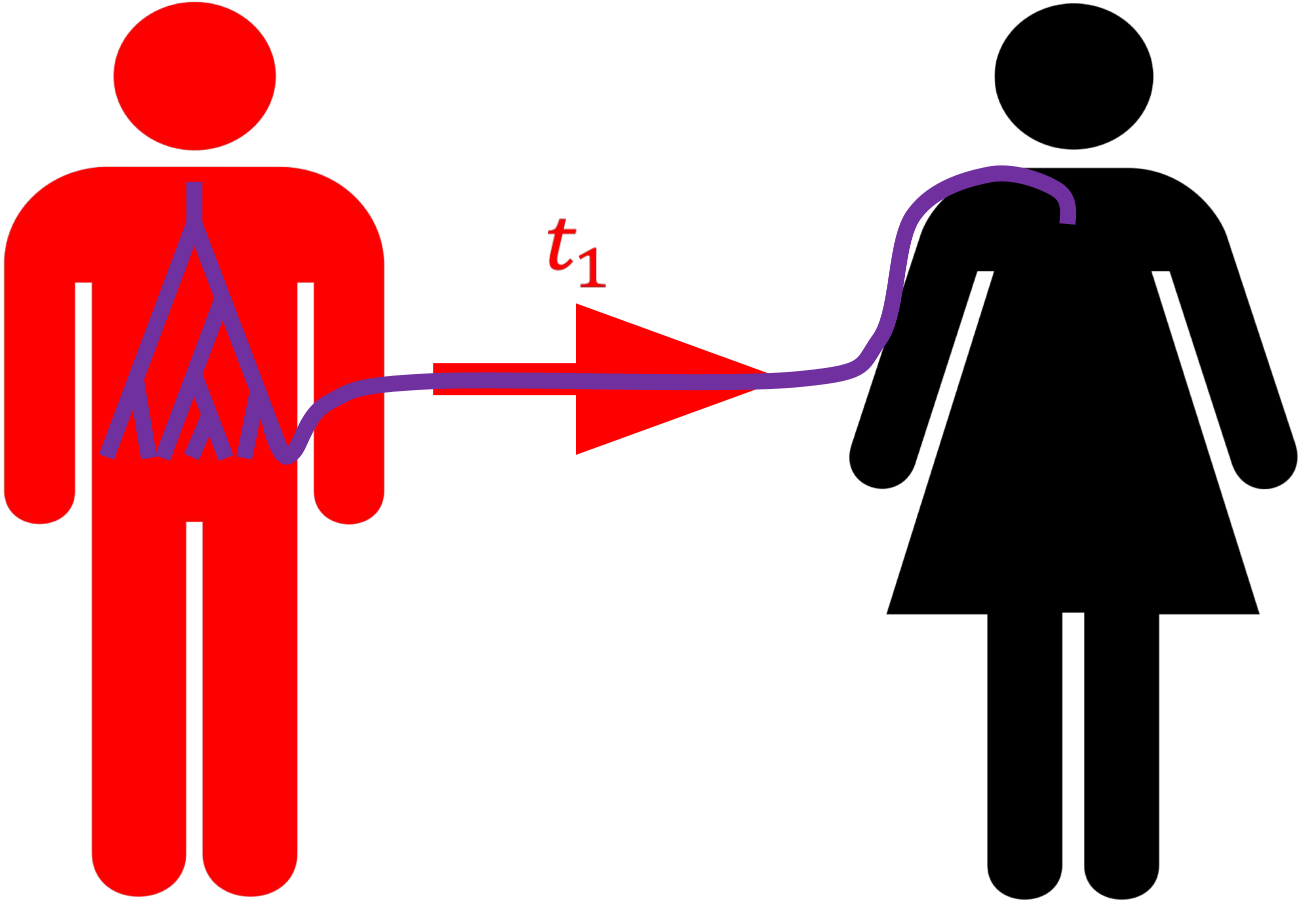


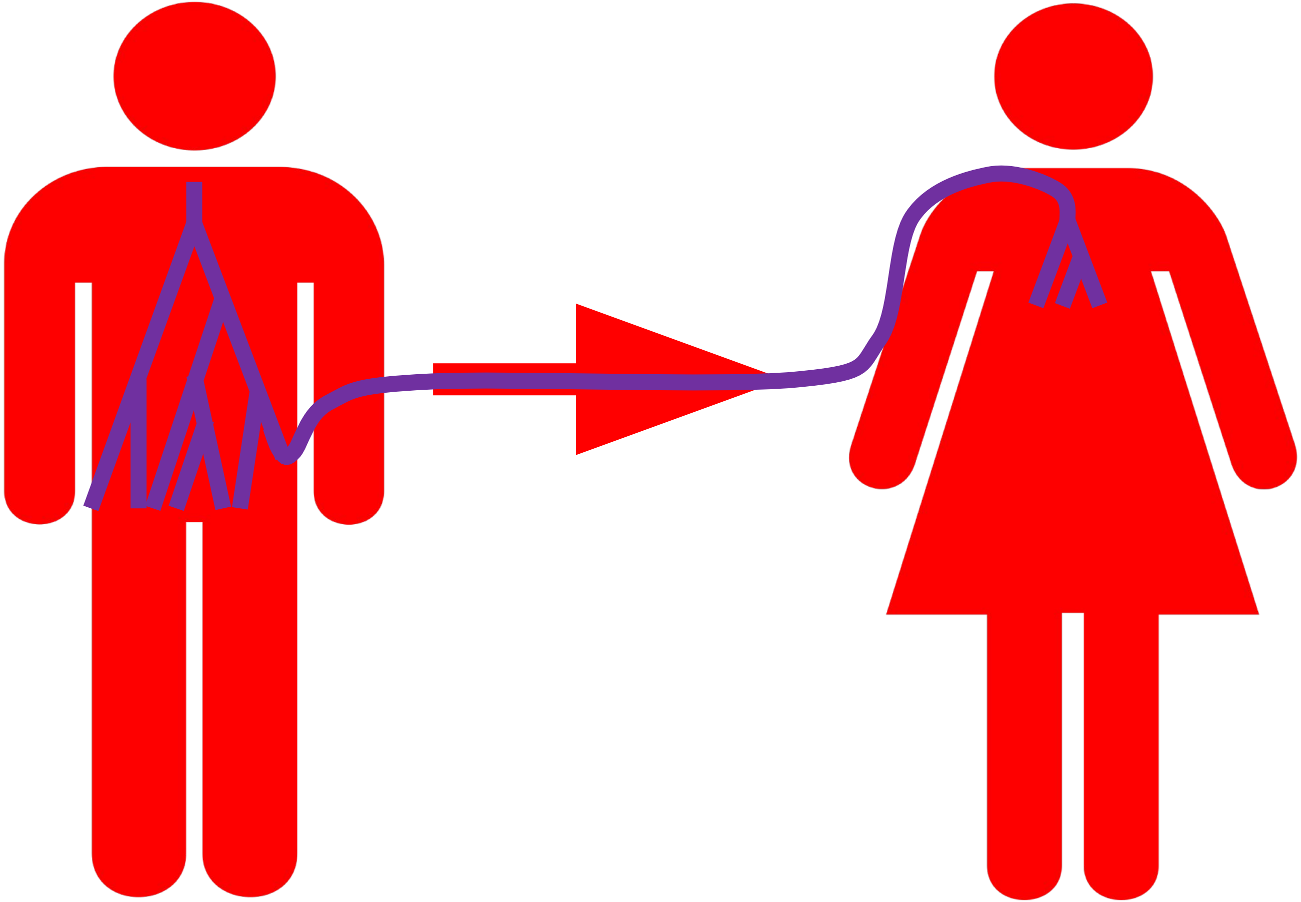
Nodes: Individuals
Edges: Transmissions

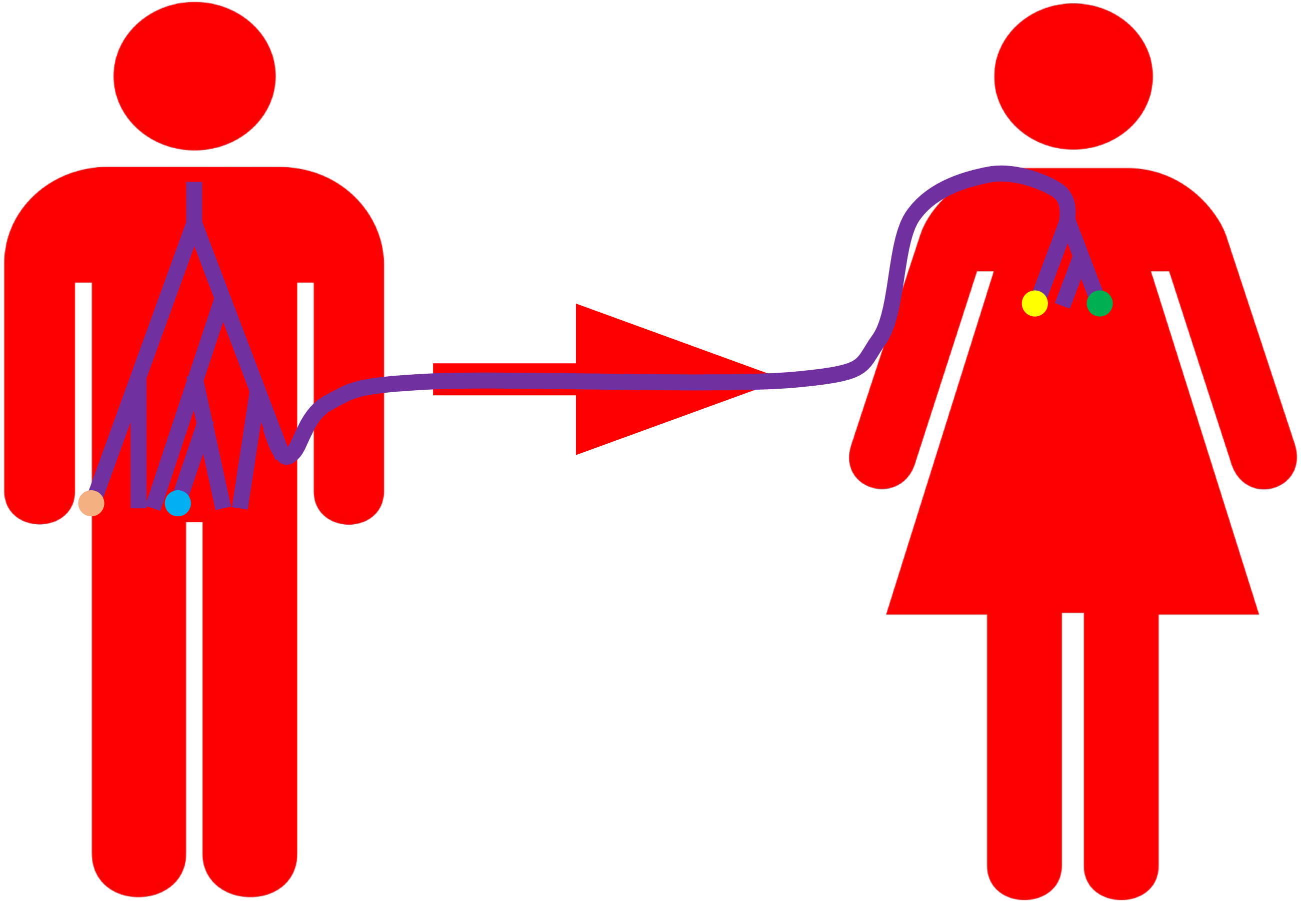
Individual Transmission Event

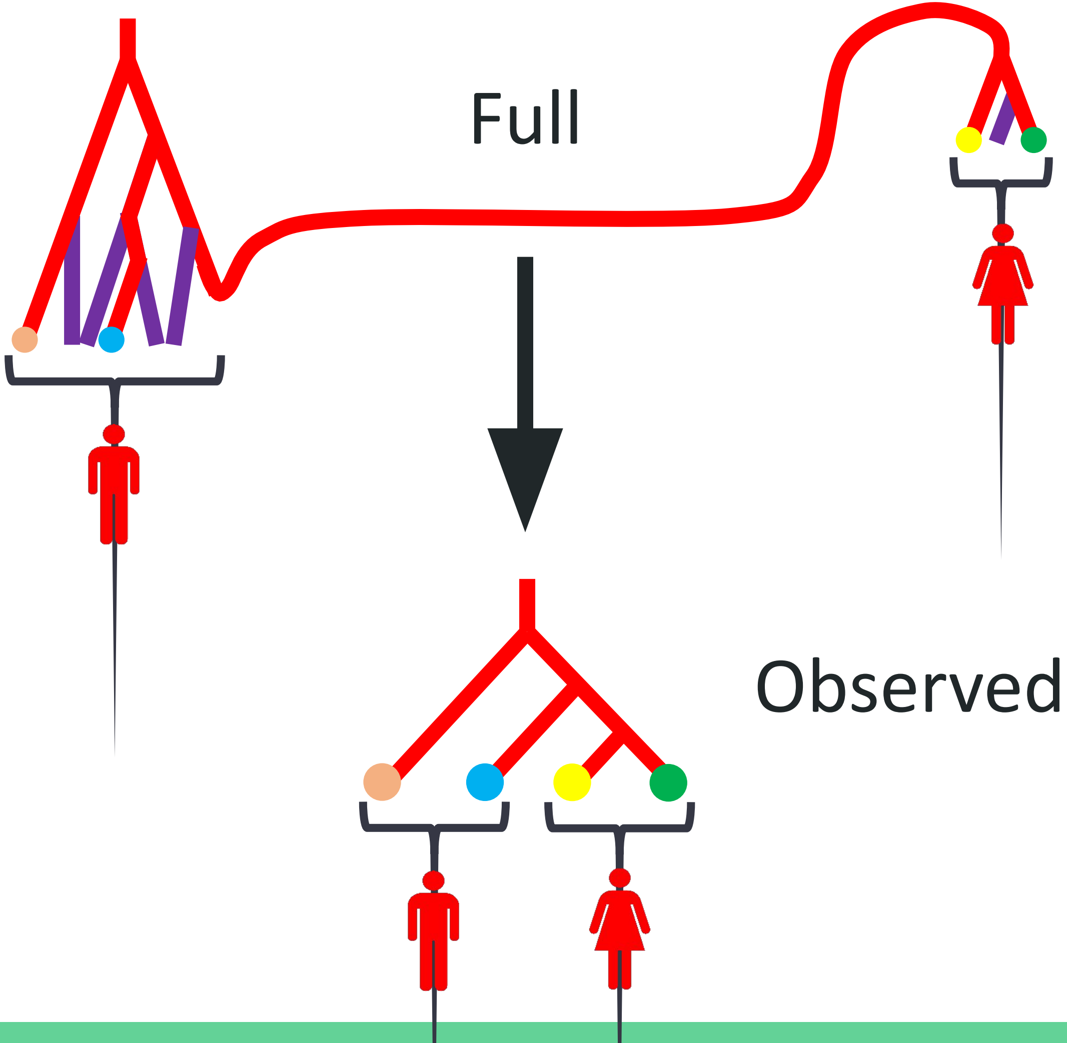





t_1

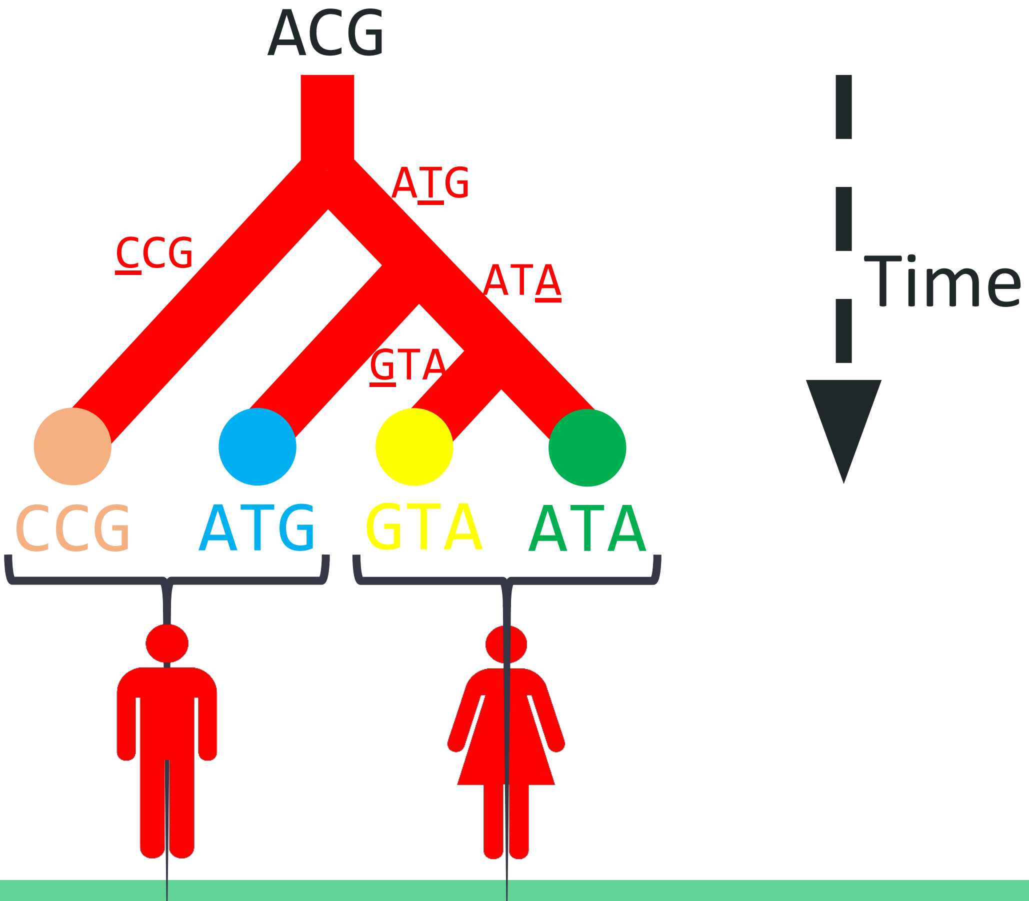






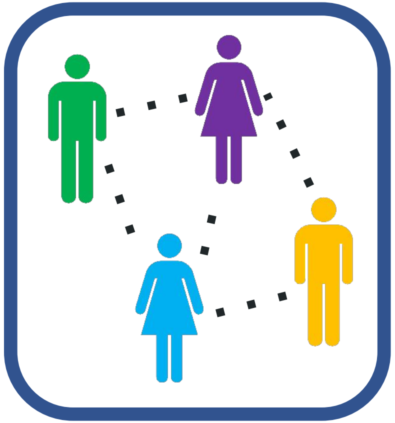
Full

Observed


ACG
ATG
CCG
ATA
GTA
Time
CCG
ATG
GTA
ATA

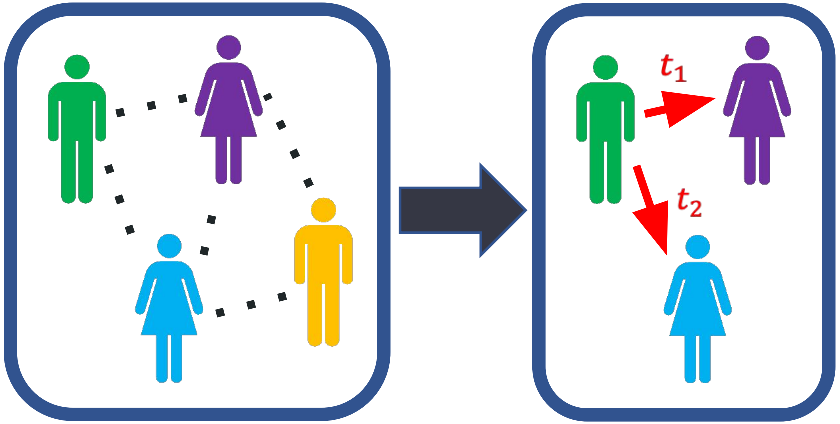
Generative Process

Generative Process



Contact Network

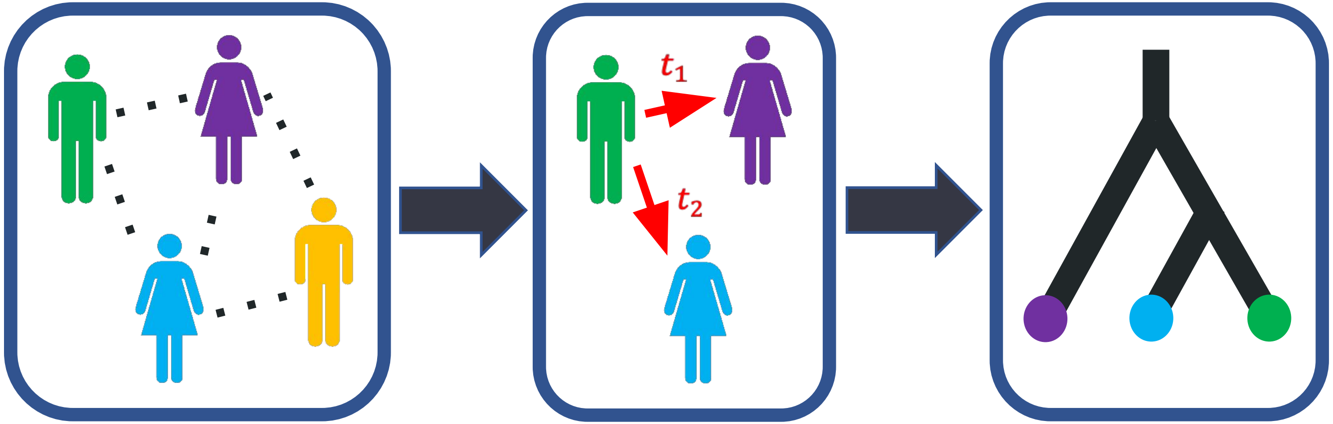
Generative Process



Contact Network

Transmission Network

Generative Process

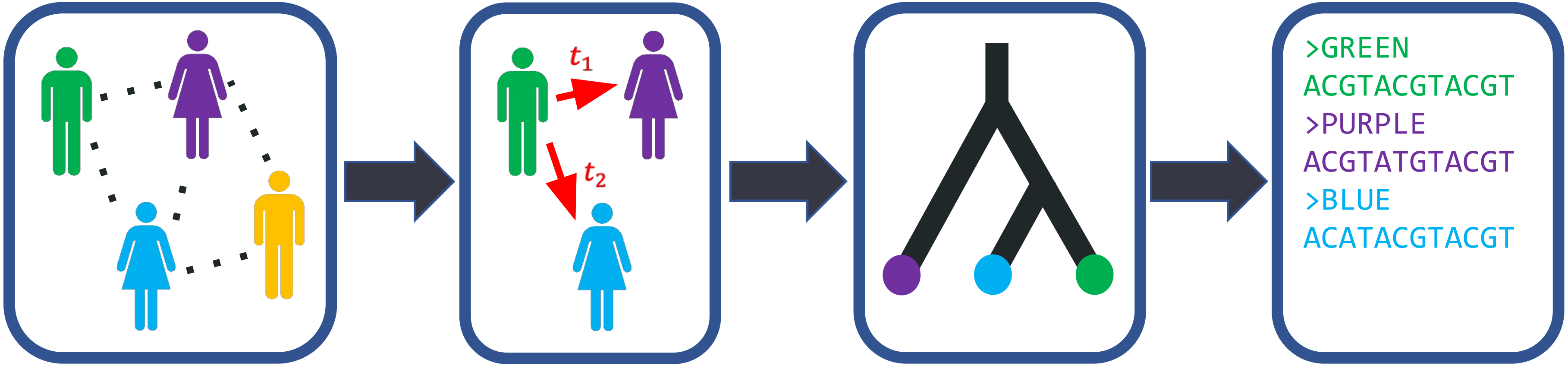


Contact Network

Transmission Network

Viral Phylogeny

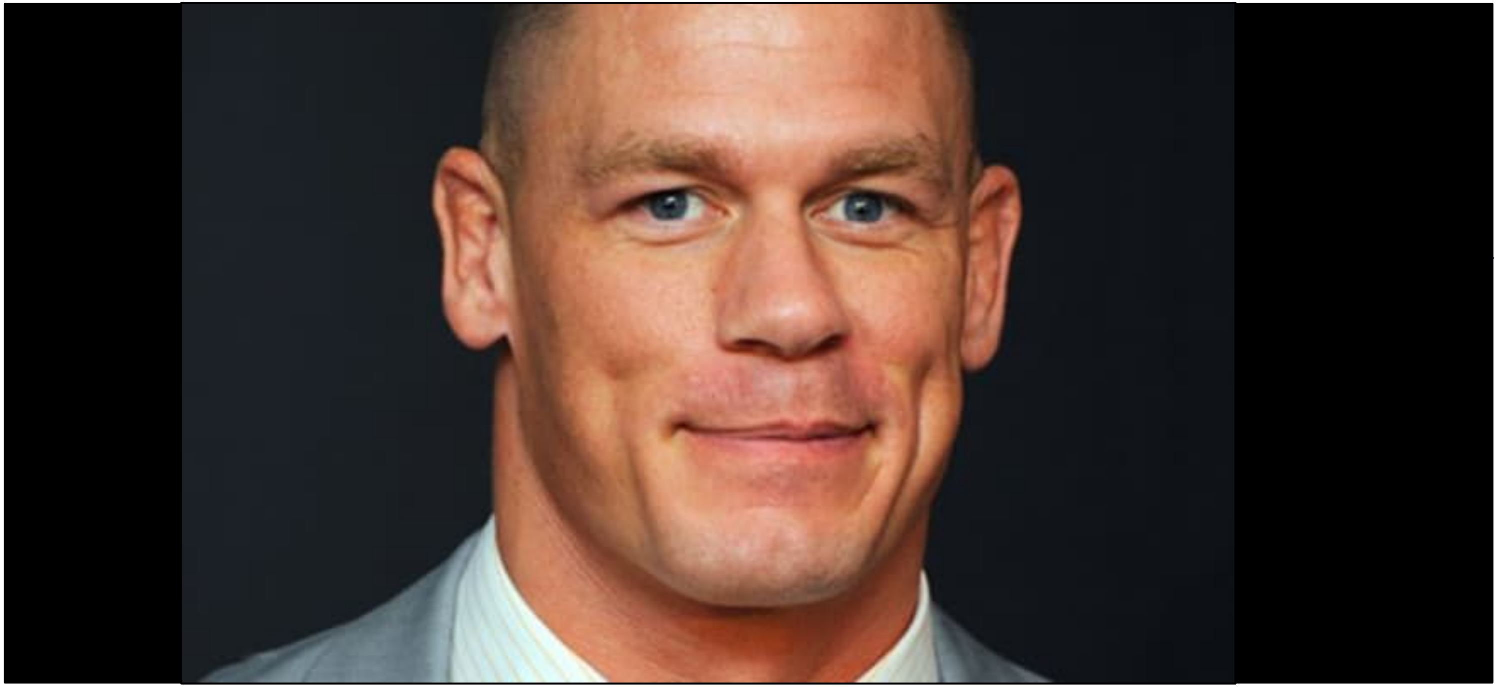
Generative Process



```
>GREEN
ACGTACGTACGT
>PURPLE
ACGTATGTACGT
>BLUE
ACATACGTACGT
```

Contact Network → Transmission Network → Viral Phylogeny → Viral Sequences

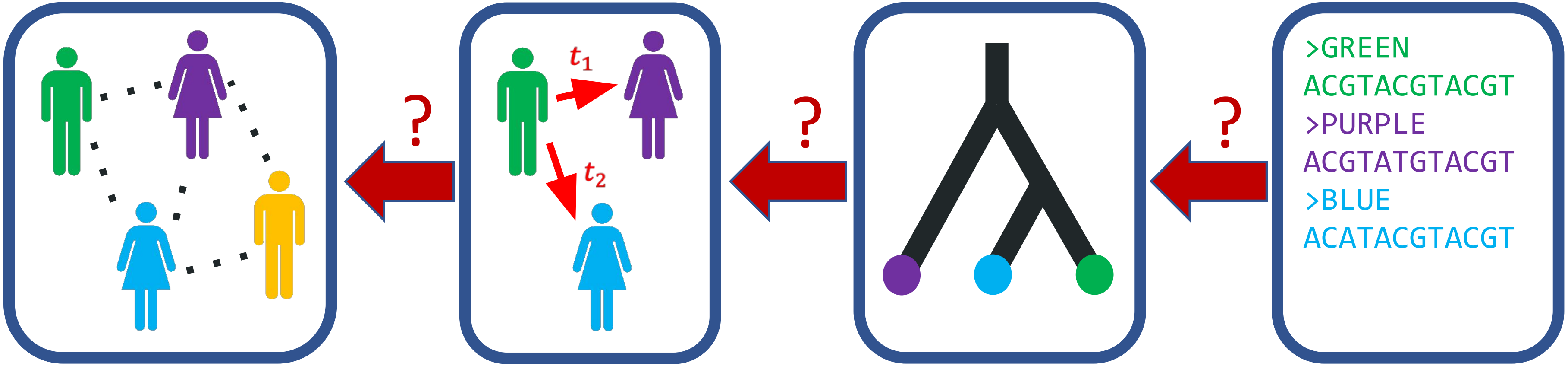
Generative Process



```
>GREEN
ACGTACGTACGT
>PURPLE
ACGTATGTACGT
>BLUE
ACATACGTACGT
```

Viral Sequences

Inference



```
>GREEN
ACGTACGTACGT
>PURPLE
ACGTATGTACGT
>BLUE
ACATACGTACGT
```

Contact Network

Transmission Network

Viral Phylogeny

Viral Sequences

Viral Molecular Epidemiology

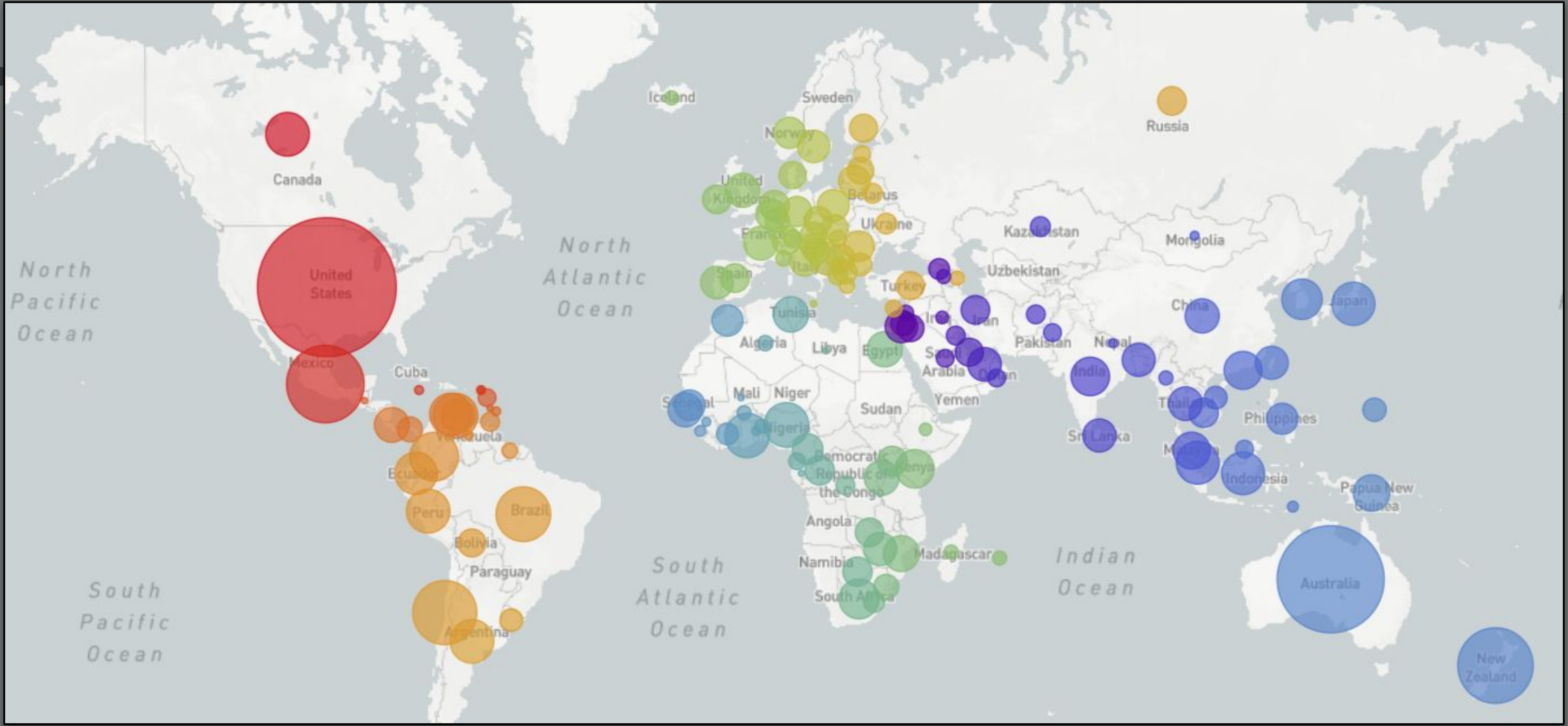
Viral Molecular Epidemiology

- We can use properties of the evolution of viruses to study a viral epidemic

Viral Molecular Epidemiology

- We can use properties of the evolution of viruses to study a viral epidemic
  - How did the virus spread across our communities or across the world?

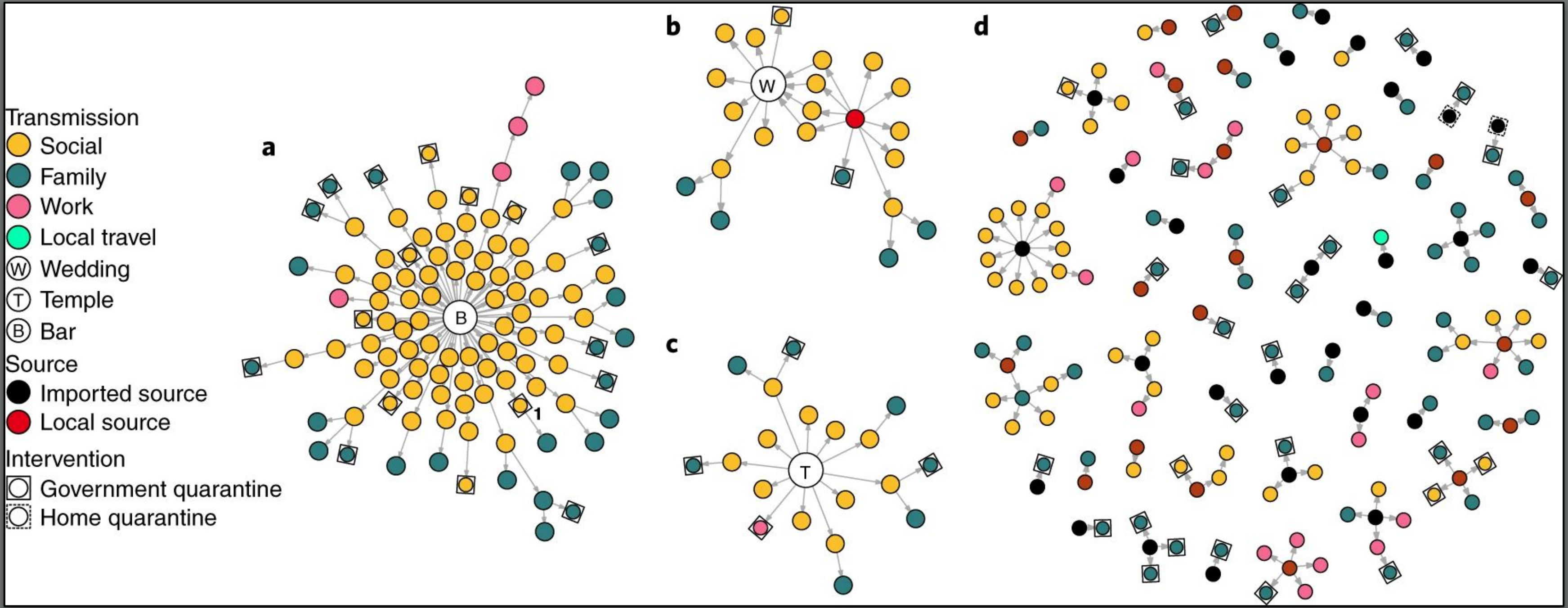
Viral Molecular Epidemiology



Viral Molecular Epidemiology

- We can use properties of the evolution of viruses to study a viral epidemic
  - How did the virus spread across our communities or across the world?
  - What “transmission clusters” exist within our population?

Viral Molecular Epidemiology

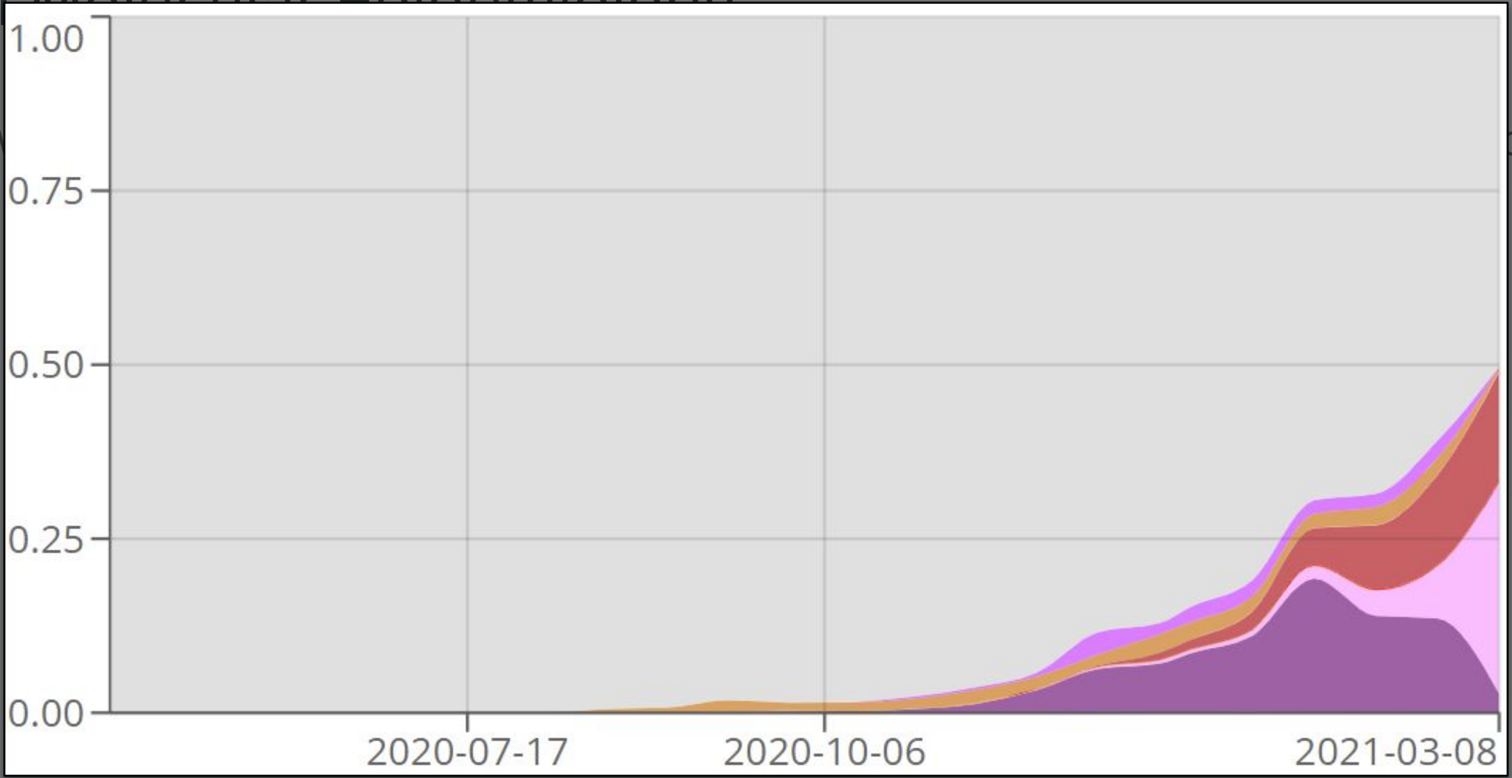


Viral Molecular Epidemiology

- We can use properties of the evolution of viruses to study a viral epidemic
  - How did the virus spread across our communities or across the world?
  - What “transmission clusters” exist within our population?
  - How is the virus mutating across the epidemic?

Viral Molecular Epidemiology

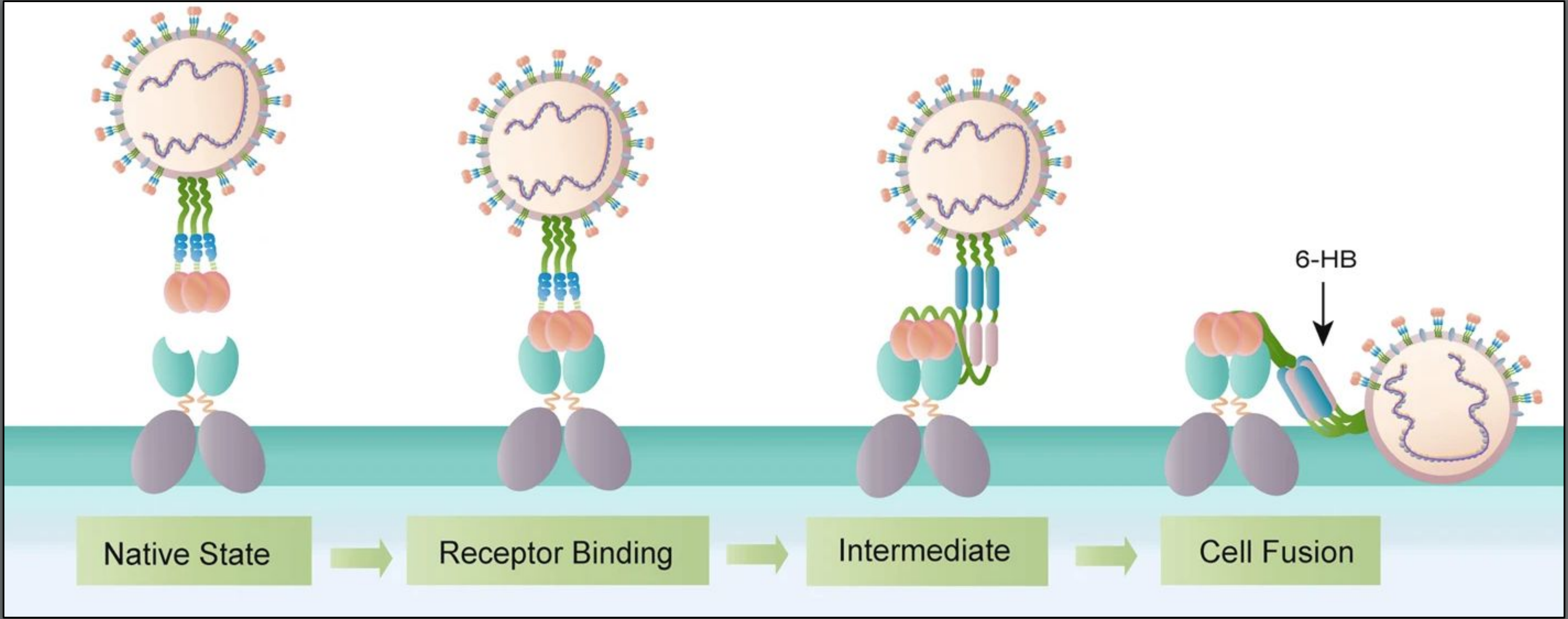
- W c



Viral Molecular Epidemiology

- We can use properties of the evolution of viruses to study a viral epidemic
    - How did the virus spread across our communities or across the world?
    - What “transmission clusters” exist within our population?
    - How is the virus mutating across the epidemic?
    - What is the molecular mechanism by which the virus invades our cells?

Viral Molecular Epidemiology

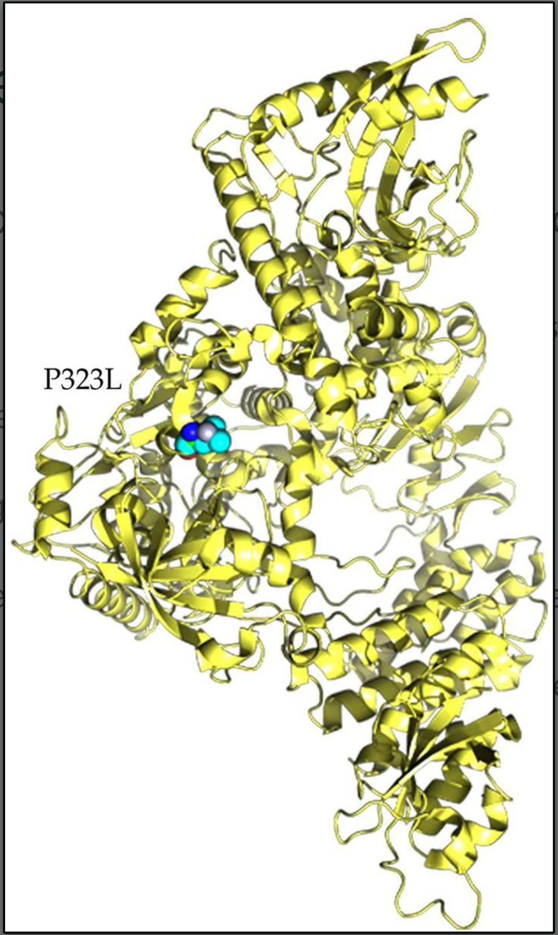


Viral Molecular Epidemiology

- We can use properties of the evolution of viruses to study a viral epidemic
    - How did the virus spread across our communities or across the world?
    - What “transmission clusters” exist within our population?
    - How is the virus mutating across the epidemic?
    - What is the molecular mechanism by which the virus invades our cells?
    - Are any of the mutations impacting the infectiousness of the virus?

Viral Molecular Epide

- We can use properties o o study a viral epidemic
  - ○ How did the virus spread the world?
  - ○ What “transmission cluste
  - ○ How is the virus mutating
  - ○ What is the molecular me es our cells?
  - ○ Are any of the mutations e virus?



Viral Molecular Epidemiology

- We can use properties of the evolution of viruses to study a viral epidemic
  - How did the virus spread across our communities or across the world?
  - What “transmission clusters” exist within our population?
  - How is the virus mutating across the epidemic?
  - What is the molecular mechanism by which the virus invades our cells?
  - Are any of the mutations impacting the infectiousness of the virus?
- How can we translate such questions into formal computational problems?

Outline

- Introduction to Viral Molecular Epidemiology
- **Sequencing the First Viral Genome**
- Annotating a Viral Genome
- Sequencing in the Midst of an Epidemic
- Aligning Viral Genome Sequences
- Phylogenetic Inference and Transmission Clustering

stack of NY Times, June 27, 2000

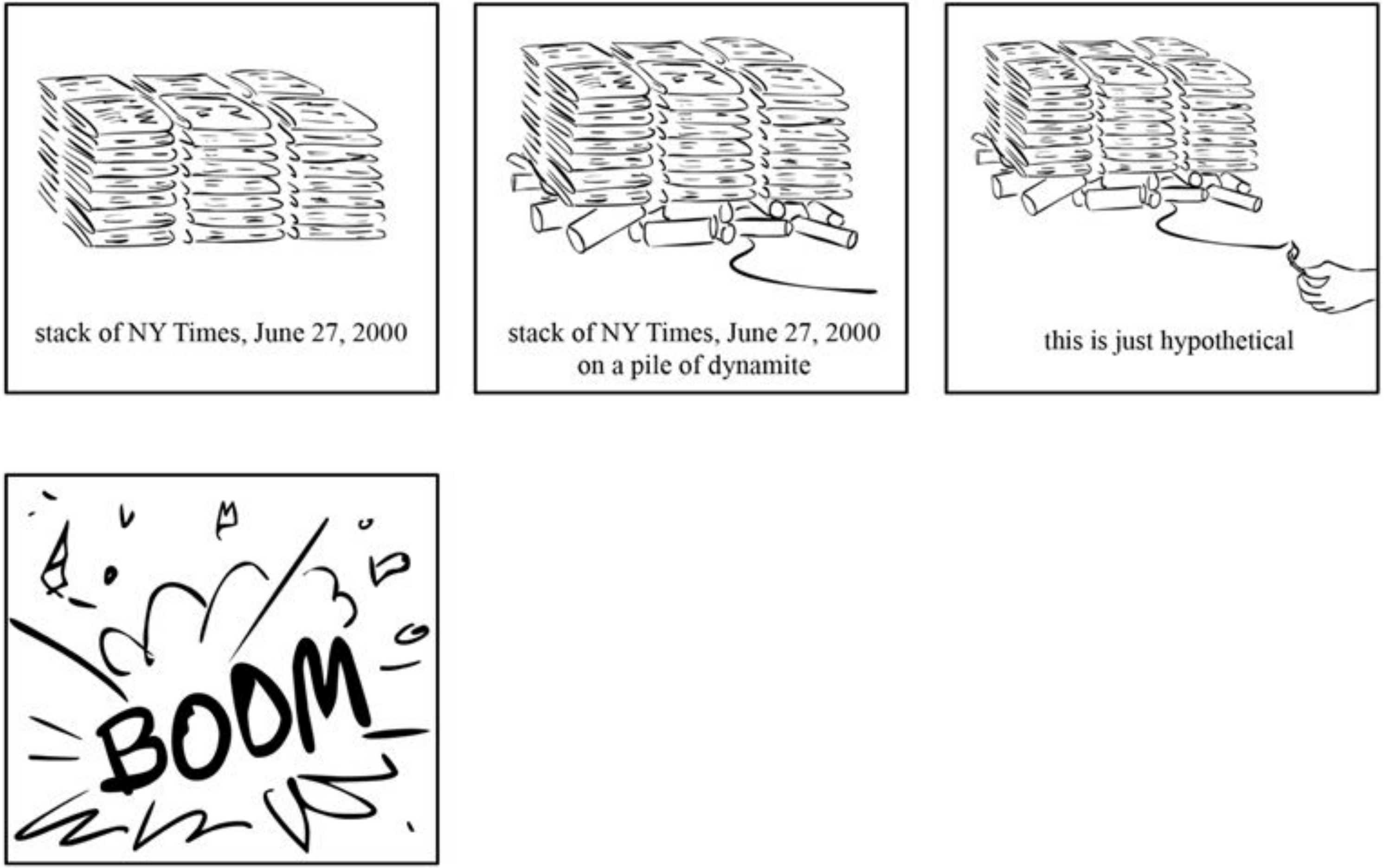
stack of NY Times, June 27, 2000

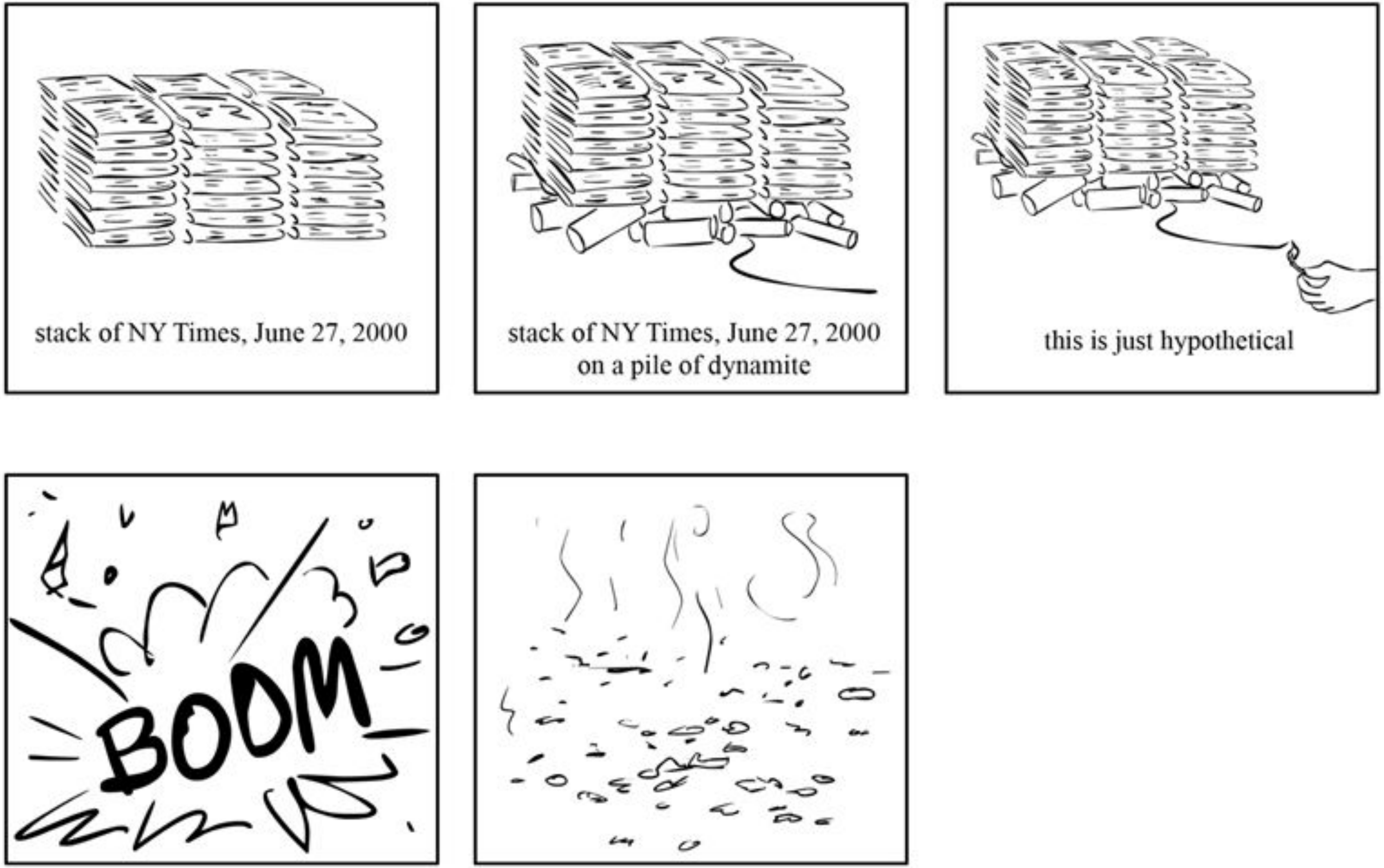
stack of NY Times, June 27, 2000
on a pile of dynamite

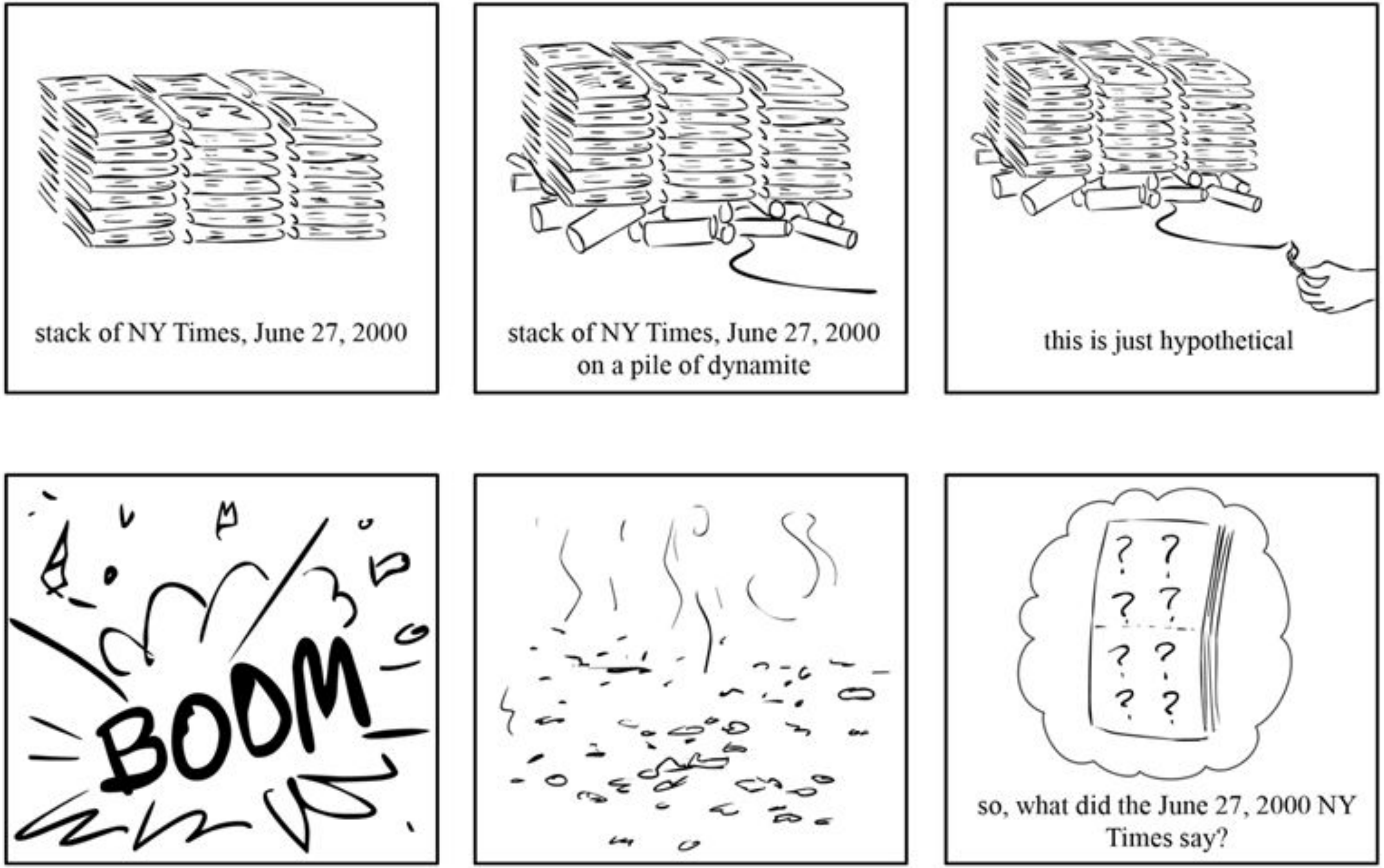
stack of NY Times, June 27, 2000

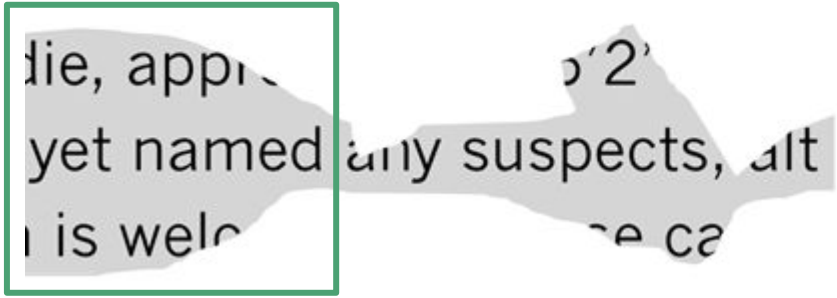
stack of NY Times, June 27, 2000
on a pile of dynamite

this is just hypothetical


stack of NY Times, June 27, 2000
stack of NY Times, June 27, 2000 on a pile of dynamite
this is just hypothetical
BOOM


stack of NY Times, June 27, 2000
stack of NY Times, June 27, 2000 on a pile of dynamite
this is just hypothetical
BOOM

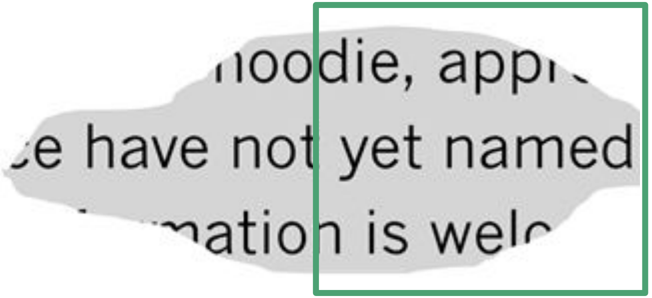

stack of NY Times, June 27, 2000
stack of NY Times, June 27, 2000 on a pile of dynamite
this is just hypothetical
BOOM
so, what did the June 27, 2000 NY Times say?



die, appr ’2’

yet named any suspects, alt

is welc ce ca



hoodie, appr

ce have not yet named

mation is welc

stack of NY Times, June 27, 2000

stack of NY Times, June 27, 2000

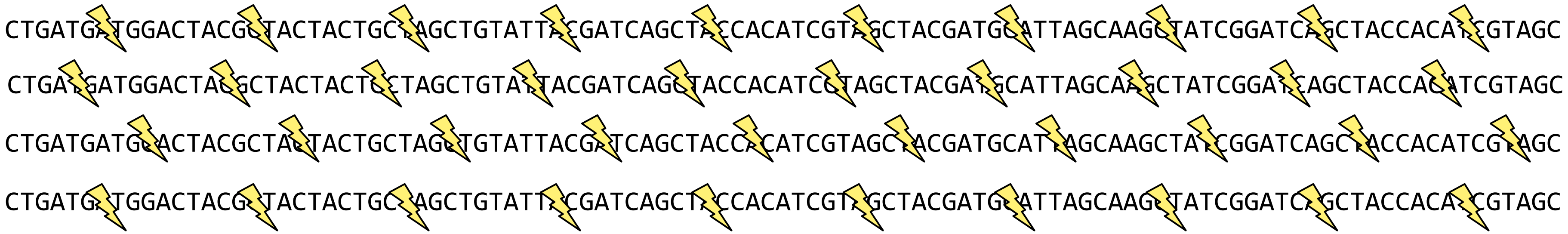
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

stack of NY Times, June 27, 2000



CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

stack of NY Times, June 27, 2000

CTGATG TGGACTACG TACTACTGC AGCTGTATT CGATCAGCT CCACATCGT GCTACGATG ATTAGCAAG TATCGGATC GCTACCACA CGTAGC

CTGA GATGGACTA GCTACTACT CTAGCTGTA TACGATCAG TACCACAT GTAGCTACGA GCATTAGC AGCTATCGGA CAGCTACCA ATCGTAGC

CTGATGATG ACTACGCT CTACTGCTAG TGTATTACG TCAGCTACC CATCGTAGC ACGATGCAT AGCAAGCTA CGGATCAGC ACCACATCG AGC

CTGAT ATGGACTAC CTACTACTG TAGCTGTATT CGATCAGC ACCACATCGT GCTACGATG ATTAGCAA CTATCGGATCA CTACCAC TCGTAGC

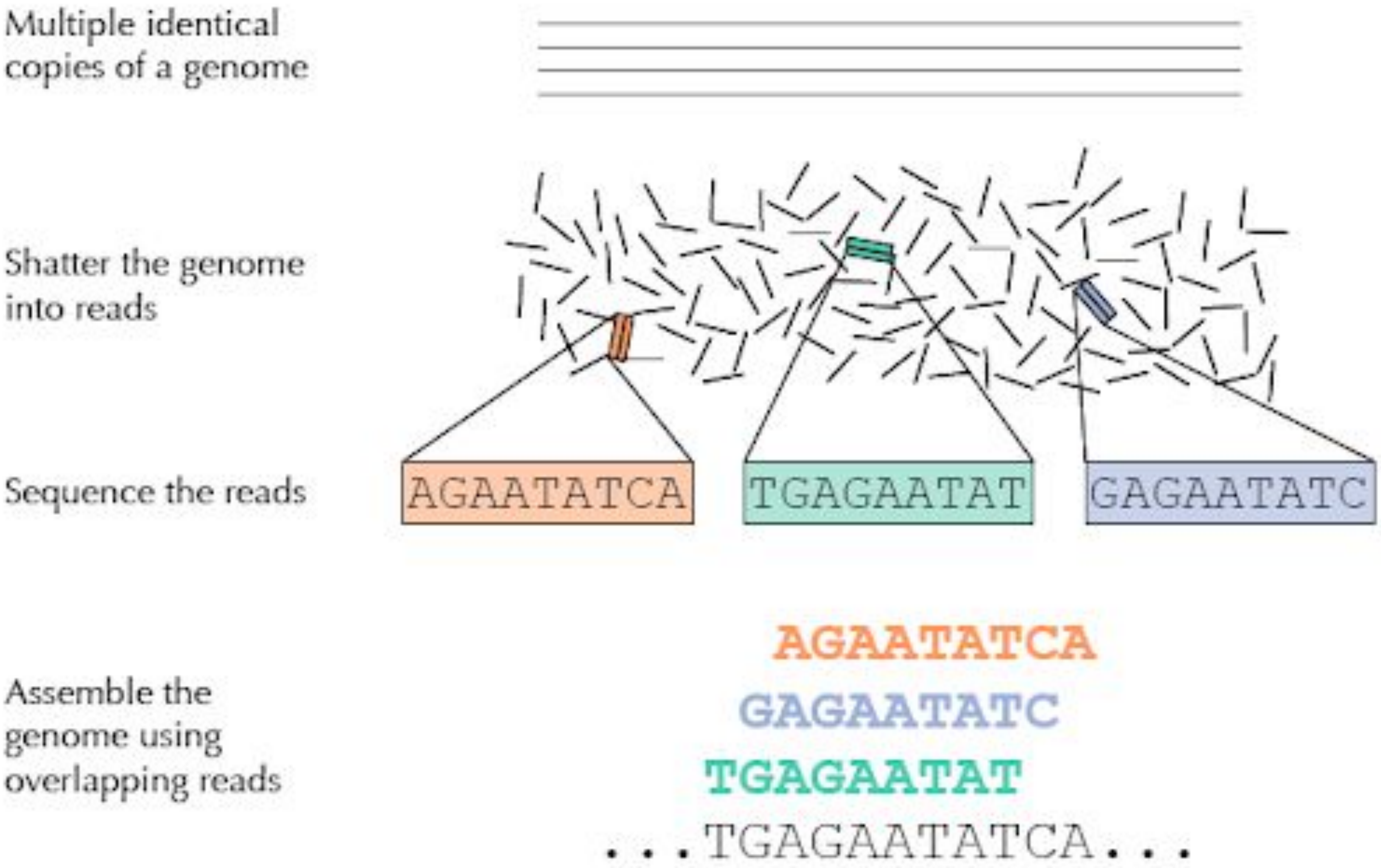
stack of NY Times, June 27, 2000

TACTACTGC CGATCAGCT CCACATCGT GCTACGATG TATCGGATC CGTAGC

CTAGCTGTA TACGATCAG GCATTAGC CAGCTACCA

CTGATGATG ACTACGCT TGTATTACG TCAGCTACC CATCGTAGC AGCAAGCTA ACCACATCG

ATGGACTAC CTACTACTG CGATCAGC GCTACGATG CTATCGGATCA CTACCAC TCGTAGC



Multiple identical copies of a genome

Shatter the genome into reads

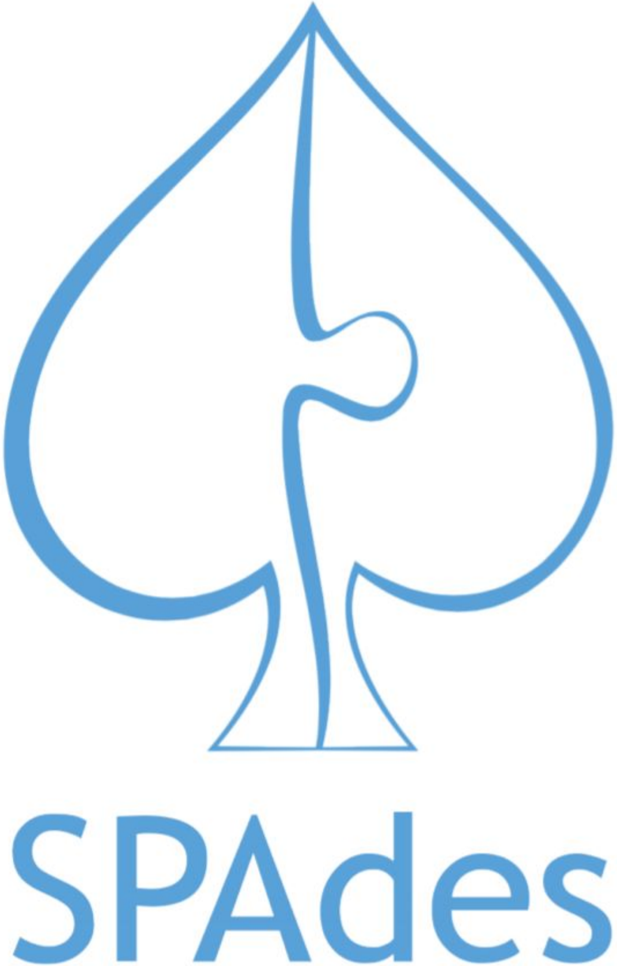
Sequence the reads

AGAATATCA TGAGAATAT GAGAATATC

Assemble the genome using overlapping reads

```
   AGAATATCA
  GAGAATATC
 TGAGAATAT
...TGAGAATATCA...
```

SPAdes Assembler



Center for Algorithmic Biotechnology, St Petersburg University | https://cab.spbu.ru/software/spades/

Outline

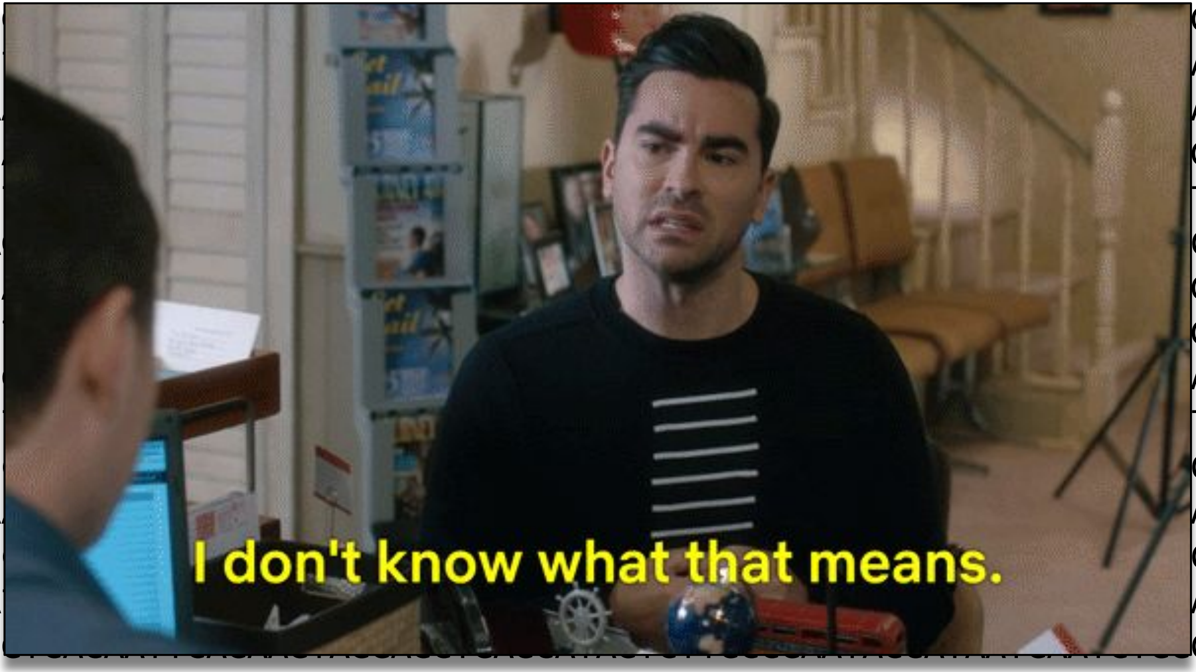
- Introduction to Viral Molecular Epidemiology
- Sequencing the First Viral Genome
- **Annotating a Viral Genome**
- Sequencing in the Midst of an Epidemic
- Aligning Viral Genome Sequences
- Phylogenetic Inference and Transmission Clustering

Assembled Genome

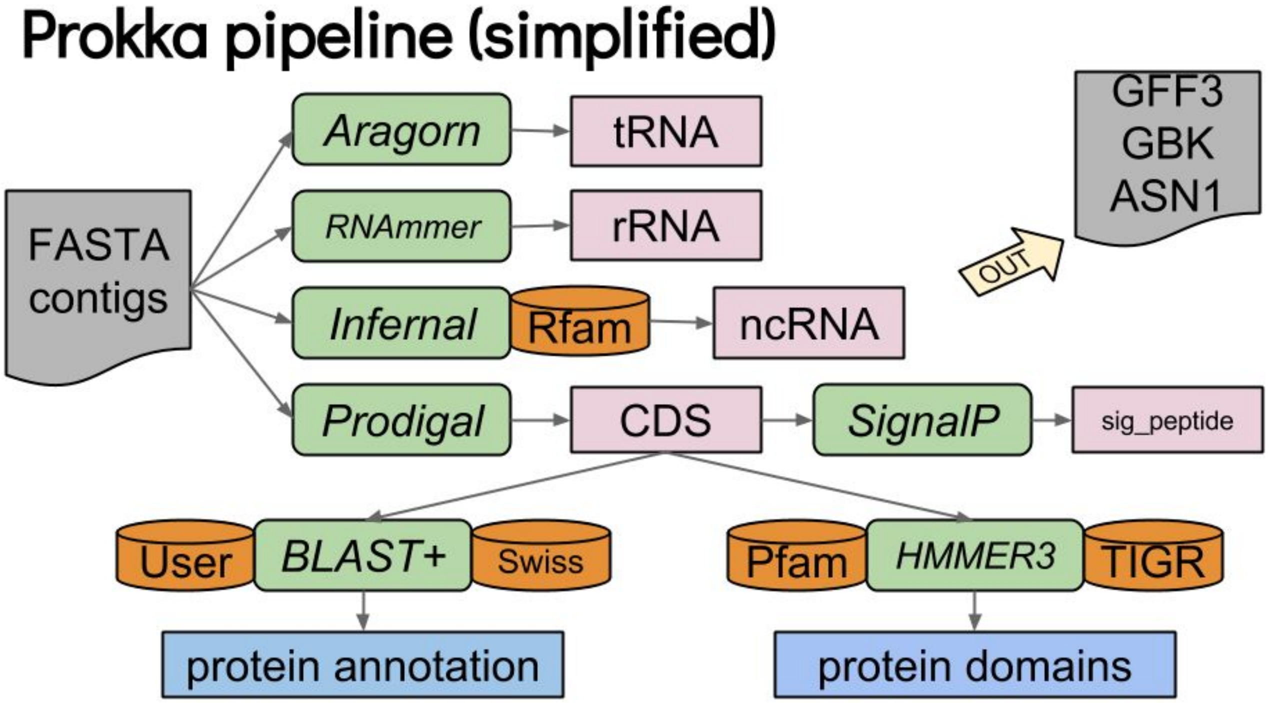
```
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGG
CTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTT
CTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAG
CCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGAC
TCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAAC
TTGAACAGCCCTATGTGTTCATCAAACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGG
CATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGT
AAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGATCCTTATG
AAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTA
TGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTG
AAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCC
CTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTT
GCGTCACCAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCAGACGGGCGATTTTG
TTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCACTACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAA
AATTTATTGTCCAGCATGTCACAATTCAGAAGTAGGACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGGCTTGAAAACCATTCTT
CGTAAGGGTGGTCGCACTATTGCCTTTGGAGGCTGTGTGTTCTCTTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCACGTGCTA
GCGCTAACATAGGTTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACAACCTTCTTGAAATACTCCAAAAAGA...
```

Assembled Genome

```
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGG
CTGTCACTCGGCTGCAT                                                          GTAACTCGTCTATCTT
CTGCAGGCTGCTTACGG                                                          AAGGTAAGATGGAGAG
CCTTGTCCCTGGTTTCA                                                          ACGTGGCTTTGGAGAC
TCCGTGGAGGAGGTCTT                                                          GGCGTTTTGCCTCAAC
TTGAACAGCCCTATGTG                                                          TAGCAGAACTCGAAGG
CATTCAGTACGGTCGTA                                                          CAAGGTTCTTCTTCGT
AAGAACGGTAATAAAGG                                                          GGCACTGATCCTTATG
AAGATTTTCAAGAAAAC                                                          GGGCATACACTCGCTA
TGTCGATAACAACTTCT                                                          AGCTTCATGCACTTTG
TCCGAACAACTGGACTT                                                          TACACGGAACGTTCTG
AAAAGAGCTATGAATTG                                                          CAAATTTTGTATTTCC
CTTAAATTCCATAATCA                                                          ATCTGTCTATCCAGTT
GCGTCACCAAATGAATG                                                          CAGACGGGCGATTTTG
TTAAAGCCACTTGCGAA                                                          AAAATGCTGTTGTTAA
AATTTATTGTCCAGCAT                                                          CTTGAAAACCATTCTT
CGTAAGGGTGGTCGCACTATTGCCTTTGGAGGCTGTGTGTTCTCTTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCACGTGCTA
GCGCTAACATAGGTTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACAACCTTCTTGAAATACTCCAAAAGA...
```



Prokka: Gene Prediction and Functional Annotation



Outline

- Introduction to Viral Molecular Epidemiology
- Sequencing the First Viral Genome
- Annotating a Viral Genome
- **Sequencing in the Midst of an Epidemic**
- Aligning Viral Genome Sequences
- Phylogenetic Inference and Transmission Clustering

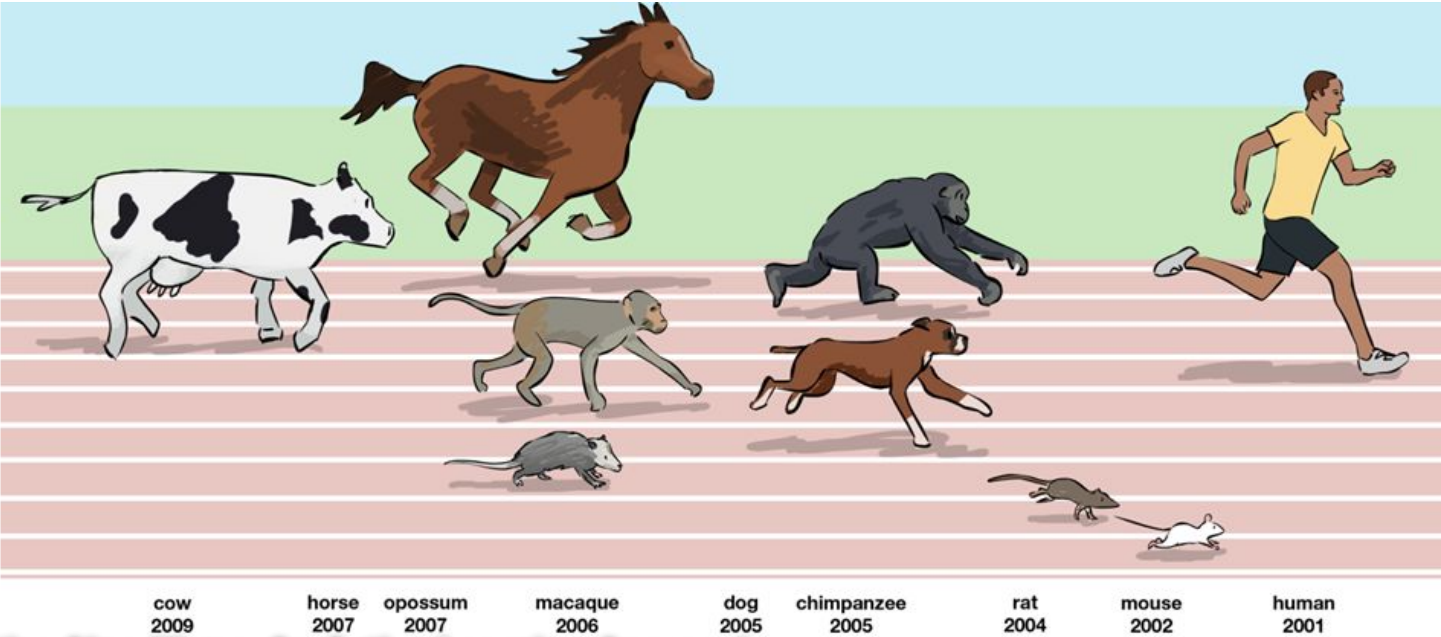
Reference Genomes

Reference Genomes

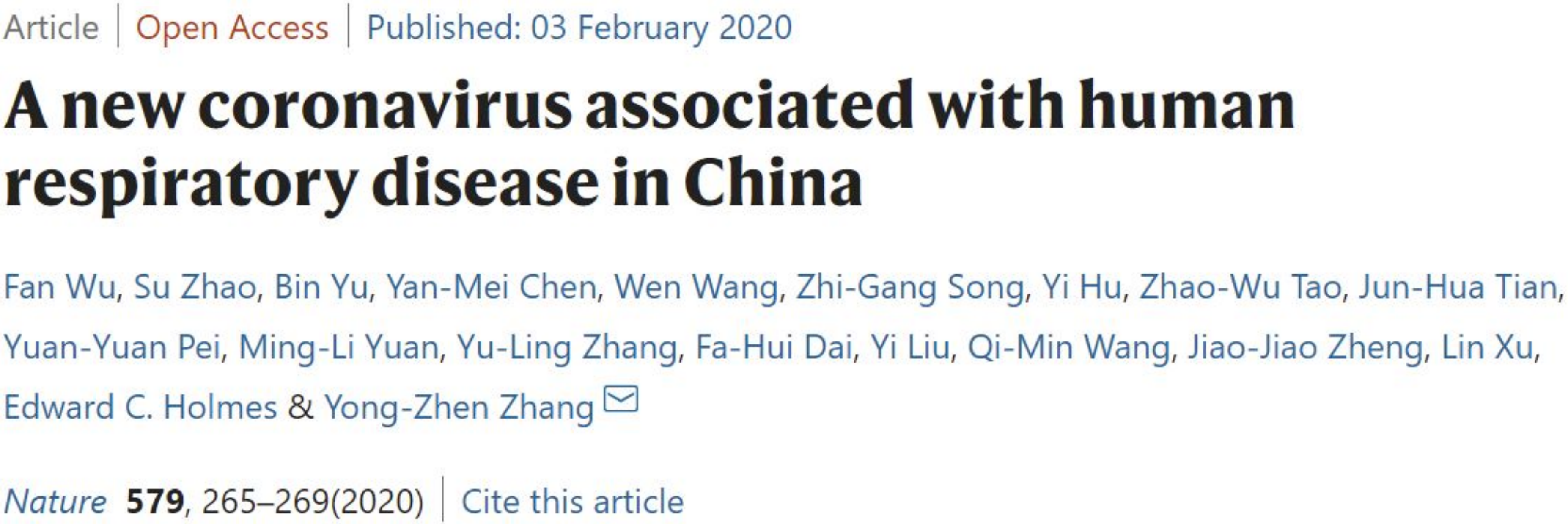
- **Reference Genome:** A high-confidence assembled genome sequence that is constructed as a representative example of an individual organism

Reference Genomes

- **Reference Genome:** A high-confidence assembled genome sequence that is constructed as a representative example of an individual organism



The SARS-CoV-2 Reference Genome



Article | Open Access | Published: 03 February 2020

A new coronavirus associated with human respiratory disease in China

Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C. Holmes & Yong-Zhen Zhang ✉

Nature **579**, 265–269(2020) | Cite this article

Mapping Reads to the Reference

Mapping Reads to the Reference

- In *de novo* assembly, we inferred relationships between reads from overlaps

Mapping Reads to the Reference

- In *de novo* assembly, we inferred relationships between reads from overlaps
- In other words, we compare reads against *each other*

Mapping Reads to the Reference

- In *de novo* assembly, we inferred relationships between reads from overlaps
- In other words, we compare reads against *each other*

Mapping Reads to the Reference

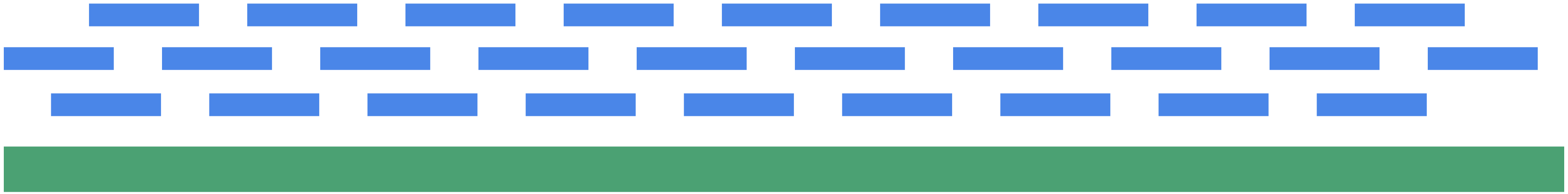
- In *de novo* assembly, we inferred relationships between reads from overlaps
- In other words, we compare reads against *each other*

Mapping Reads to the Reference

- In *de novo* assembly, we inferred relationships between reads from overlaps
- In other words, we compare reads against *each other*

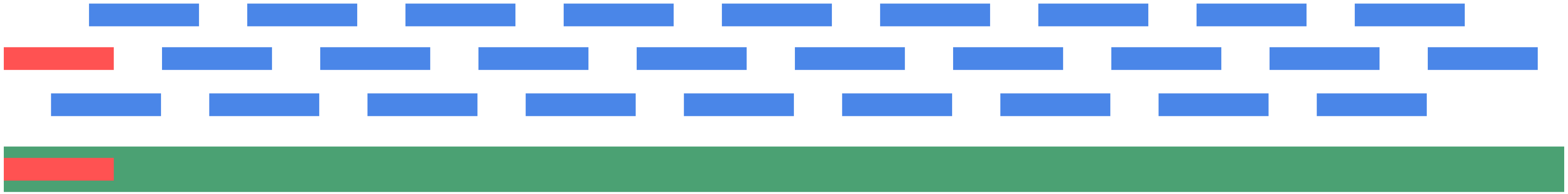
Mapping Reads to the Reference

- In *de novo* assembly, we inferred relationships between reads from overlaps
- In other words, we compare reads against *each other*
- What if we instead compare reads against the ***reference genome***?



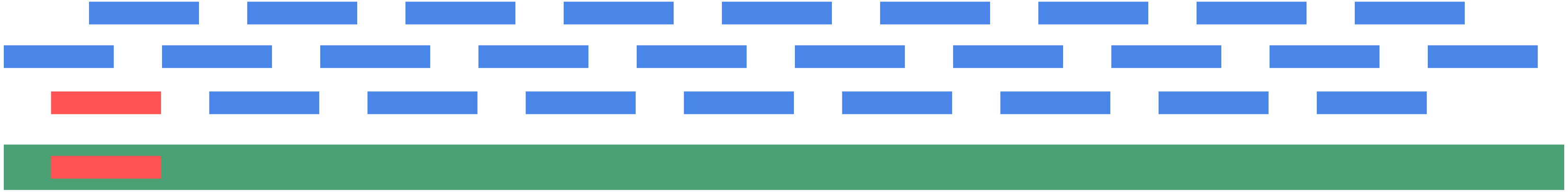
Mapping Reads to the Reference

- In *de novo* assembly, we inferred relationships between reads from overlaps
- In other words, we compare reads against *each other*
- What if we instead compare reads against the ***reference genome***?



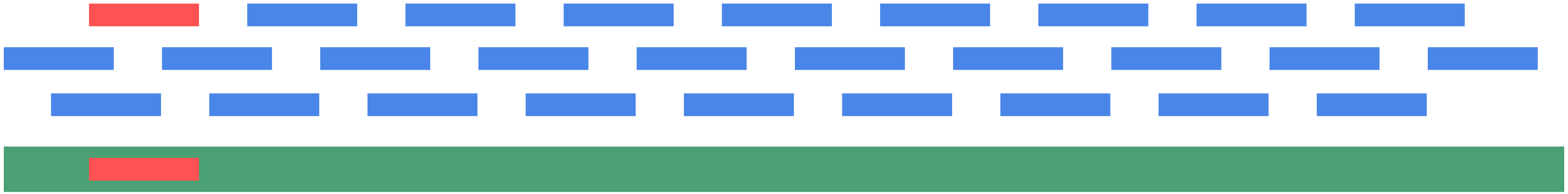
Mapping Reads to the Reference

- In *de novo* assembly, we inferred relationships between reads from overlaps
- In other words, we compare reads against *each other*
- What if we instead compare reads against the ***reference genome***?



Mapping Reads to the Reference

- In *de novo* assembly, we inferred relationships between reads from overlaps
- In other words, we compare reads against *each other*
- What if we instead compare reads against the ***reference genome***?



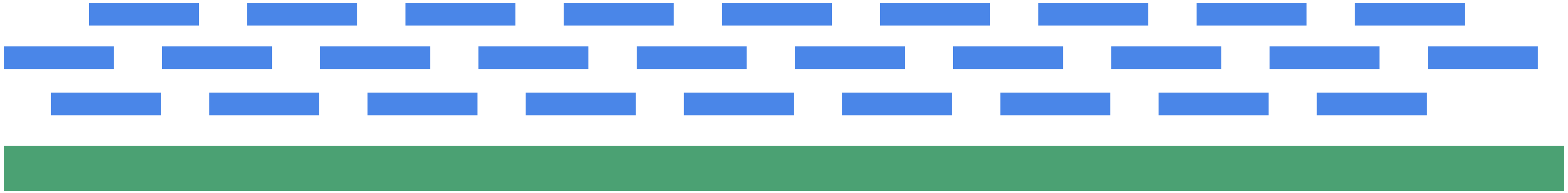
Read Mappers

- **Minimap2:** https://github.com/lh3/minimap2
- **Unimap:** https://github.com/lh3/unimap
- **BWA:** https://github.com/lh3/bwa
- **Bowtie 2:** https://github.com/BenLangmead/bowtie2
- Many more

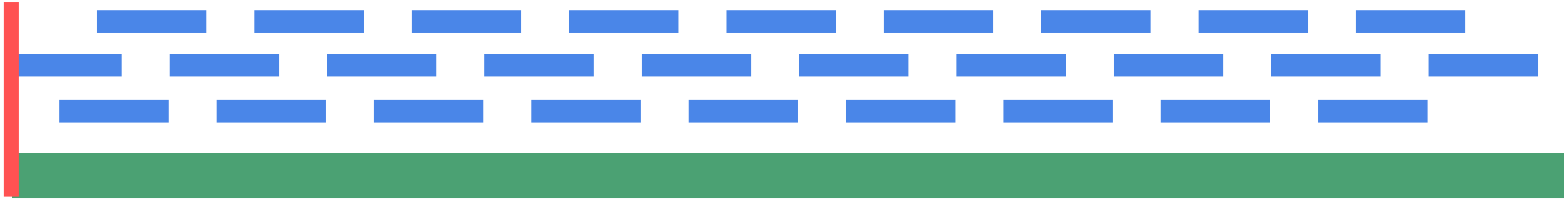
Read Mappers

- **Minimap2:** https://github.com/lh3/minimap2
- **Unimap:** https://github.com/lh3/unimap
- **BWA:** https://github.com/lh3/bwa
- **Bowtie 2:** https://github.com/BenLangmead/bowtie2
- Many more

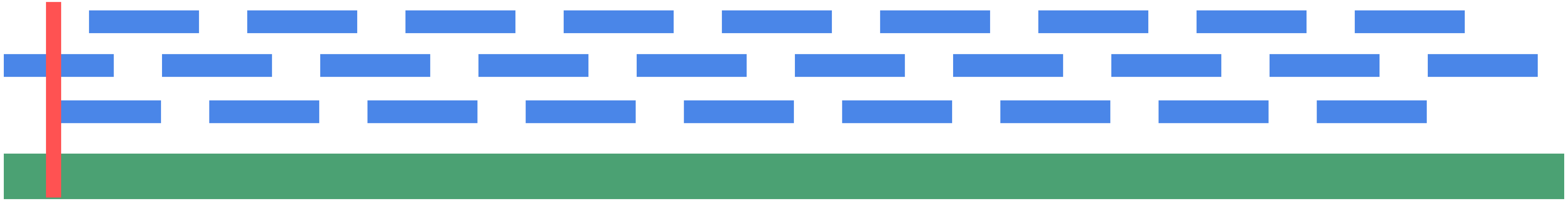
Consensus Sequence



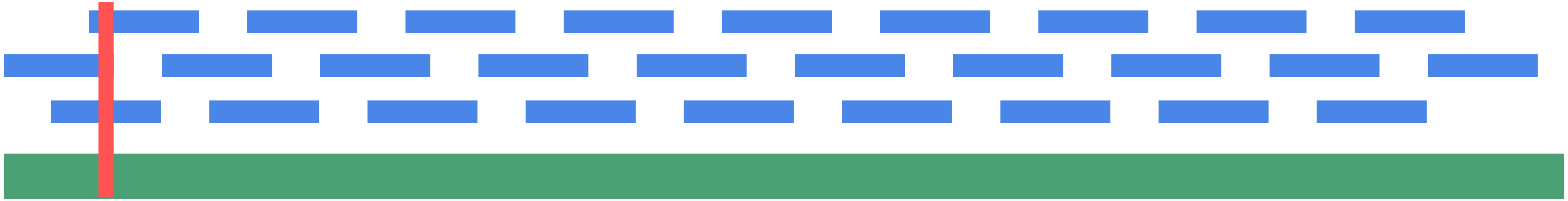
Consensus Sequence



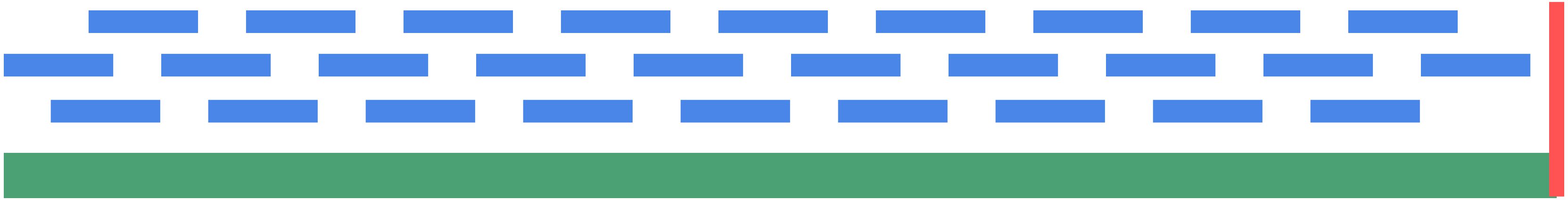
Consensus Sequence



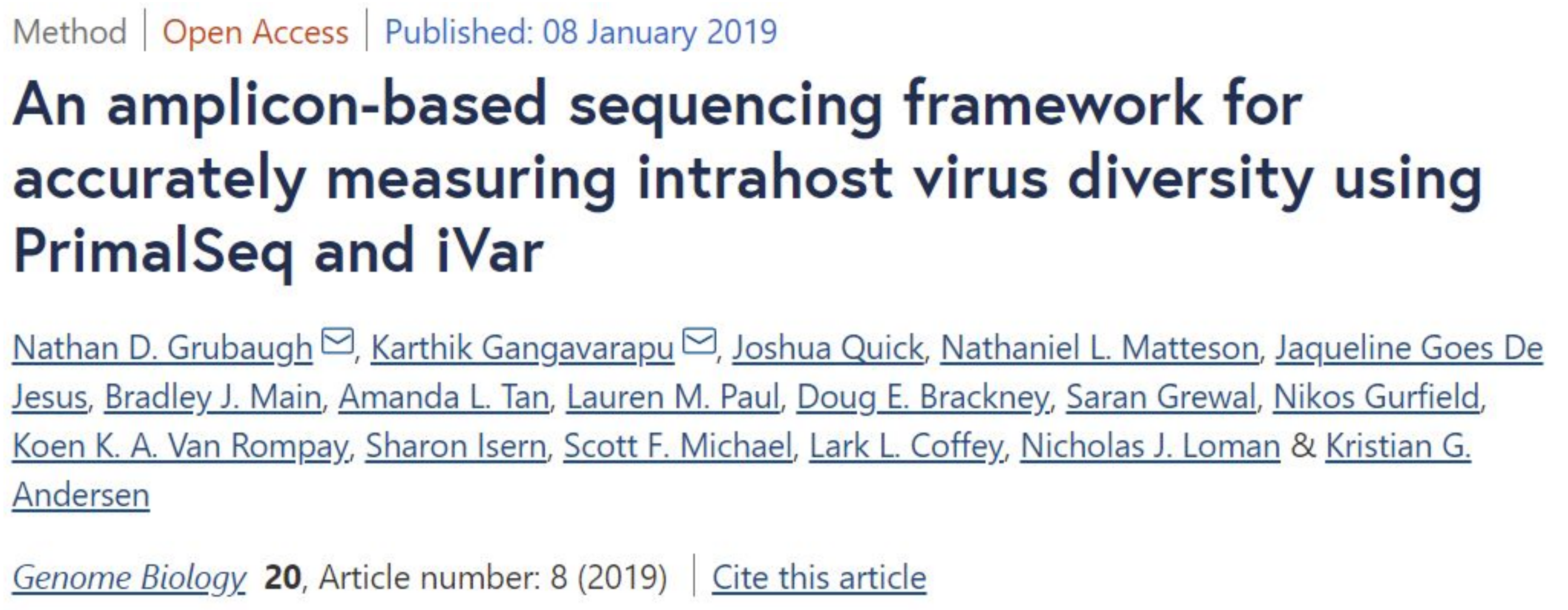
Consensus Sequence



Consensus Sequence



Consensus Sequence



Method | Open Access | Published: 08 January 2019

An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar

Nathan D. Grubaugh, Karthik Gangavarapu, Joshua Quick, Nathaniel L. Matteson, Jaqueline Goes De Jesus, Bradley J. Main, Amanda L. Tan, Lauren M. Paul, Doug E. Brackney, Saran Grewal, Nikos Gurfield, Koen K. A. Van Rompay, Sharon Isern, Scott F. Michael, Lark L. Coffey, Nicholas J. Loman & Kristian G. Andersen

Genome Biology **20**, Article number: 8 (2019) | Cite this article

Outline

- Introduction to Viral Molecular Epidemiology
- Sequencing the First Viral Genome
- Annotating a Viral Genome
- Sequencing in the Midst of an Epidemic
- **Aligning Viral Genome Sequences**
- Phylogenetic Inference and Transmission Clustering

Multiple Sequence Alignment

Multiple Sequence Alignment

- We want to align *n* genome sequences, each with length *k*

Multiple Sequence Alignment

- We want to align *n* genome sequences, each with length *k*
  - **COVID-19:** $k = 29{,}000$ and $n > 3$ million

Multiple Sequence Alignment

- We want to align *n* genome sequences, each with length *k*
  - **COVID-19:** *k* = 29,000 and *n* > 3 million
- Finding an alignment guaranteed to be optimal has time complexity **O(k^n)**

Multiple Sequence Alignment

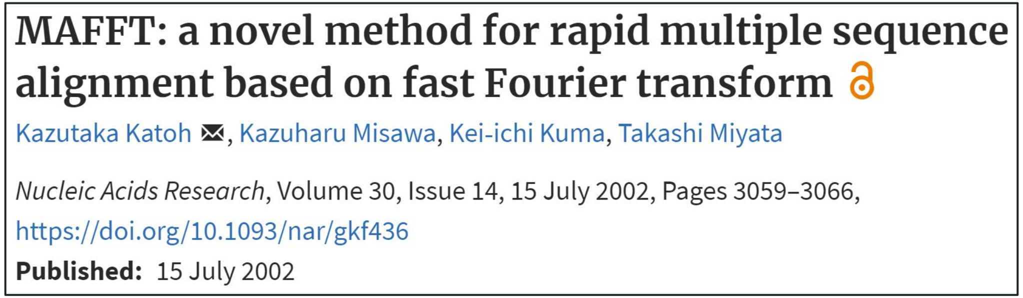
- We want to align *n* genome sequences, each with length *k*
  - **COVID-19:** $k = 29{,}000$ and $n > 3$ million
- Finding an alignment guaranteed to be optimal has time complexity **O(*k*^{n})**
  - Assuming each operation takes 1 ns, that's **longer than the existence of the universe**

Multiple Sequence Alignment

- We want to align *n* genome sequences, each with length *k*
    - **COVID-19:** $k = 29{,}000$ and $n > 3$ million
- Finding an alignment guaranteed to be optimal has time complexity **O(*k*^{n})**
    - Assuming each operation takes 1 ns, that's **longer than the existence of the universe**
- Heuristics exist that give *approximate* alignments much faster

Multiple Sequence Alignment

- We want to align *n* genome sequences, each with length *k*
  - **COVID-19:** k = 29,000 and n > 3 million
- Finding an alignment guaranteed to be optimal has time complexity $\mathbf{O(k^n)}$
  - Assuming each operation takes 1 ns, that's **longer than the existence of the universe**
- Heuristics exist that give *approximate* alignments much faster
  - Not guaranteed to be optimal, but have pretty good accuracy

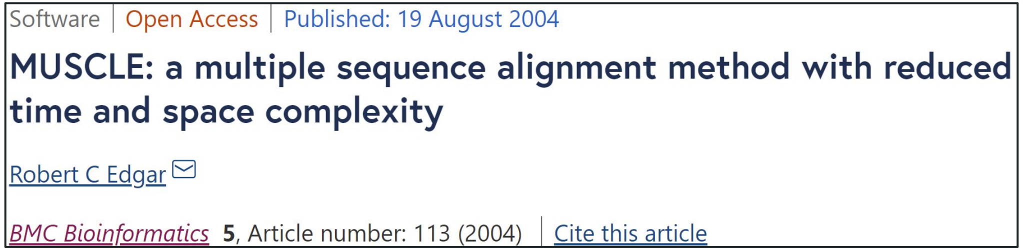


MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform

Kazutaka Katoh ✉, Kazuharu Misawa, Kei-ichi Kuma, Takashi Miyata

Nucleic Acids Research, Volume 30, Issue 14, 15 July 2002, Pages 3059–3066, https://doi.org/10.1093/nar/gkf436

Published: 15 July 2002

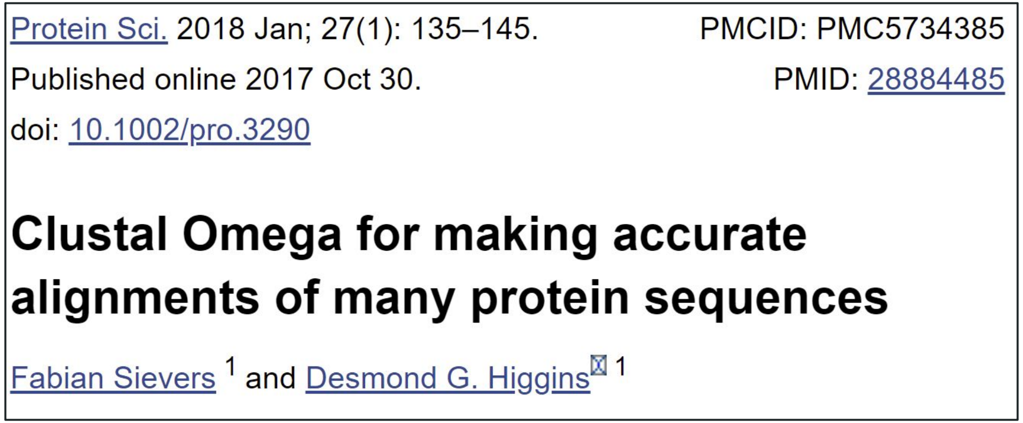


Software | Open Access | Published: 19 August 2004

MUSCLE: a multiple sequence alignment method with reduced time and space complexity

Robert C Edgar ✉

BMC Bioinformatics **5**, Article number: 113 (2004) | Cite this article



Protein Sci. 2018 Jan; 27(1): 135–145. PMCID: PMC5734385

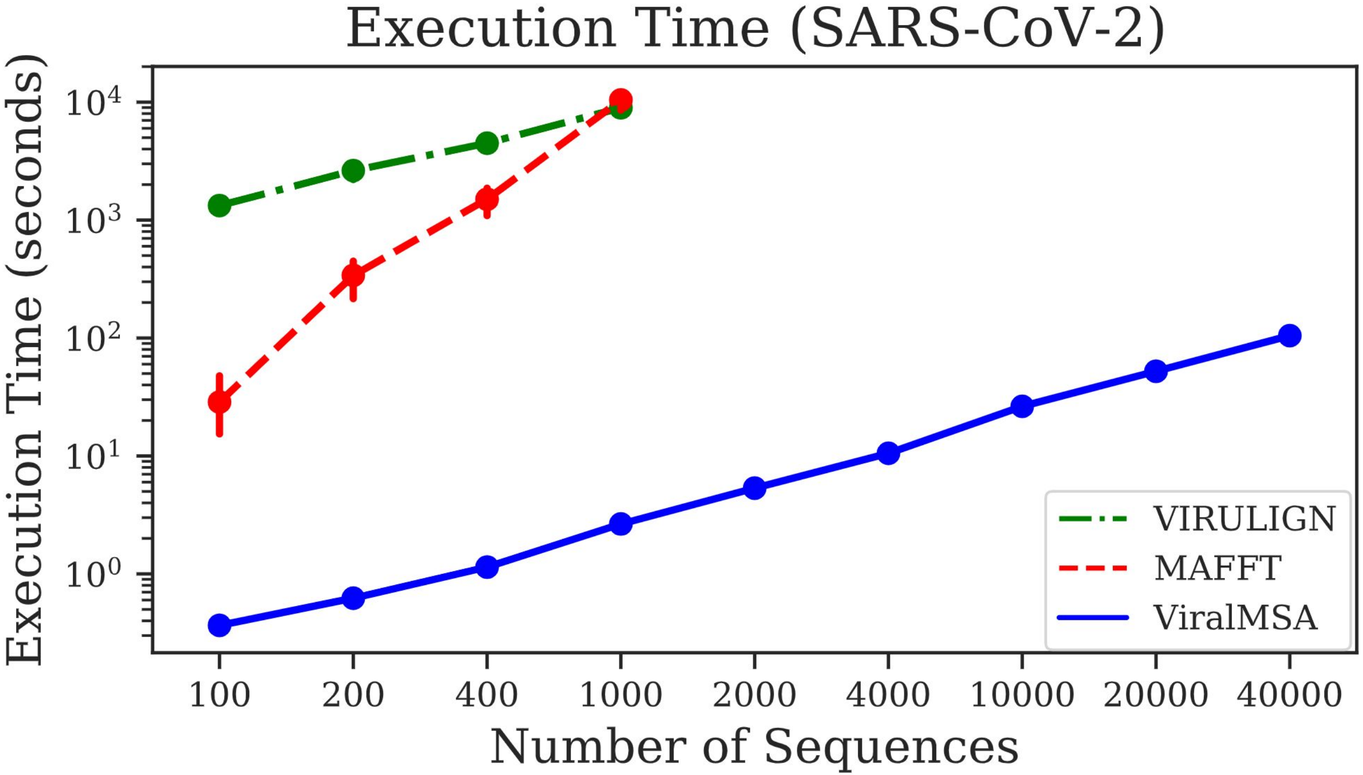
Published online 2017 Oct 30. PMID: 28884485

doi: 10.1002/pro.3290

Clustal Omega for making accurate alignments of many protein sequences

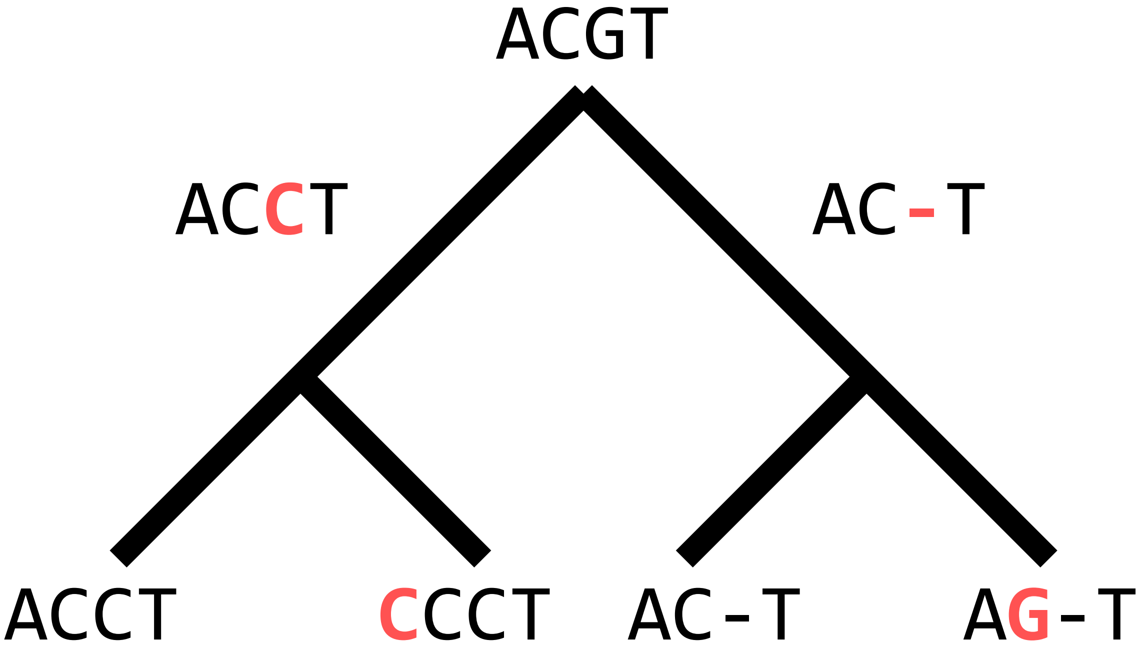
Fabian Sievers [1] and Desmond G. Higgins✉ [1]

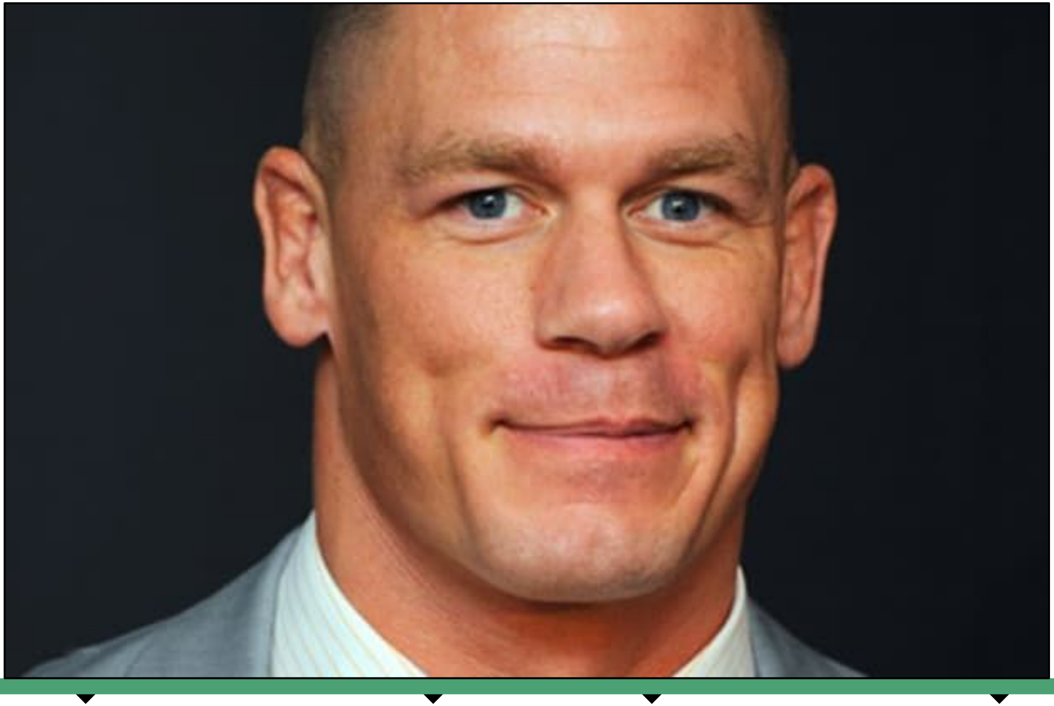
Align-to-Reference Approach



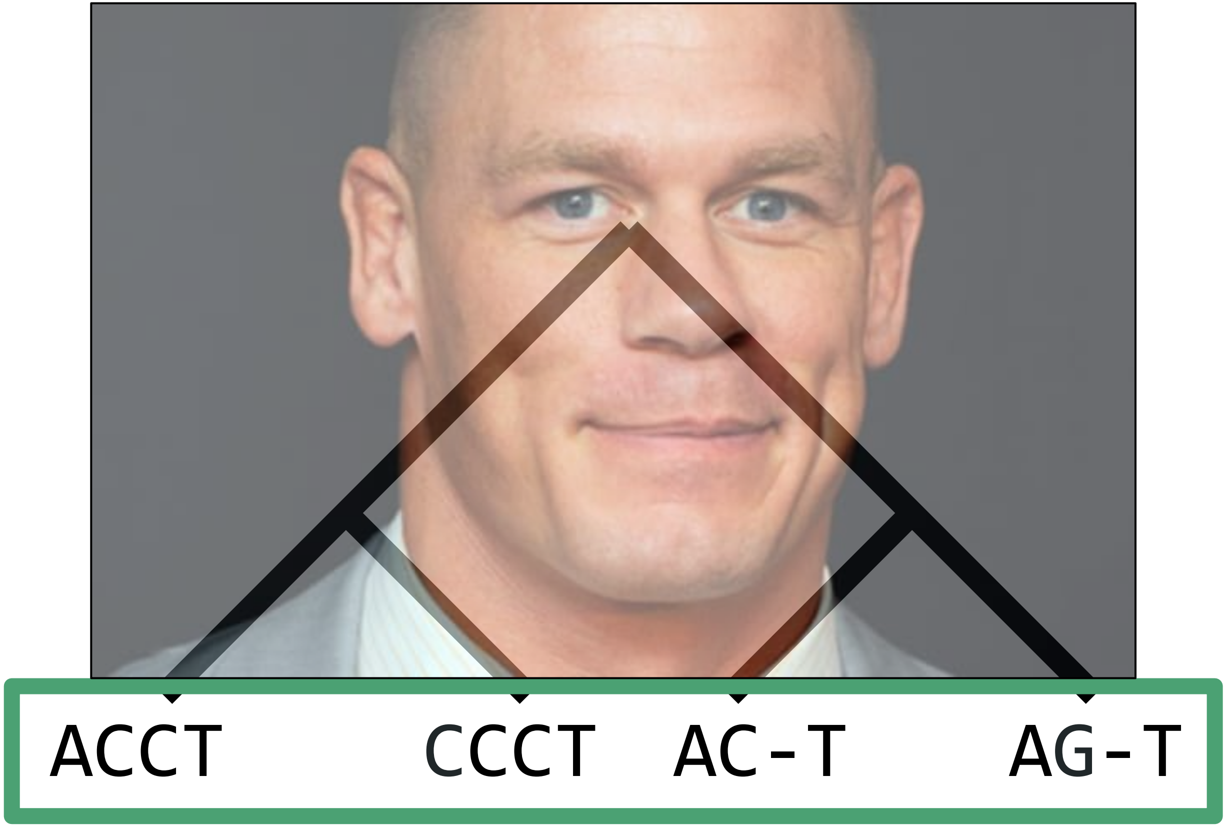
Outline

- Introduction to Viral Molecular Epidemiology
- Sequencing the First Viral Genome
- Annotating a Viral Genome
- Sequencing in the Midst of an Epidemic
- Aligning Viral Genome Sequences
- **Phylogenetic Inference and Transmission Clustering**


ACGT
ACCT
AC-T
ACCT
CCCT
AC-T
AG-T



ACCT CCCT AC-T AG-T



ACCT CCCT AC-T AG-T

Current State-of-the-Art Phylogenetic Inference Tools

Current State-of-the-Art Phylogenetic Inference Tools

- The most popular at the moment is **IQ-TREE 2**

Current State-of-the-Art Phylogenetic Inference Tools

- The most popular at the moment is **IQ-TREE 2**
  - Typically the best trade-off between accuracy and speed

Current State-of-the-Art Phylogenetic Inference Tools

- The most popular at the moment is **IQ-TREE 2**
  - Typically the best trade-off between accuracy and speed
  - http://www.iqtree.org/ and https://github.com/Cibiv/IQ-TREE

Current State-of-the-Art Phylogenetic Inference Tools

- The most popular at the moment is **IQ-TREE 2**
  - Typically the best trade-off between accuracy and speed
  - http://www.iqtree.org/ and https://github.com/Cibiv/IQ-TREE
- The fastest tool, but generally lower-accuracy, is **FastTree 2**

Current State-of-the-Art Phylogenetic Inference Tools

- The most popular at the moment is **IQ-TREE 2**
  - Typically the best trade-off between accuracy and speed
  - http://www.iqtree.org/ and https://github.com/Cibiv/IQ-TREE
- The fastest tool, but generally lower-accuracy, is **FastTree 2**
  - http://www.microbesonline.org/fasttree/

Current State-of-the-Art Phylogenetic Inference Tools

- The most popular at the moment is **IQ-TREE 2**
  - Typically the best trade-off between accuracy and speed
  - http://www.iqtree.org/ and https://github.com/Cibiv/IQ-TREE
- The fastest tool, but generally lower-accuracy, is **FastTree 2**
  - http://www.microbesonline.org/fasttree/
- The slowest tool, but generally highest-accuracy, is **RAxML-NG**

Current State-of-the-Art Phylogenetic Inference Tools

- The most popular at the moment is **IQ-TREE 2**
  - Typically the best trade-off between accuracy and speed
  - http://www.iqtree.org/ and https://github.com/Cibiv/IQ-TREE
- The fastest tool, but generally lower-accuracy, is **FastTree 2**
  - http://www.microbesonline.org/fasttree/
- The slowest tool, but generally highest-accuracy, is **RAxML-NG**
  - https://cme.h-its.org/exelixis/software.html and https://github.com/amkozlov/raxml-ng

Dating a Phylogeny

Dating a Phylogeny

- When we infer a phylogeny, branch lengths are in unit of **number of mutations**

Dating a Phylogeny

- When we infer a phylogeny, branch lengths are in unit of **number of mutations**
  - “How many **mutations** occurred from parent to child?”

Dating a Phylogeny

- When we infer a phylogeny, branch lengths are in unit of **number of mutations**
    - “How many **mutations** occurred from parent to child?”
- It would be more useful if branch lengths were in unit of **time**

Dating a Phylogeny

- When we infer a phylogeny, branch lengths are in unit of **number of mutations**
    - “How many **mutations** occurred from parent to child?”
- It would be more useful if branch lengths were in unit of **time**
    - “How much **time** has passed from parent to child?

Dating a Phylogeny

- When we infer a phylogeny, branch lengths are in unit of **number of mutations**
    - “How many **mutations** occurred from parent to child?”
- It would be more useful if branch lengths were in unit of **time**
    - “How much **time** has passed from parent to child?
- We can infer a mutation phylogeny using just sequences

Dating a Phylogeny

- When we infer a phylogeny, branch lengths are in unit of **number of mutations**
    - “How many **mutations** occurred from parent to child?”
- It would be more useful if branch lengths were in unit of **time**
    - “How much **time** has passed from parent to child?
- We can infer a mutation phylogeny using just sequences
- With time information (e.g. sample collection dates), we can **scale** branches

Dating a Phylogeny

- When we infer a phylogeny, branch lengths are in unit of **number of mutations**
    - “How many **mutations** occurred from parent to child?”
- It would be more useful if branch lengths were in unit of **time**
    - “How much **time** has passed from parent to child?
- We can infer a mutation phylogeny using just sequences
- With time information (e.g. sample collection dates), we can **scale** branches
    - We can estimate **mutation rates** (# mutations per time) and scale the branches accordingly

Tools for Dating a Phylogeny

Tools for Dating a Phylogeny

- One of the more popular tools is **TreeTime**

Tools for Dating a Phylogeny

- One of the more popular tools is **TreeTime**
    - https://github.com/neherlab/treetime

Tools for Dating a Phylogeny

- One of the more popular tools is **TreeTime**
  - https://github.com/neherlab/treetime
  - Bonus: It can also do Ancestral State Reconstruction!

Tools for Dating a Phylogeny

- One of the more popular tools is **TreeTime**
  - https://github.com/neherlab/treetime
  - Bonus: It can also do Ancestral State Reconstruction!
- Another popular tool is **LSD2**

Tools for Dating a Phylogeny

- One of the more popular tools is **TreeTime**
  - https://github.com/neherlab/treetime
  - Bonus: It can also do Ancestral State Reconstruction!
- Another popular tool is **LSD2**
  - https://github.com/tothuhien/lsd2

Tools for Dating a Phylogeny

- One of the more popular tools is **TreeTime**
  - https://github.com/neherlab/treetime
  - Bonus: It can also do Ancestral State Reconstruction!
- Another popular tool is **LSD2**
  - https://github.com/tothuhien/lsd2
- Less popular but of interest is **LogDate**

Tools for Dating a Phylogeny

- One of the more popular tools is **TreeTime**
  - https://github.com/neherlab/treetime
  - Bonus: It can also do Ancestral State Reconstruction!
- Another popular tool is **LSD2**
  - https://github.com/tothuhien/lsd2
- Less popular but of interest is **LogDate**
  - https://github.com/uym2/LogDate

Assigning Genomes to Lineages

Assigning Genomes to Lineages

- A **lineage** is simply a subtree of the overall phylogeny

Assigning Genomes to Lineages

- A **lineage** is simply a subtree of the overall phylogeny
    - Biologically, it’s all viral sequences that inherited all mutations up to that point

Assigning Genomes to Lineages

- A **lineage** is simply a subtree of the overall phylogeny
    - Biologically, it’s all viral sequences that inherited all mutations up to that point
- A **strain** is a significant deviation from the ancestral virus

Assigning Genomes to Lineages

- A **lineage** is simply a subtree of the overall phylogeny
  - Biologically, it’s all viral sequences that inherited all mutations up to that point
- A **strain** is a significant deviation from the ancestral virus
  - I couldn’t find a firm/specific definition for what exactly is considered “significant”

Assigning Genomes to Lineages

- A **lineage** is simply a subtree of the overall phylogeny
  - Biologically, it’s all viral sequences that inherited all mutations up to that point
- A **strain** is a significant deviation from the ancestral virus
  - I couldn’t find a firm/specific definition for what exactly is considered “significant”
  - It seems to generally have implications of significant functional/phenotypic differences

Assigning Genomes to Lineages

- A **lineage** is simply a subtree of the overall phylogeny
  - Biologically, it's all viral sequences that inherited all mutations up to that point
- A **strain** is a significant deviation from the ancestral virus
  - I couldn't find a firm/specific definition for what exactly is considered "significant"
  - It seems to generally have implications of significant functional/phenotypic differences
- For SARS-CoV-2, people use **pangolin** to assign genomes to lineages

Assigning Genomes to Lineages

- A **lineage** is simply a subtree of the overall phylogeny
  - Biologically, it's all viral sequences that inherited all mutations up to that point
- A **strain** is a significant deviation from the ancestral virus
  - I couldn't find a firm/specific definition for what exactly is considered "significant"
  - It seems to generally have implications of significant functional/phenotypic differences
- For SARS-CoV-2, people use **pangolin** to assign genomes to lineages
  - https://github.com/cov-lineages/pangolin

Studying the Prevalence of Mutations

Studying the Prevalence of Mutations

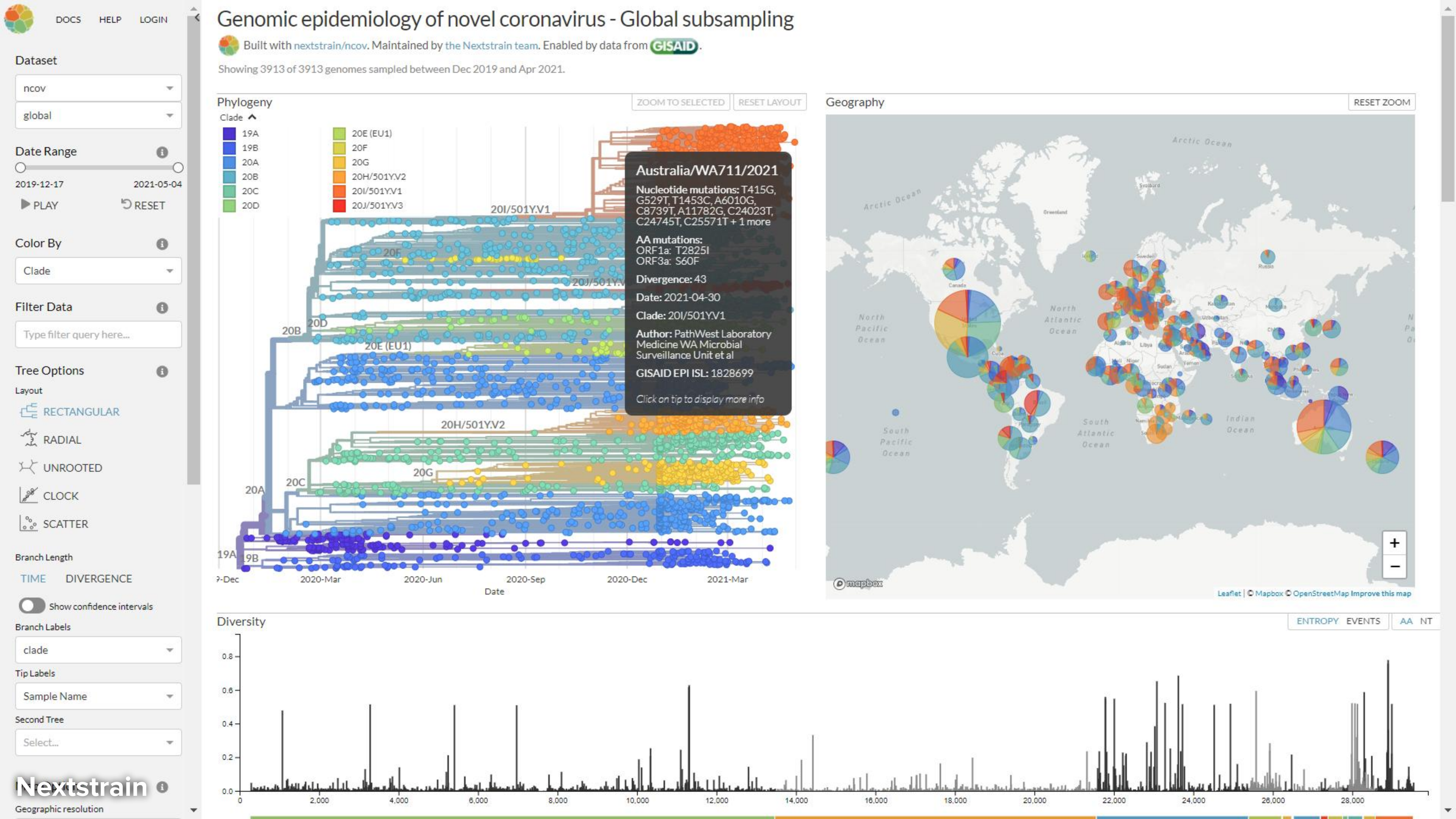
- **Nextstrain** is the most popular tool for tracking viral mutations

Studying the Prevalence of Mutations

- **Nextstrain** is the most popular tool for tracking viral mutations
    - Allows you to visualize the phylogeny + mutations + demographic data

Studying the Prevalence of Mutations

- **Nextstrain** is the most popular tool for tracking viral mutations
  - Allows you to visualize the phylogeny + mutations + demographic data
  - https://nextstrain.org/sars-cov-2



Genomic epidemiology of novel coronavirus - Global subsampling

Built with nextstrain/ncov. Maintained by the Nextstrain team. Enabled by data from GISAID.

Showing 3913 of 3913 genomes sampled between Dec 2019 and Apr 2021.

Dataset: ncov / global

Date Range: 2019-12-17 – 2021-05-04

Color By: Clade

Filter Data: Type filter query here...

Layout: RECTANGULAR

Branch Length: TIME

Branch Labels: clade

Tip Labels: Sample Name

Second Tree: Select...

Australia/WA711/2021

Nucleotide mutations: T415G, G529T, T1453C, A6010G, C8739T, A11782G, C24023T, C24745T, C25571T + 1 more

AA mutations:
ORF1a: T2825I
ORF3a: S60F

Divergence: 43

Date: 2021-04-30

Clade: 20I/501Y.V1

Author: PathWest Laboratory Medicine WA Microbial Surveillance Unit et al

GISAID EPI ISL: 1828699

Click on tip to display more info

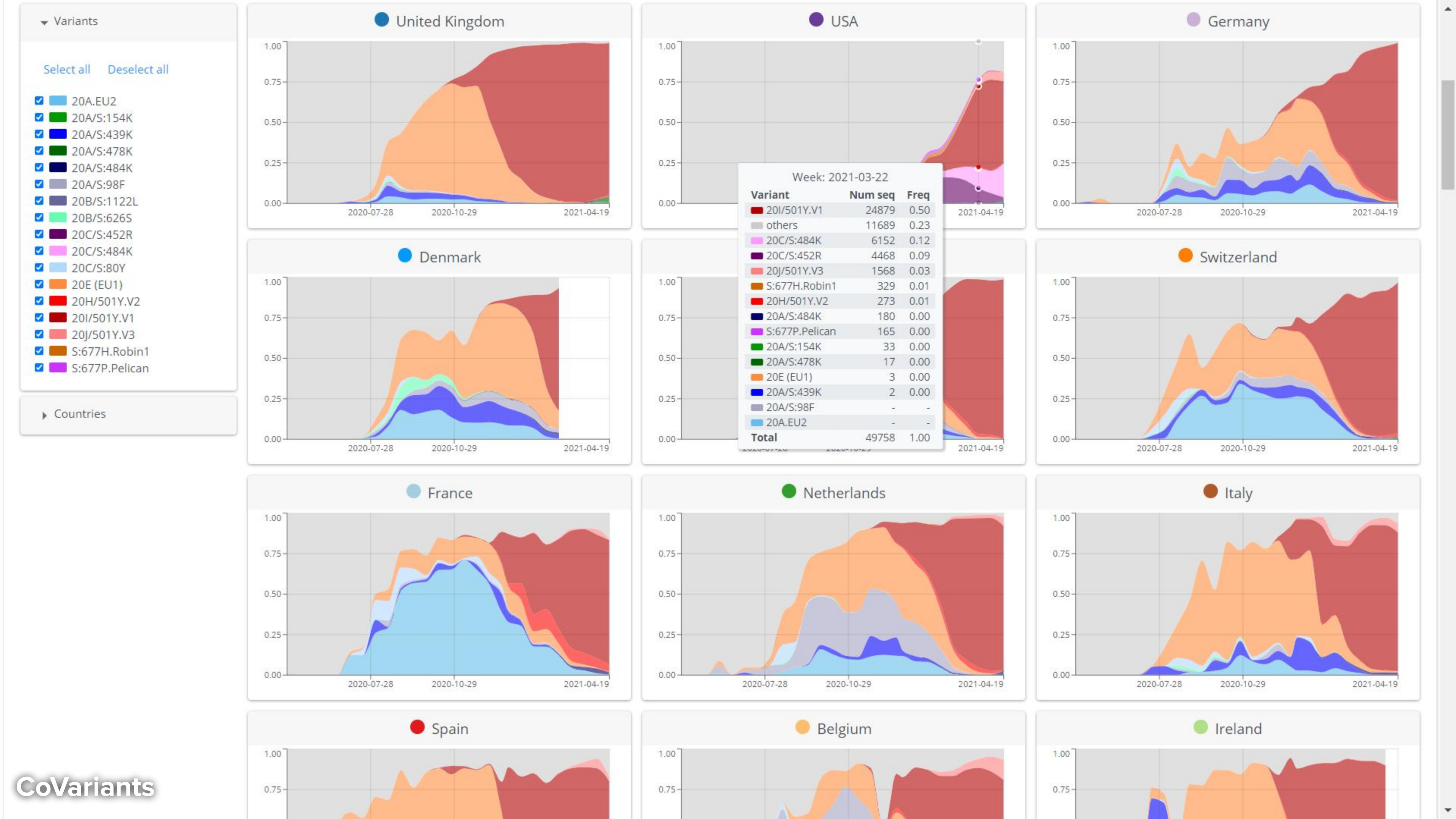
Nextstrain

Studying the Prevalence of Mutations

- **Nextstrain** is the most popular tool for tracking viral mutations
  - Allows you to visualize the phylogeny + mutations + demographic data
  - https://nextstrain.org/sars-cov-2
- **CoVariants** tracks SARS-CoV-2 variants + mutations over time

Studying the Prevalence of Mutations

- **Nextstrain** is the most popular tool for tracking viral mutations
  - Allows you to visualize the phylogeny + mutations + demographic data
  - https://nextstrain.org/sars-cov-2
- **CoVariants** tracks SARS-CoV-2 variants + mutations over time
  - https://covariants.org/



Variants

Select all Deselect all

- ☑ 20A.EU2
- ☑ 20A/S:154K
- ☑ 20A/S:439K
- ☑ 20A/S:478K
- ☑ 20A/S:484K
- ☑ 20A/S:98F
- ☑ 20B/S:1122L
- ☑ 20B/S:626S
- ☑ 20C/S:452R
- ☑ 20C/S:484K
- ☑ 20C/S:80Y
- ☑ 20E (EU1)
- ☑ 20H/501Y.V2
- ☑ 20I/501Y.V1
- ☑ 20J/501Y.V3
- ☑ S:677H.Robin1
- ☑ S:677P.Pelican

Countries

Week: 2021-03-22

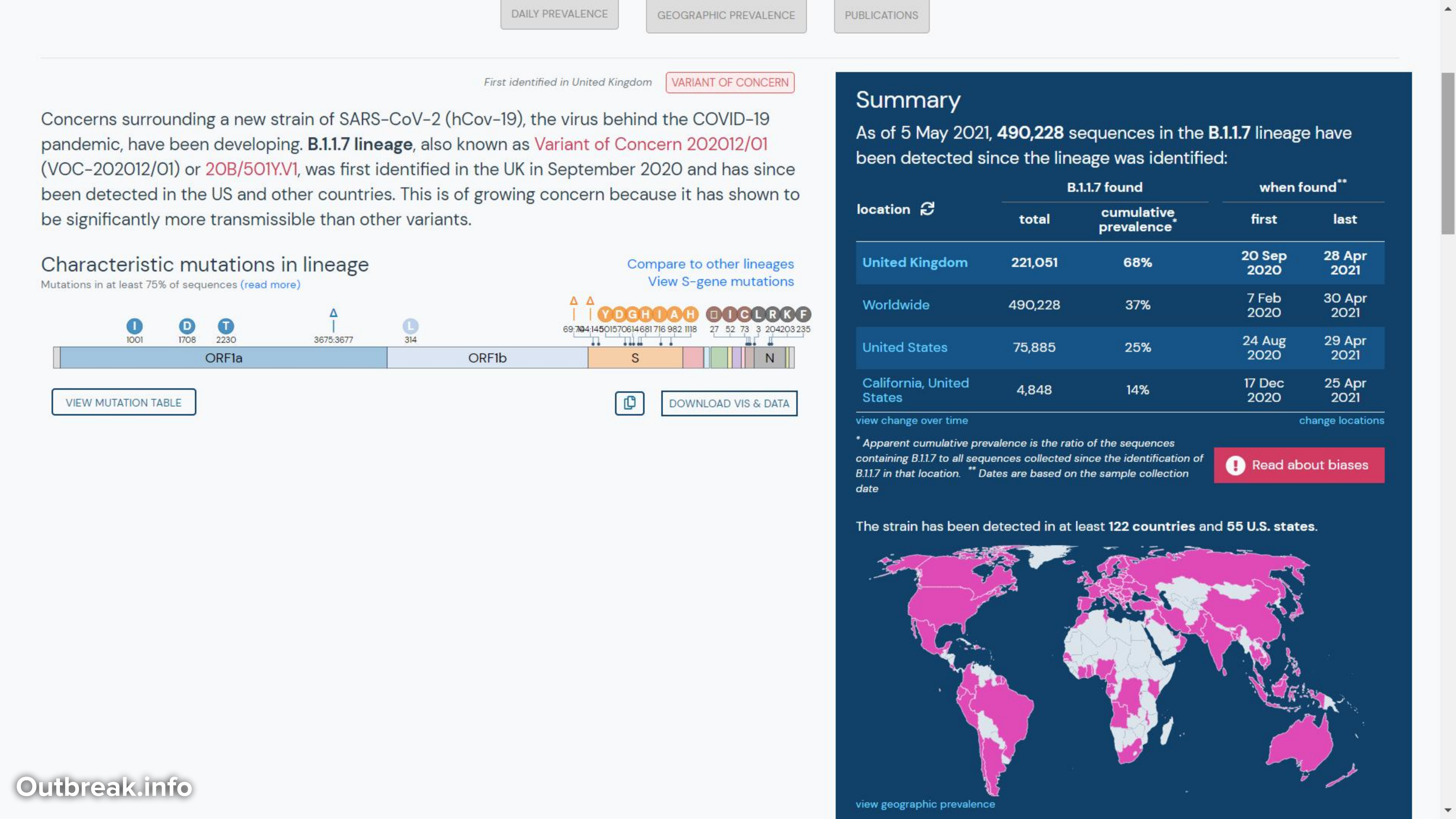
Variant	Num seq	Freq
20I/501Y.V1	24879	0.50
others	11689	0.23
20C/S:484K	6152	0.12
20C/S:452R	4468	0.09
20J/501Y.V3	1568	0.03
S:677H.Robin1	329	0.01
20H/501Y.V2	273	0.01
20A/S:484K	180	0.00
S:677P.Pelican	165	0.00
20A/S:154K	33	0.00
20A/S:478K	17	0.00
20E (EU1)	3	0.00
20A/S:439K	2	0.00
20A/S:98F	-	-
20A.EU2	-	-
Total	49758	1.00

Studying the Prevalence of Mutations

- **Nextstrain** is the most popular tool for tracking viral mutations
  - Allows you to visualize the phylogeny + mutations + demographic data
  - https://nextstrain.org/sars-cov-2
- **CoVariants** tracks SARS-CoV-2 variants + mutations over time
  - https://covariants.org/
- **Outbreak.info** tracks cases, deaths, and lineages in specific populations

Studying the Prevalence of Mutations

- **Nextstrain** is the most popular tool for tracking viral mutations
  - Allows you to visualize the phylogeny + mutations + demographic data
  - https://nextstrain.org/sars-cov-2
- **CoVariants** tracks SARS-CoV-2 variants + mutations over time
  - https://covariants.org/
- **Outbreak.info** tracks cases, deaths, and lineages in specific populations
  - https://outbreak.info/



First identified in United Kingdom VARIANT OF CONCERN

Concerns surrounding a new strain of SARS-CoV-2 (hCov-19), the virus behind the COVID-19 pandemic, have been developing. **B.1.1.7 lineage**, also known as Variant of Concern 202012/01 (VOC-202012/01) or 20B/501Y.V1, was first identified in the UK in September 2020 and has since been detected in the US and other countries. This is of growing concern because it has shown to be significantly more transmissible than other variants.

Characteristic mutations in lineage

Mutations in at least 75% of sequences (read more)

Compare to other lineages
View S-gene mutations

Summary

As of 5 May 2021, **490,228** sequences in the **B.1.1.7** lineage have been detected since the lineage was identified:

location	B.1.1.7 found		when found**	
	total	cumulative prevalence*	first	last
United Kingdom	221,051	68%	20 Sep 2020	28 Apr 2021
Worldwide	490,228	37%	7 Feb 2020	30 Apr 2021
United States	75,885	25%	24 Aug 2020	29 Apr 2021
California, United States	4,848	14%	17 Dec 2020	25 Apr 2021

view change over time
change locations

** Apparent cumulative prevalence is the ratio of the sequences containing B.1.1.7 to all sequences collected since the identification of B.1.1.7 in that location. ** Dates are based on the sample collection date*

Read about biases

The strain has been detected in at least **122 countries** and **55 U.S. states**.

view geographic prevalence

Outbreak.info

Molecular Cluster vs. Transmission Cluster

Molecular Cluster vs. Transmission Cluster

- **Molecular Cluster:** Group of persons linked by molecular (e.g. sequence) data

Molecular Cluster vs. Transmission Cluster

- **Molecular Cluster:** Group of persons linked by molecular (e.g. sequence) data
- A molecular cluster is a **subset** of a transmission cluster

Molecular Cluster vs. Transmission Cluster

- **Molecular Cluster:** Group of persons linked by molecular (e.g. sequence) data
- A molecular cluster is a **subset** of a transmission cluster
    - Not every infected person gets their molecular data collected (e.g. sequencing)

Molecular Cluster vs. Transmission Cluster

- **Molecular Cluster:** Group of persons linked by molecular (e.g. sequence) data
- A molecular cluster is a **subset** of a transmission cluster
  - Not every infected person gets their molecular data collected (e.g. sequencing)
  - A molecular cluster only consists of persons in a transmission cluster who got sequenced

Molecular Cluster vs. Transmission Cluster

- **Molecular Cluster:** Group of persons linked by molecular (e.g. sequence) data
- A molecular cluster is a **subset** of a transmission cluster
    - Not every infected person gets their molecular data collected (e.g. sequencing)
    - A molecular cluster only consists of persons in a transmission cluster who got sequenced
- We need to be conscious of people missing from a molecular cluster

Molecular Cluster vs. Transmission Cluster

- **Molecular Cluster:** Group of persons linked by molecular (e.g. sequence) data
- A molecular cluster is a **subset** of a transmission cluster
  - Not every infected person gets their molecular data collected (e.g. sequencing)
  - A molecular cluster only consists of persons in a transmission cluster who got sequenced
- We need to be conscious of people missing from a molecular cluster
  - Diagnosed but not sequenced (in *transmission* cluster, but not *molecular* cluster)

Molecular Cluster vs. Transmission Cluster

- **Molecular Cluster:** Group of persons linked by molecular (e.g. sequence) data
- A molecular cluster is a **subset** of a transmission cluster
    - Not every infected person gets their molecular data collected (e.g. sequencing)
    - A molecular cluster only consists of persons in a transmission cluster who got sequenced
- We need to be conscious of people missing from a molecular cluster
    - Diagnosed but not sequenced (in *transmission* cluster, but not *molecular* cluster)
    - Not diagnosed (in *risk network*, but not *transmission cluster*)

Molecular Clustering from Sequences

Molecular Clustering from Sequences

- Imagine I collect a viral sequence from person *x* and person *y*

Molecular Clustering from Sequences

- Imagine I collect a viral sequence from person *x* and person *y*
  - *x* and *y* were infected by the same person → $d(x,y)$ will be extremely small

Molecular Clustering from Sequences

- Imagine I collect a viral sequence from person *x* and person *y*
  - *x* and *y* were infected by the same person → $d(x,y)$ will be extremely small
  - *x* and *y* were not infected by the same person, but same group → $d(x,y)$ will be pretty small

Molecular Clustering from Sequences

- Imagine I collect a viral sequence from person *x* and person *y*
  - *x* and *y* were infected by the same person → *d*(*x*,*y*) will be extremely small
  - *x* and *y* were not infected by the same person, but same group → *d*(*x*,*y*) will be pretty small
  - *x* and *y* were infected from completely unrelated sources → *d*(*x*,*y*) wil be large

Molecular Clustering from Sequences

- Imagine I collect a viral sequence from person *x* and person *y*
  - *x* and *y* were infected by the same person → *d*(*x*,*y*) will be extremely small
  - *x* and *y* were not infected by the same person, but same group → *d*(*x*,*y*) will be pretty small
  - *x* and *y* were infected from completely unrelated sources → *d*(*x*,*y*) wil be large
- <u>Idea</u>: We can "link" *x* and *y* if *d*(*x*,*y*) is very small

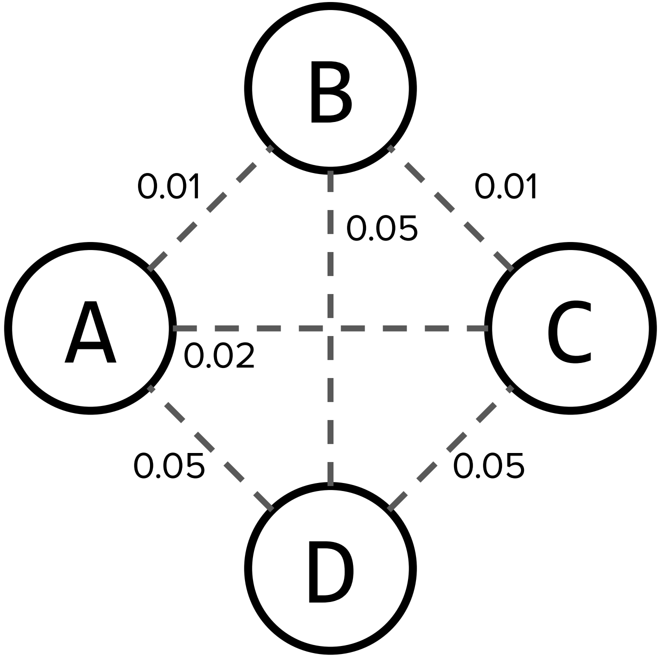
Molecular Clustering from Sequences

- Imagine I collect a viral sequence from person x and person y
  - x and y were infected by the same person → $d(x,y)$ will be extremely small
  - x and y were not infected by the same person, but same group → $d(x,y)$ will be pretty small
  - x and y were infected from completely unrelated sources → $d(x,y)$ wil be large
- Idea: We can “link” x and y if $d(x,y)$ is very small
  - For all pairs of individuals, “link” them if their pairwise sequence distance is small

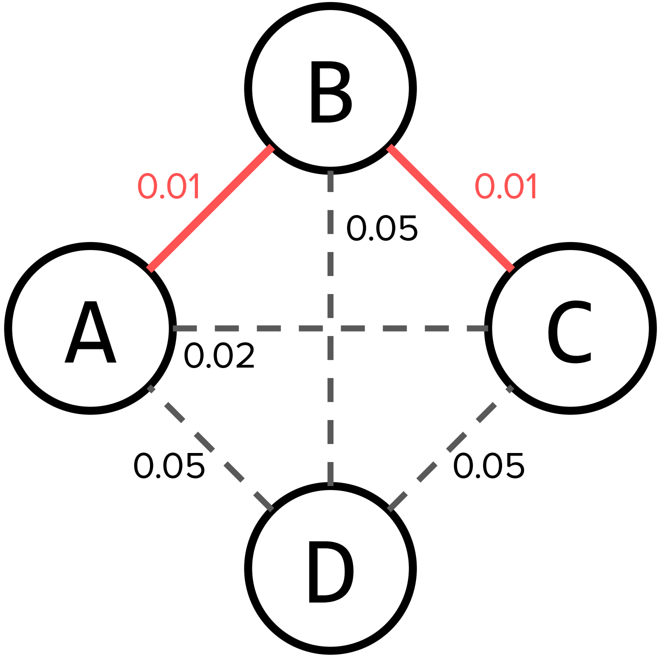
Molecular Clustering from Sequences

- Imagine I collect a viral sequence from person *x* and person *y*
  - *x* and *y* were infected by the same person → *d*(*x*,*y*) will be extremely small
  - *x* and *y* were not infected by the same person, but same group → *d*(*x*,*y*) will be pretty small
  - *x* and *y* were infected from completely unrelated sources → *d*(*x*,*y*) wil be large
- Idea: We can “link” *x* and *y* if *d*(*x*,*y*) is very small
  - For all pairs of individuals, “link” them if their pairwise sequence distance is small
  - Each resulting chain of links defines a molecular cluster

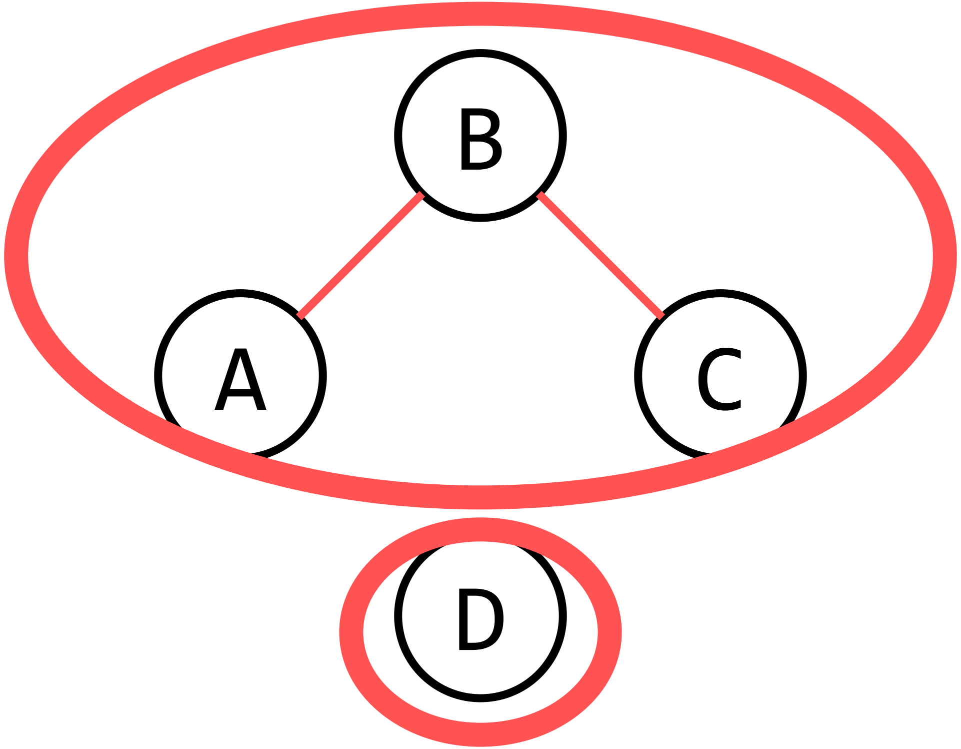
Molecular Clustering from Sequences

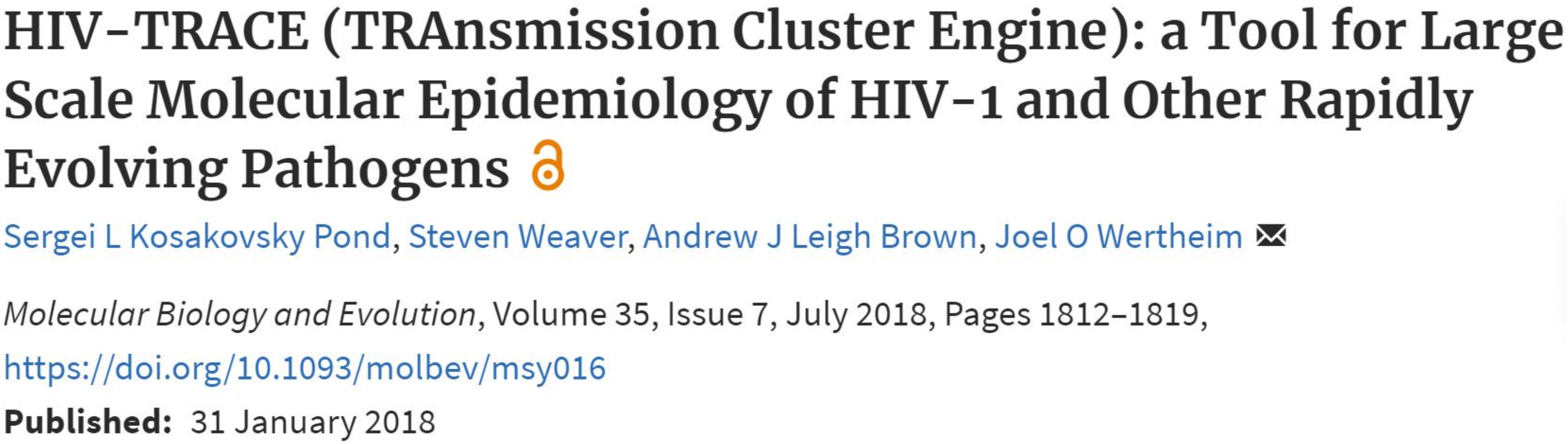


Molecular Clustering from Sequences



Molecular Clustering from Sequences





HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens

Sergei L Kosakovsky Pond, Steven Weaver, Andrew J Leigh Brown, Joel O Wertheim ✉

Molecular Biology and Evolution, Volume 35, Issue 7, July 2018, Pages 1812–1819, https://doi.org/10.1093/molbev/msy016

Published: 31 January 2018

https://github.com/veg/hivtrace

Molecular Clustering from a Phylogeny

Molecular Clustering from a Phylogeny

- The viral phylogeny is heavily constrained by the transmission network

Molecular Clustering from a Phylogeny

- The viral phylogeny is heavily constrained by the transmission network
    - *x* and *y* were infected by the same person ➔ *x* and *y* will be very close in the phylogeny

Molecular Clustering from a Phylogeny

- The viral phylogeny is heavily constrained by the transmission network
  - *x* and *y* were infected by the same person ➔ *x* and *y* will be very close in the phylogeny
  - *x* and *y* were not infected by the same person, but same group ➔ *x* and *y* will be pretty close

Molecular Clustering from a Phylogeny

- The viral phylogeny is heavily constrained by the transmission network
    - *x* and *y* were infected by the same person ➔ *x* and *y* will be very close in the phylogeny
    - *x* and *y* were not infected by the same person, but same group ➔ *x* and *y* will be pretty close
    - *x* and *y* were infected from completely unrelated sources ➔ *x* and *y* will be far apart

Molecular Clustering from a Phylogeny

- The viral phylogeny is heavily constrained by the transmission network
    - *x* and *y* were infected by the same person ➔ *x* and *y* will be very close in the phylogeny
    - *x* and *y* were not infected by the same person, but same group ➔ *x* and *y* will be pretty close
    - *x* and *y* were infected from completely unrelated sources ➔ *x* and *y* will be far apart
- Idea: We can define clusters using relationships within the phylogeny

Molecular Clustering from a Phylogeny

- The viral phylogeny is heavily constrained by the transmission network
  - *x* and *y* were infected by the same person → *x* and *y* will be very close in the phylogeny
  - *x* and *y* were not infected by the same person, but same group → *x* and *y* will be pretty close
  - *x* and *y* were infected from completely unrelated sources → *x* and *y* will be far apart
- Idea: We can define clusters using relationships within the phylogeny
  - There are quite a few ways to define “clusters” given a phylogeny

Molecular Clustering from a Phylogeny

- The viral phylogeny is heavily constrained by the transmission network
  - *x* and *y* were infected by the same person ➔ *x* and *y* will be very close in the phylogeny
  - *x* and *y* were not infected by the same person, but same group ➔ *x* and *y* will be pretty close
  - *x* and *y* were infected from completely unrelated sources ➔ *x* and *y* will be far apart
- Idea: We can define clusters using relationships within the phylogeny
  - There are quite a few ways to define "clusters" given a phylogeny
  - Subtree with maximum pairwise distance below some threshold? Cutting the tree in some way?

Molecular Clustering from a Phylogeny

- The viral phylogeny is heavily constrained by the transmission network
  - *x* and *y* were infected by the same person ➔ *x* and *y* will be very close in the phylogeny
  - *x* and *y* were not infected by the same person, but same group ➔ *x* and *y* will be pretty close
  - *x* and *y* were infected from completely unrelated sources ➔ *x* and *y* will be far apart
- Idea: We can define clusters using relationships within the phylogeny
  - There are quite a few ways to define “clusters” given a phylogeny
  - Subtree with maximum pairwise distance below some threshold? Cutting the tree in some way?
  - Single-linkage like in HIV-TRACE (but using pairwise distances from the tree)?

TreeCluster: Clustering biological sequences using phylogenetic trees

Metin Balaban, Niema Moshiri, Uyen Mai, Xingfan Jia, Siavash Mirarab ✉

Published: August 22, 2019 • https://doi.org/10.1371/journal.pone.0221068

https://github.com/niemasd/TreeCluster

Questions?