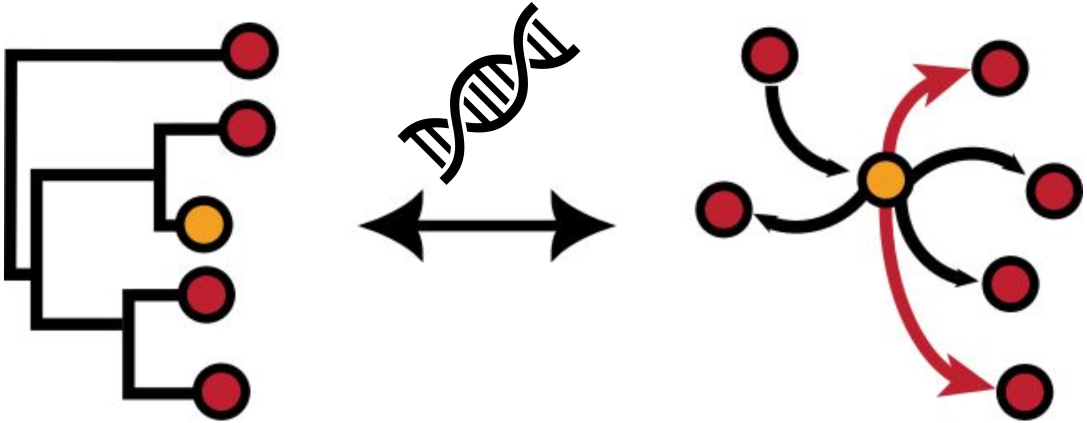
EPAH6052: Pathogen Genomic Epidemiology

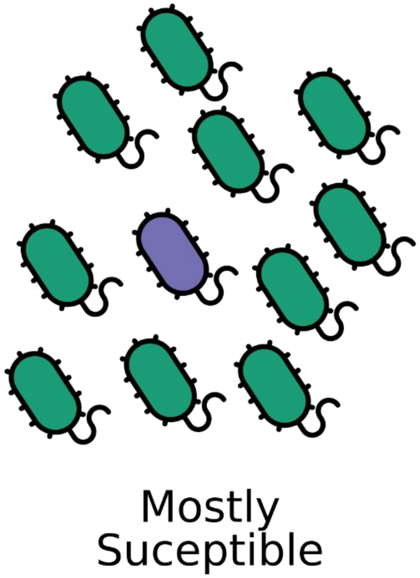


Finlay Maguire
Dalhousie University
Shared Hospital Laboratory
Public Health Alliance for Genomic Epidemiology

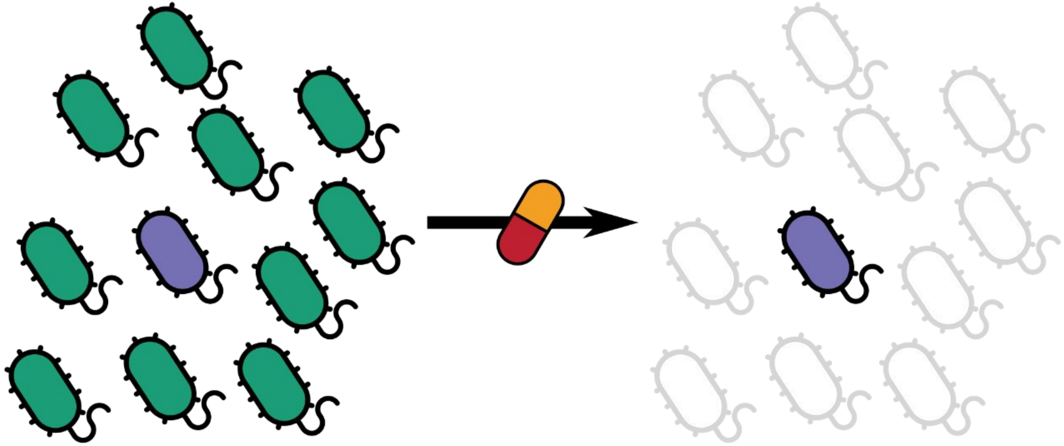
Outcomes

- Explain how evolution and infectious disease epidemiology are connected
- Identify what additional information genomic data can provide
- Give example of how genomics can be used for diagnostics
- Explain how you can infer an evolutionary tree (phylogeny)
- Provide examples of processes which determine the shape of a phylogeny
- Articulate the role of Bayesian models in genomic epidemiology
- List at least 3 ways in which a phylogeny can be used in epidemiology

Agents of infectious diseases undergo evolution

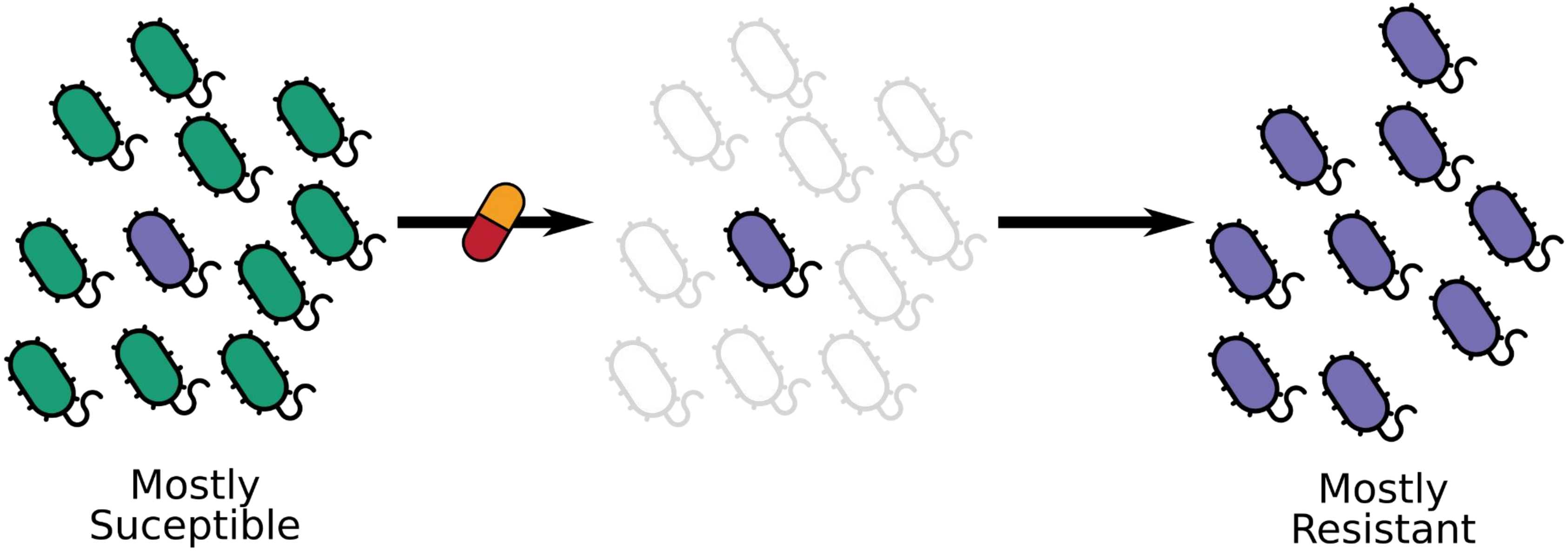


Agents of infectious diseases undergo evolution

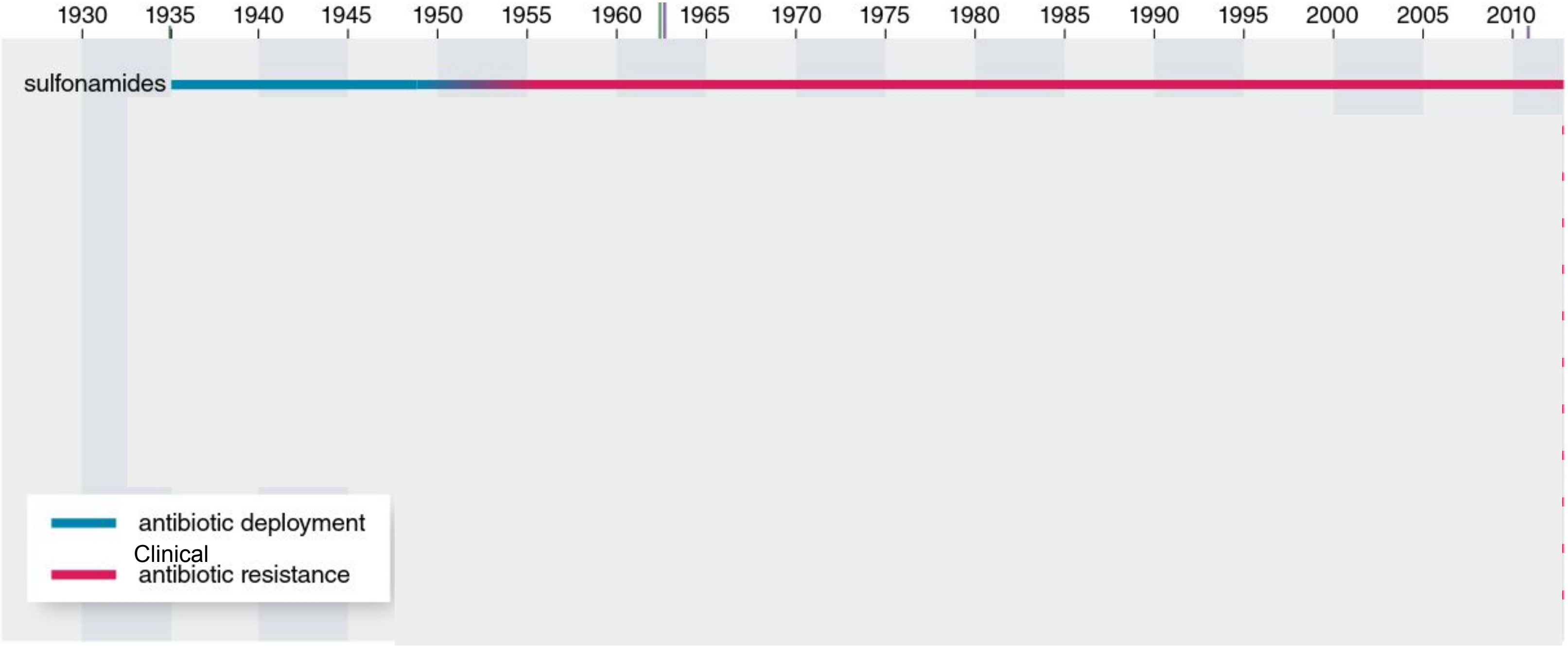


Mostly
Suceptible

Agents of infectious diseases undergo evolution

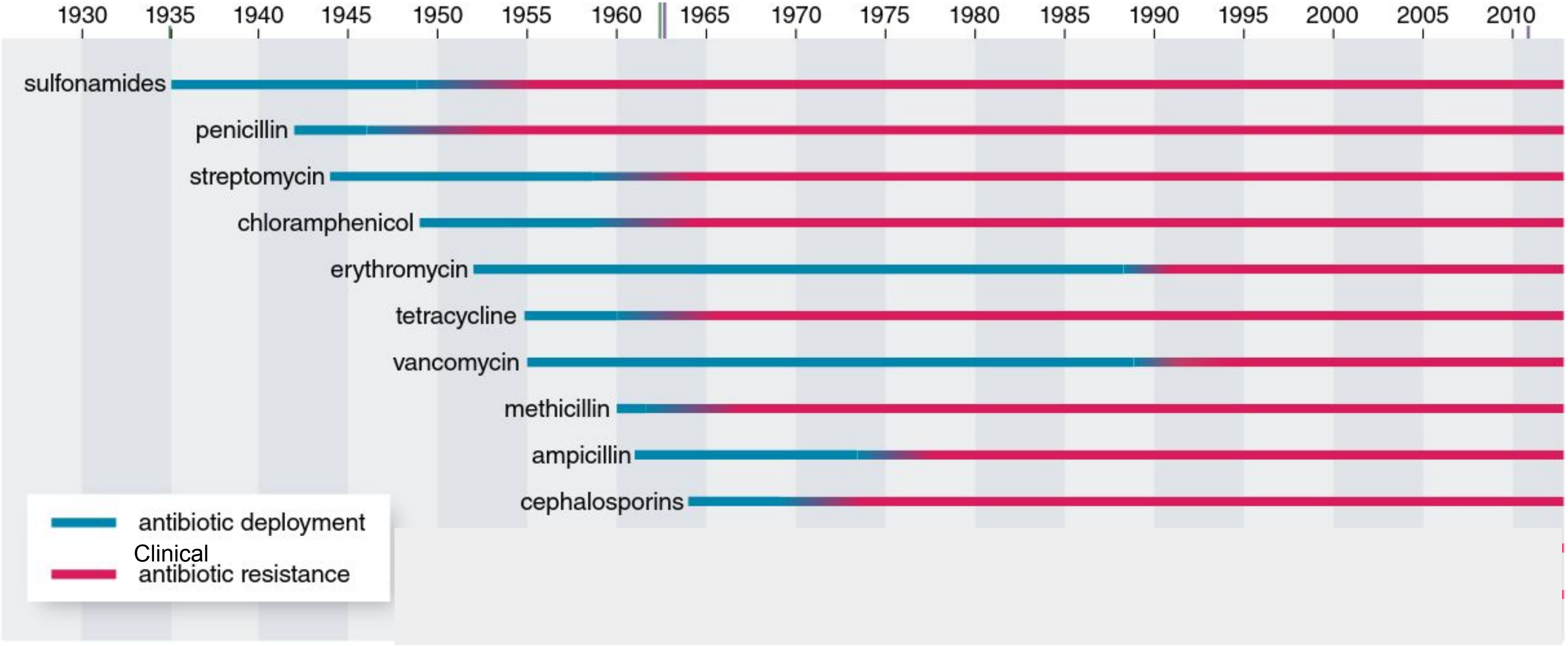


Pathogen evolution impacts treatment



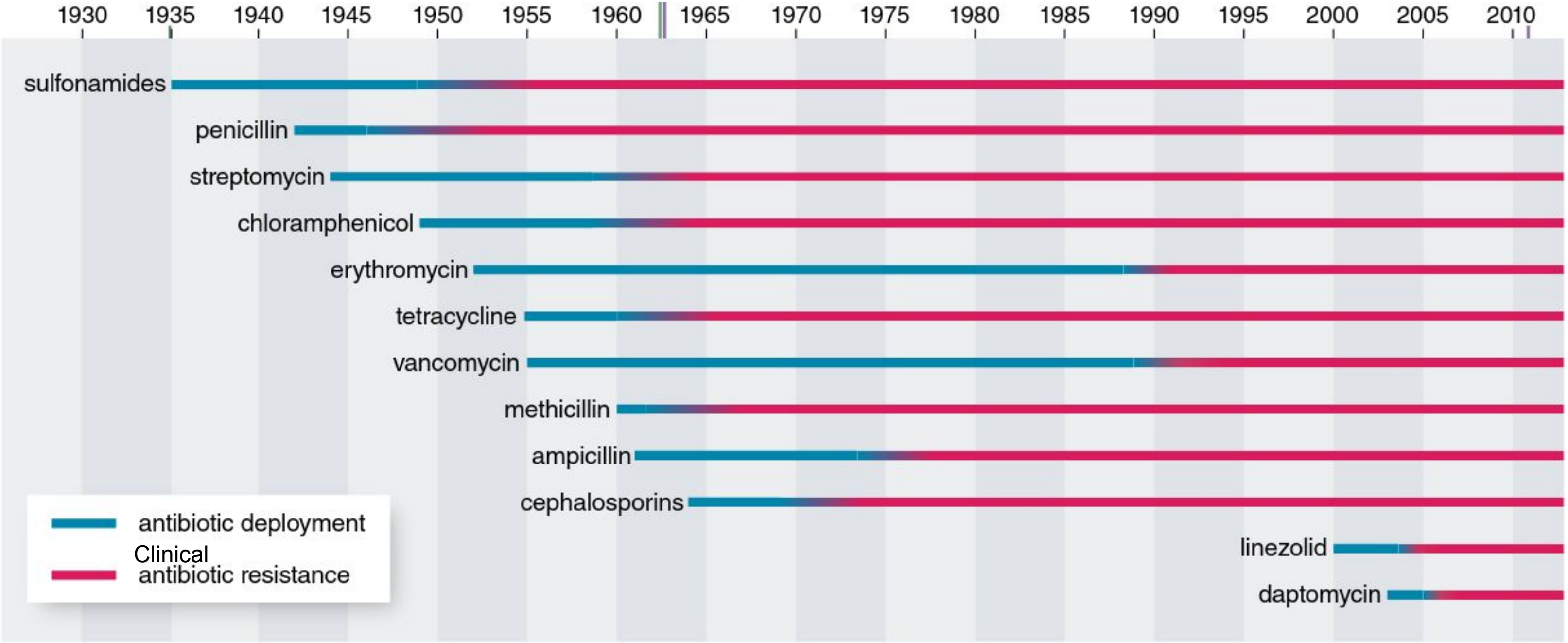
10.1511/2014.106.42

Pathogen evolution impacts treatment



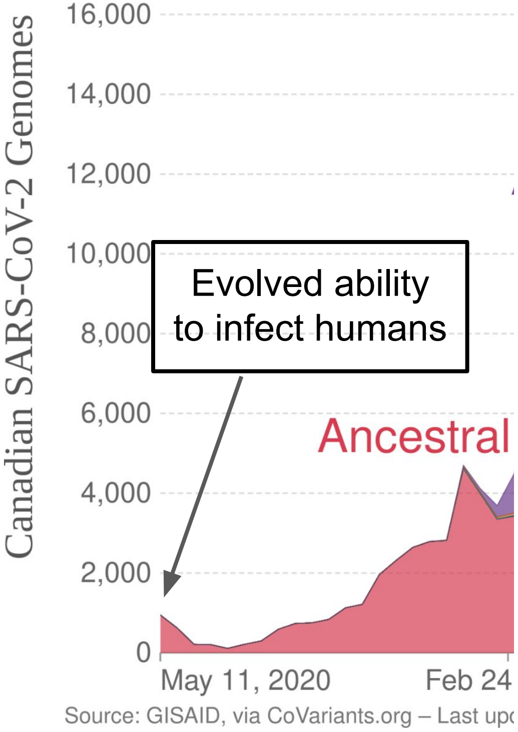
10.1511/2014.106.42

Pathogen evolution impacts treatment



10.1511/2014.106.42

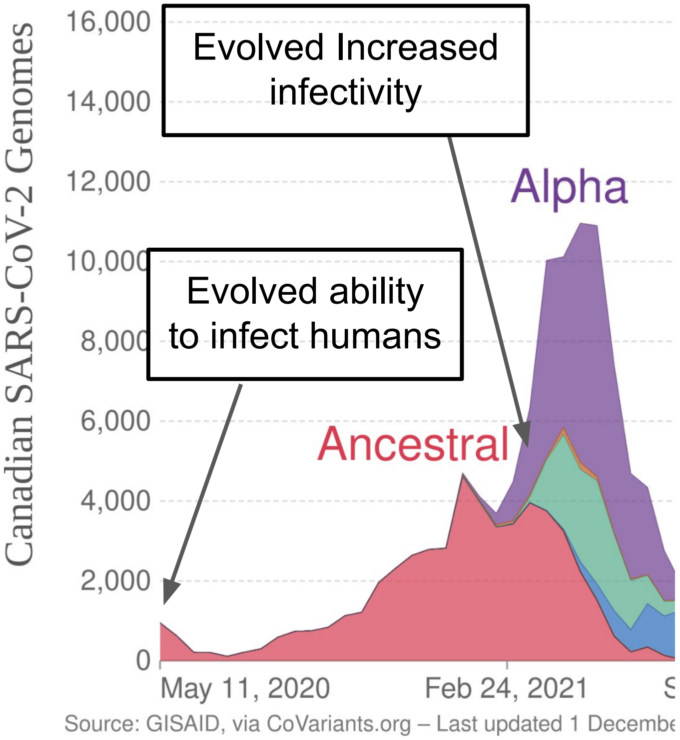
Evolution drives cases: SARS-CoV-2 waves



Evolved ability to infect humans

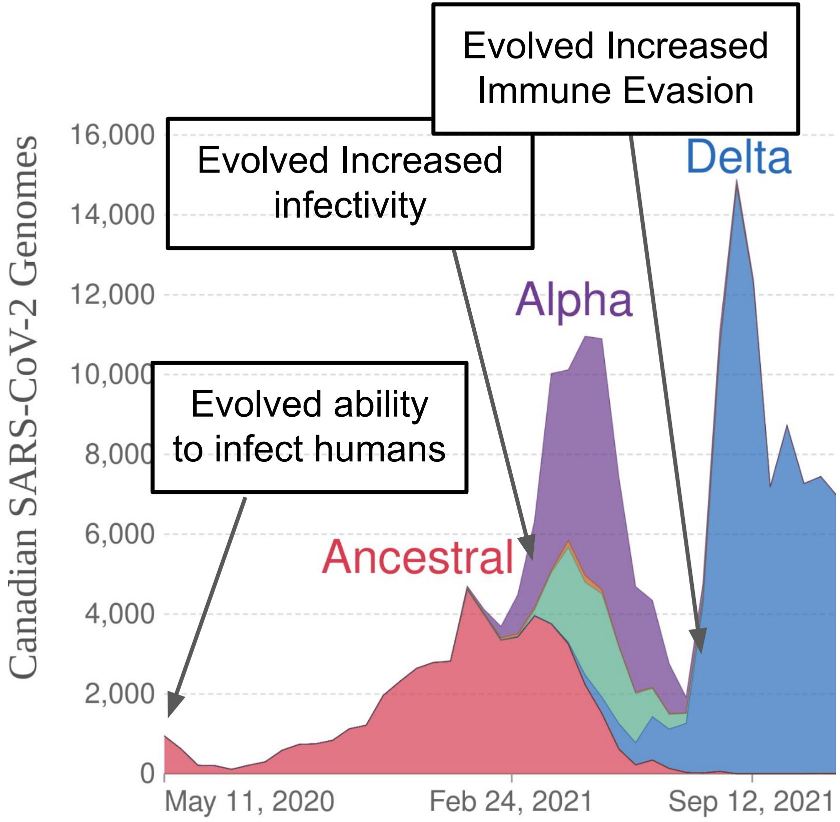
Source: GISAID, via CoVariants.org – Last upc

Evolution drives cases: SARS-CoV-2 waves



Source: GISAID, via CoVariants.org – Last updated 1 Decembe

Evolution drives cases: SARS-CoV-2 waves



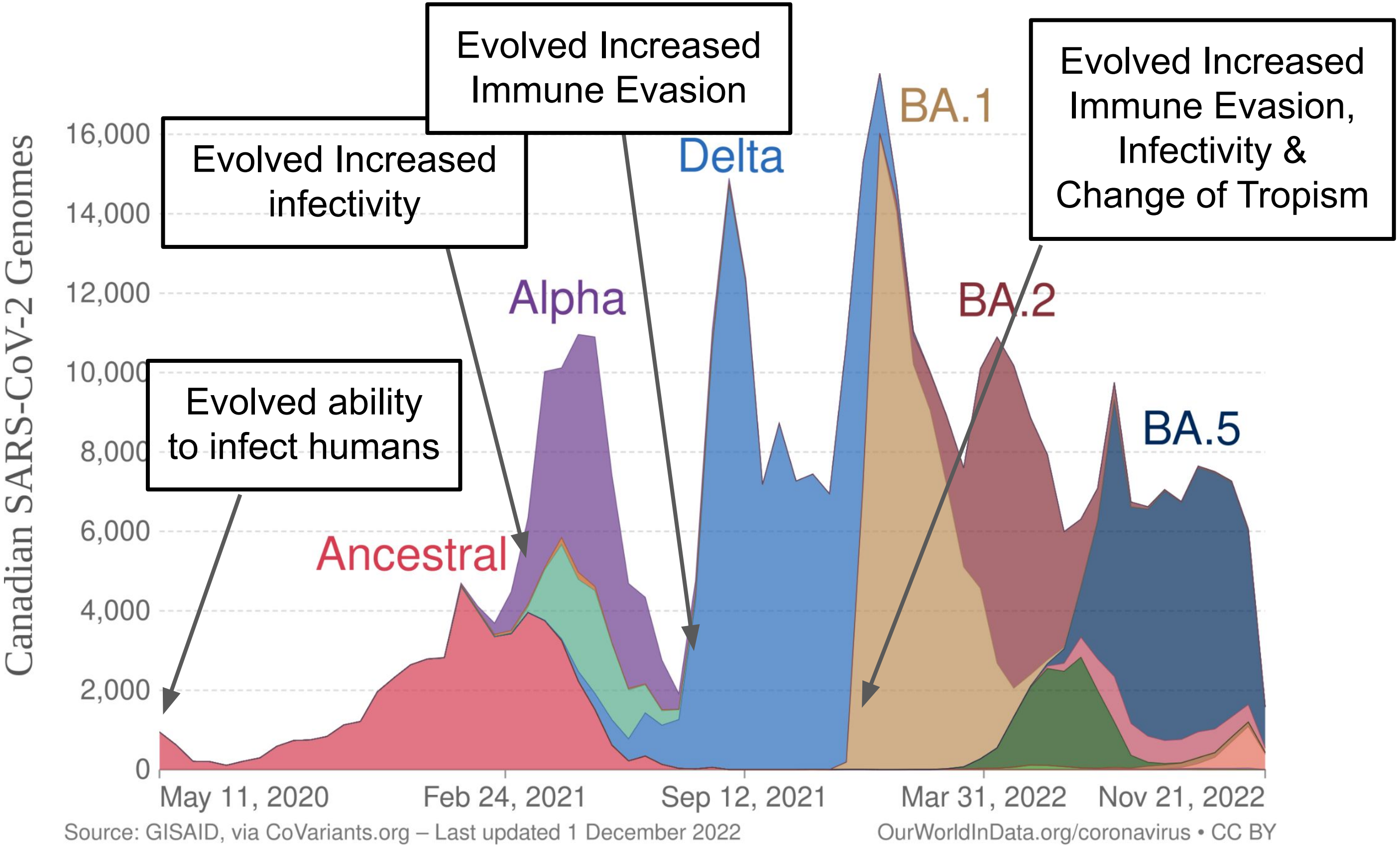
Evolved Increased Immune Evasion

Evolved Increased infectivity

Evolved ability to infect humans

Source: GISAID, via CoVariants.org – Last updated 1 December 2022

Evolution drives cases: SARS-CoV-2 waves


Evolved ability to infect humans

Evolved Increased infectivity

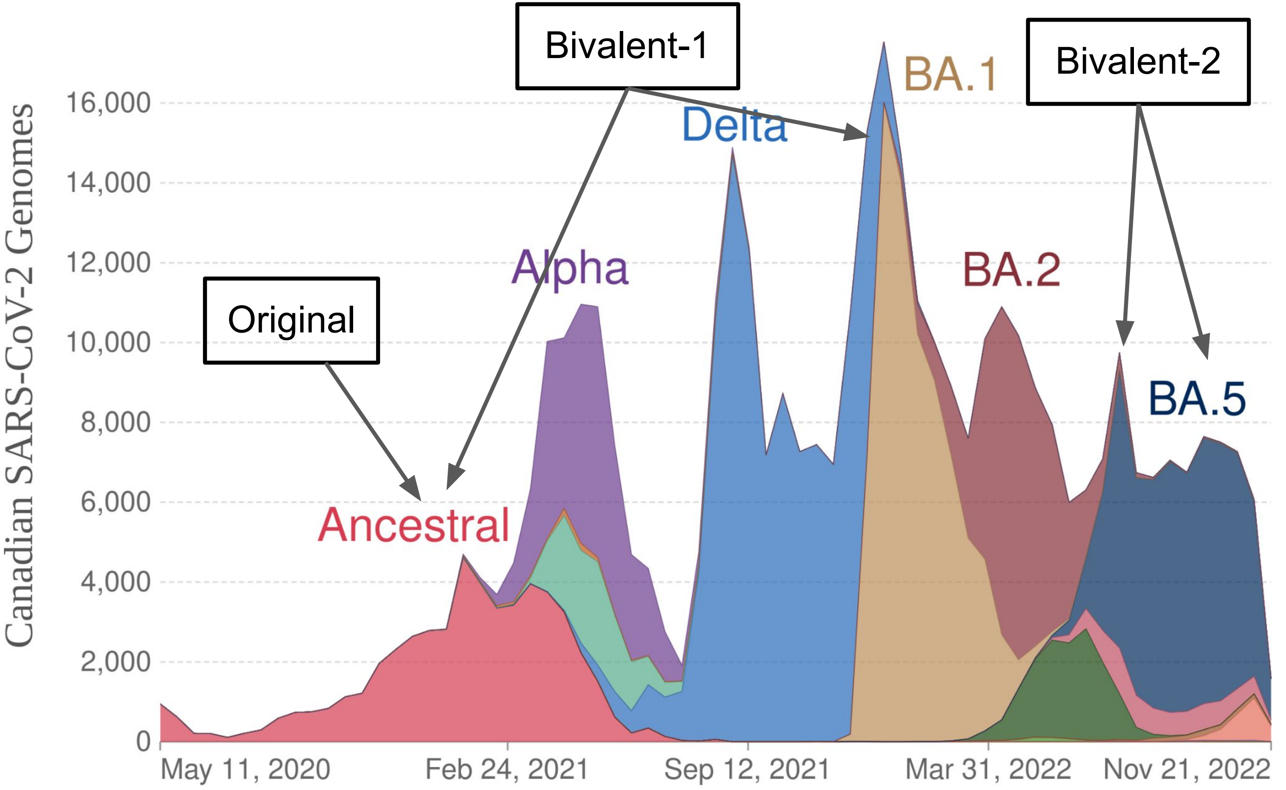
Evolved Increased Immune Evasion

Evolved Increased Immune Evasion, Infectivity & Change of Tropism

Source: GISAID, via CoVariants.org – Last updated 1 December 2022

OurWorldInData.org/coronavirus • CC BY

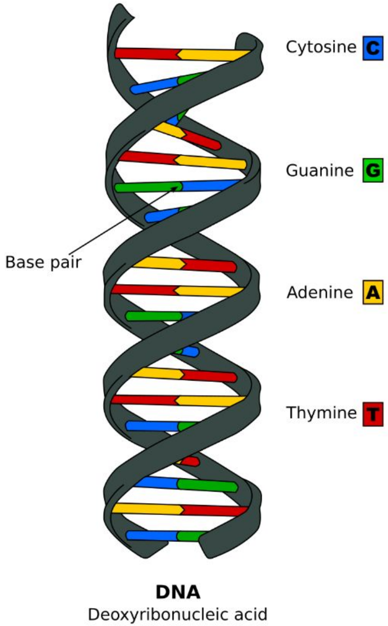
Evolution drives vaccine design/effectiveness



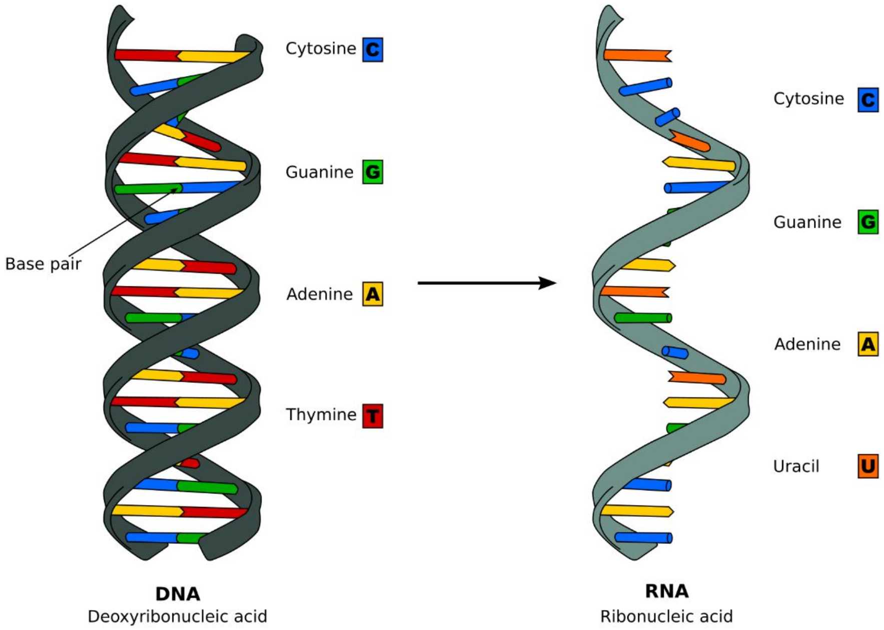
Source: GISAID, via CoVariants.org – Last updated 1 December 2022 OurWorldInData.org/coronavirus • CC BY

How can we monitor evolution?

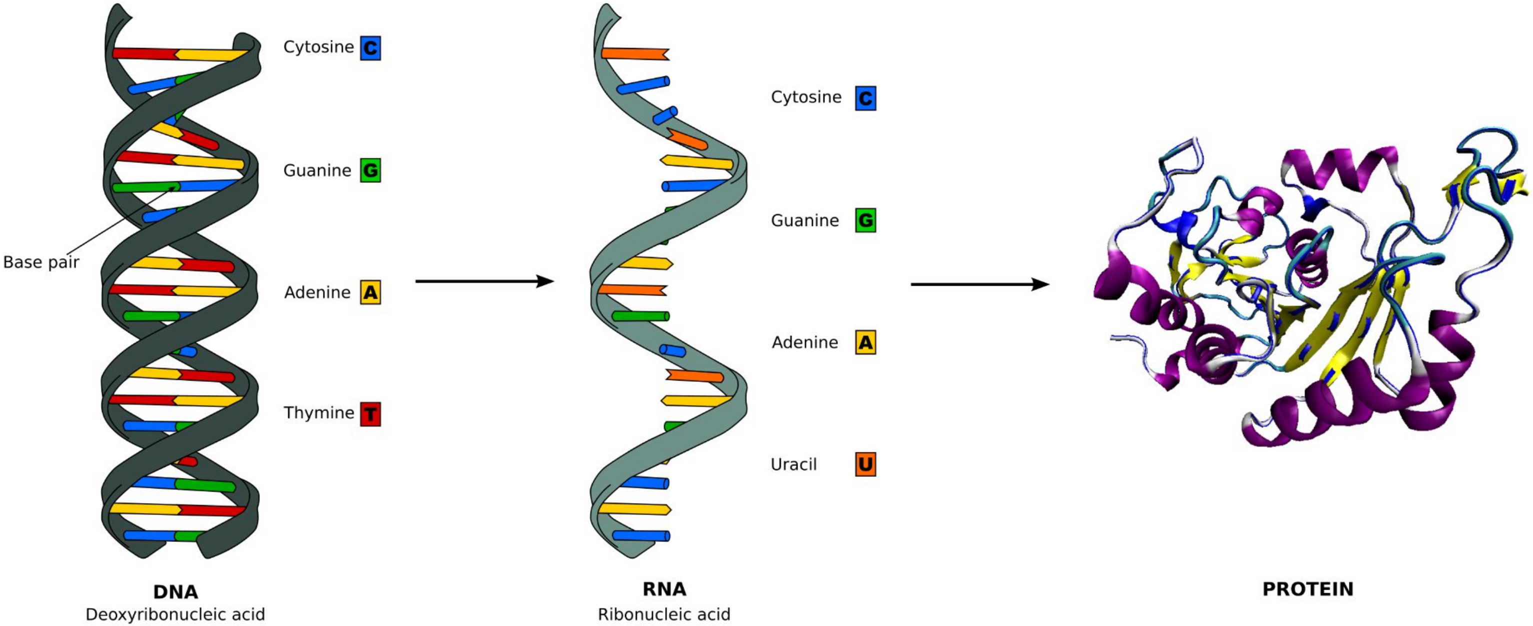
Genomes are the substrate of evolution



Genomes are the substrate of evolution

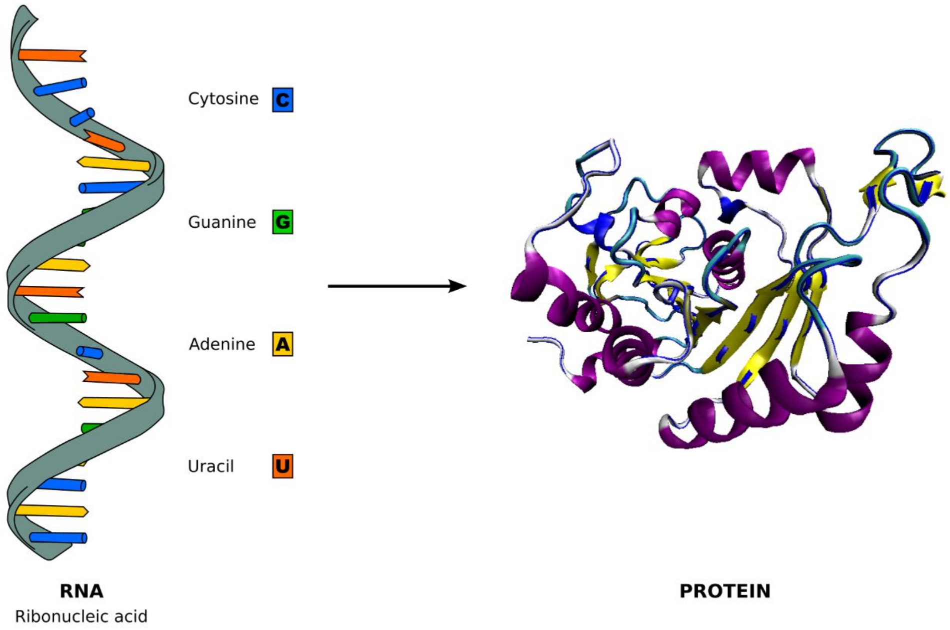


Genomes are the substrate of evolution



- DNA encodes RNA which encodes proteins

Genomes are the substrate of evolution



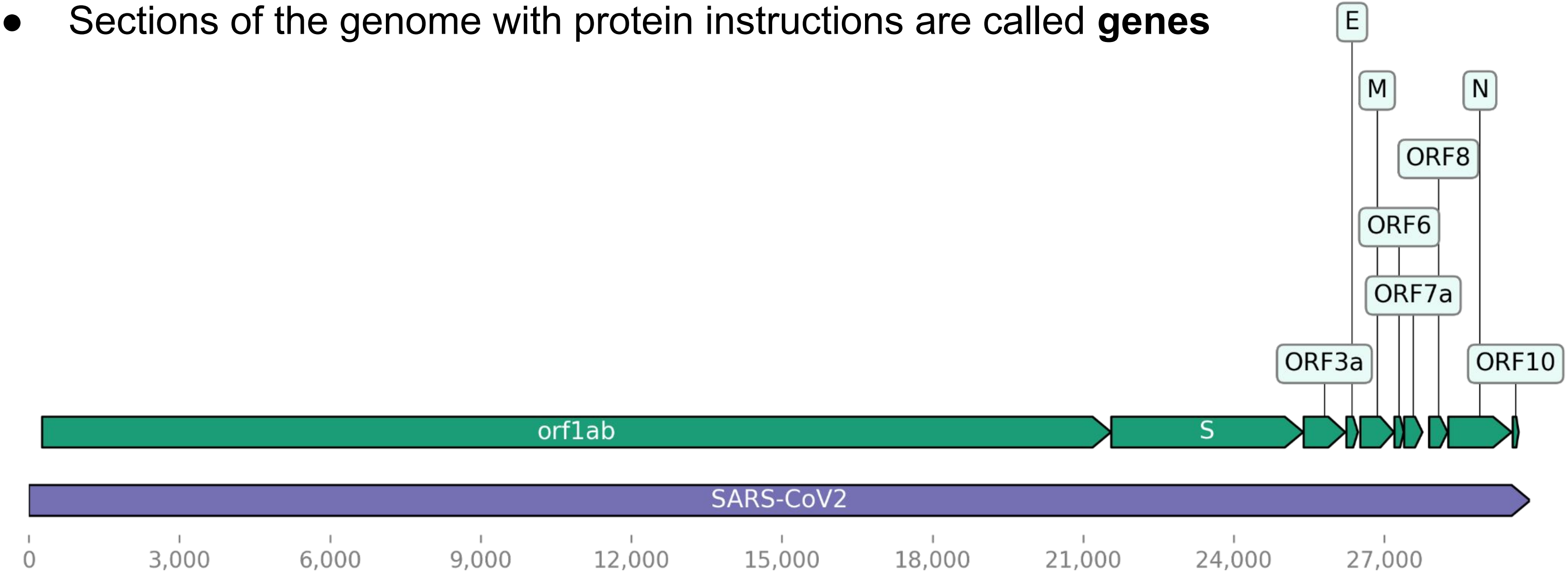
- DNA encodes RNA which encodes proteins
- Viruses like SARS-CoV-2 skip DNA

Genomes are the substrate of evolution

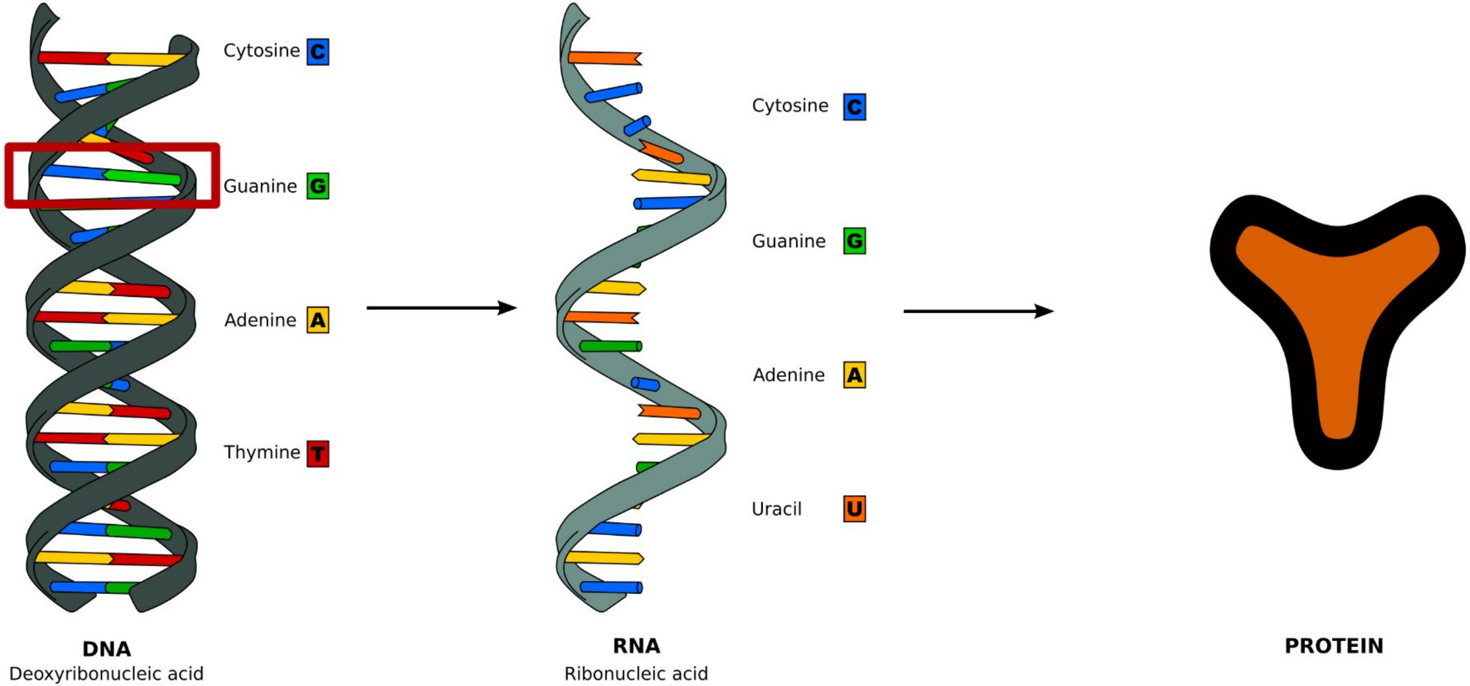
- Genomes are the complete collection of genetic instructions

Genomes are the substrate of evolution

- Genomes are the complete collection of genetic instructions
- Sections of the genome with protein instructions are called **genes**

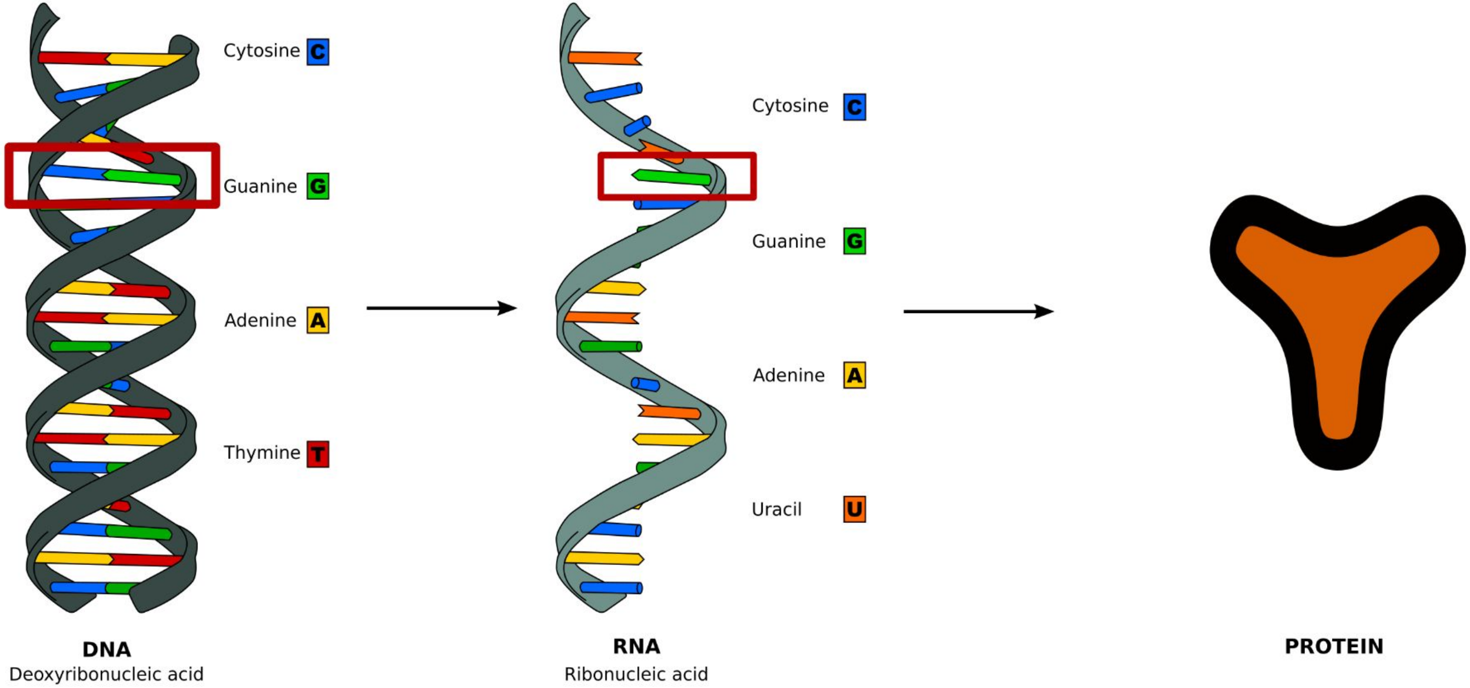


Mutations provide the variation upon which evolution acts



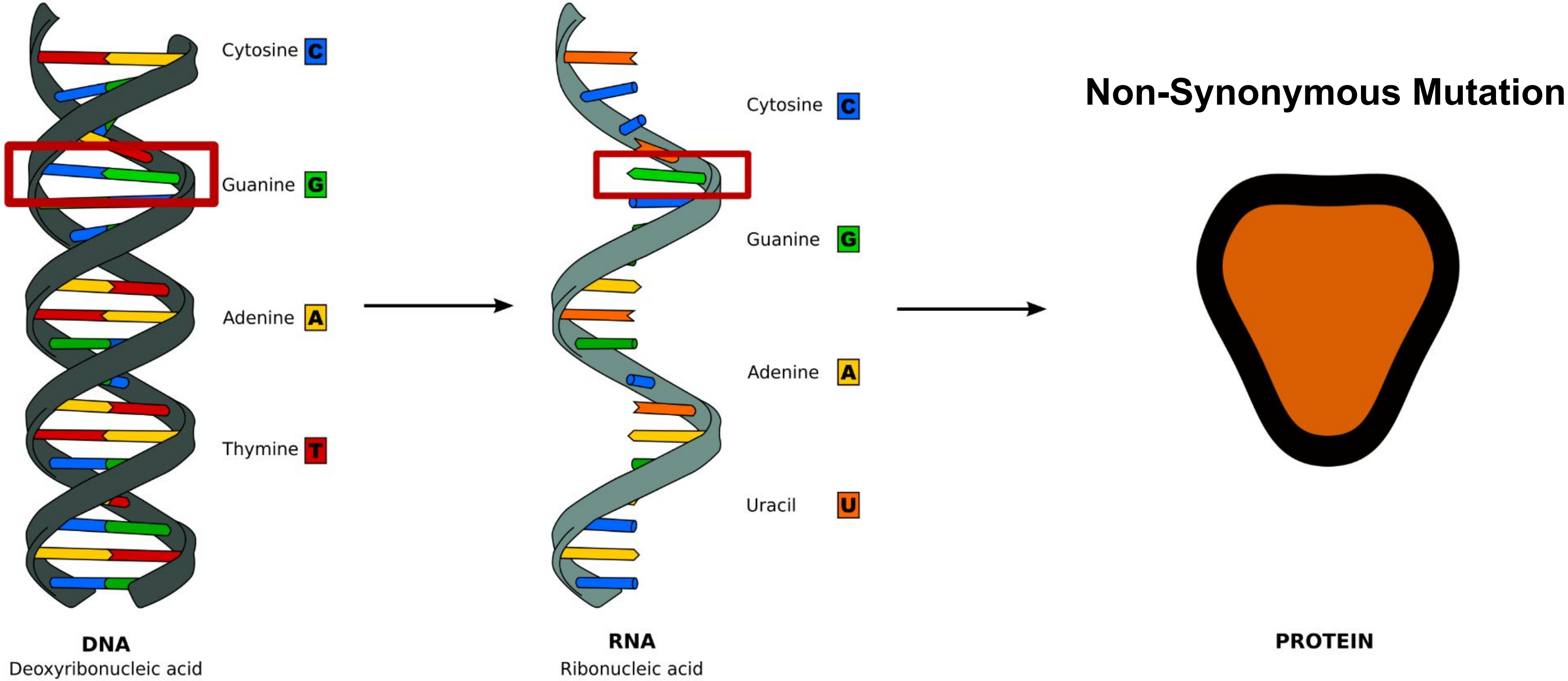
- Genome copying is error-prone

Mutations provide the variation upon which evolution acts



- Genome copying is error-prone
- Errors are called mutations

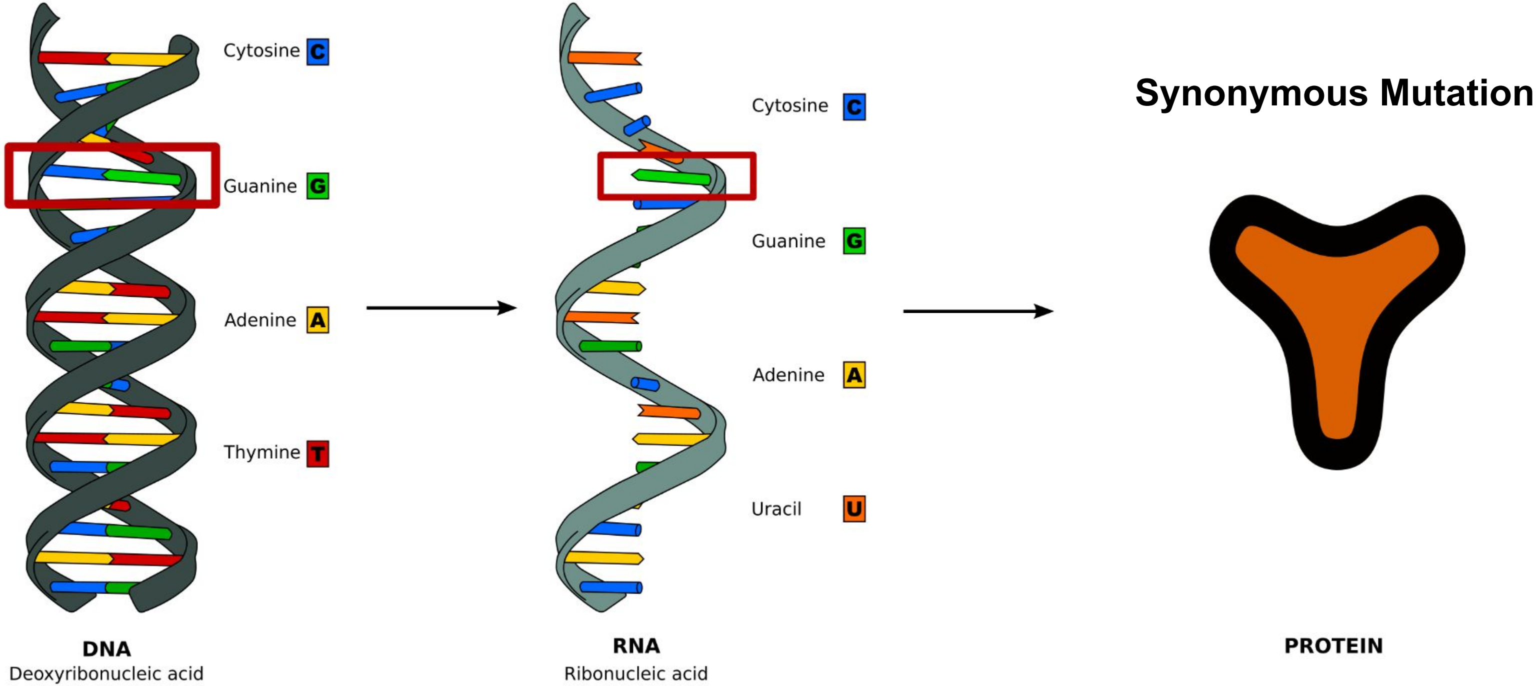
Mutations provide the variation upon which evolution acts



Non-Synonymous Mutation

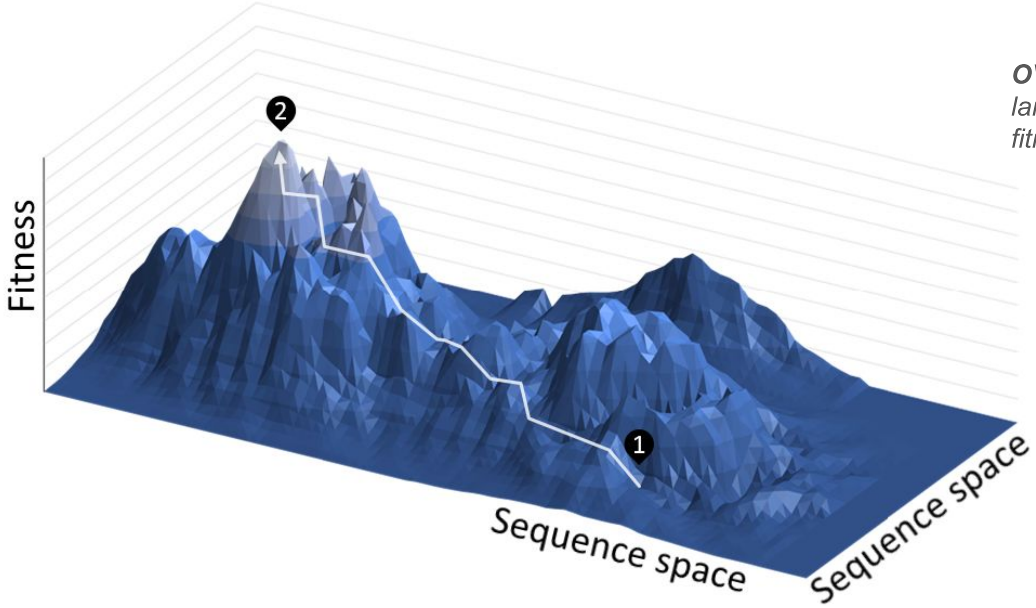
- Genome copying is error-prone
- Errors are called mutations
- Mutations can change protein sequence

Mutations provide the variation upon which evolution acts



- Genome copying is error-prone
- Errors are called mutations
- Mutations can change protein sequence - *but don't always*

Mutations are a random walk across a fitness landscape

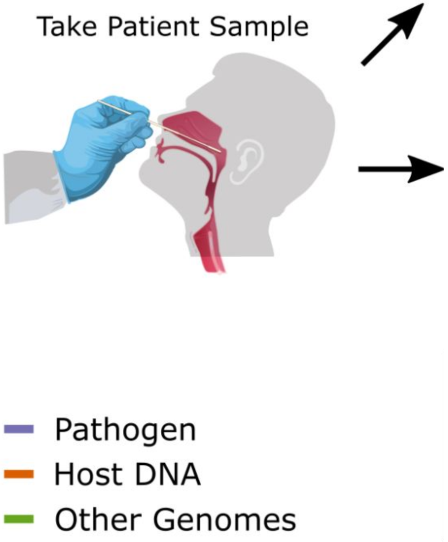


OVERSIMPLIFICATION - *actual landscapes are dynamic/changing & fitness is hard to measure*

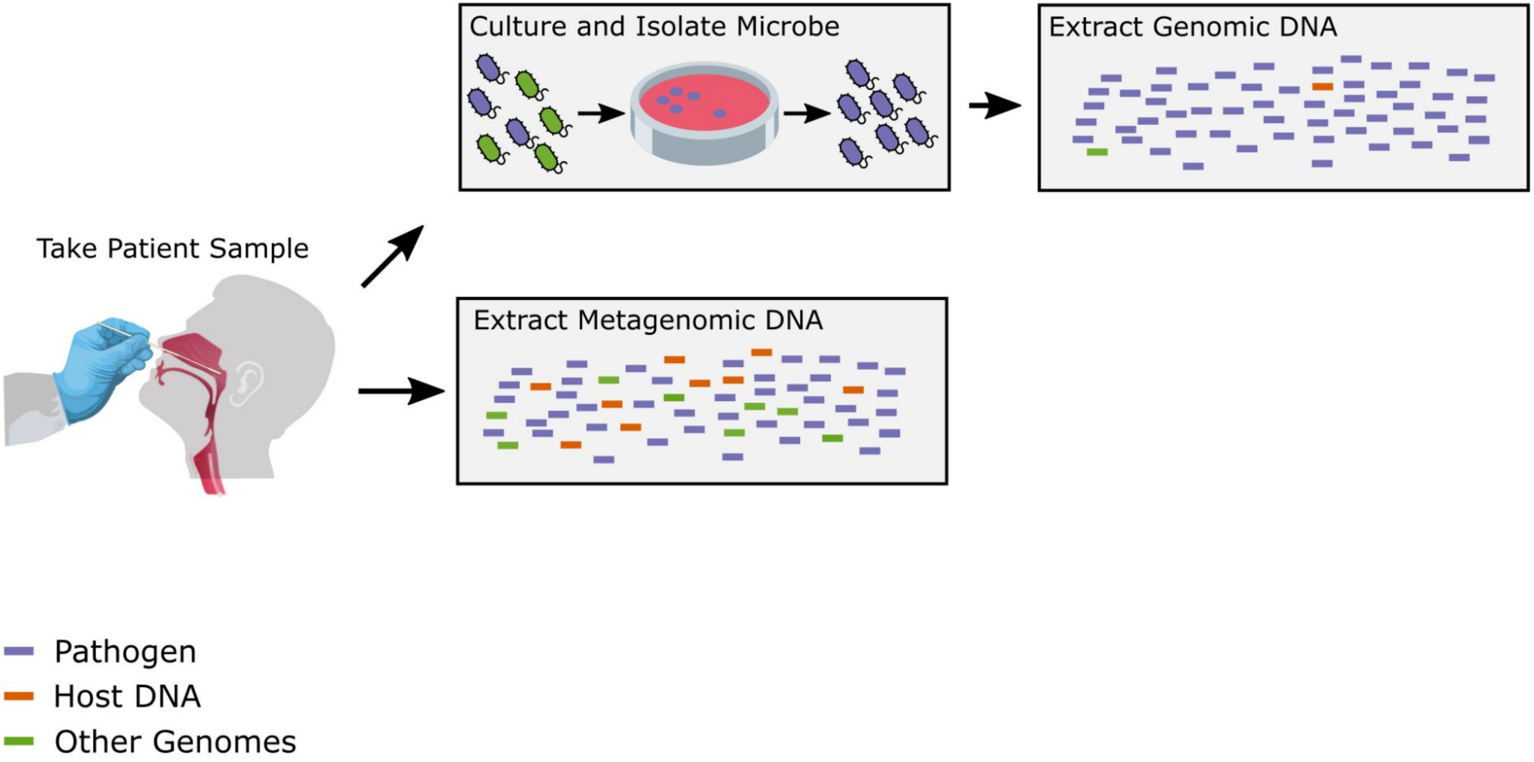
- *Fitness = quantitative representation of individual reproductive success*
- *High Fitness = more descendants (larger proportion of circulating population)*
- *Low Fitness = fewer descendants (higher chance of dying out)*

So, how do we get genomes?

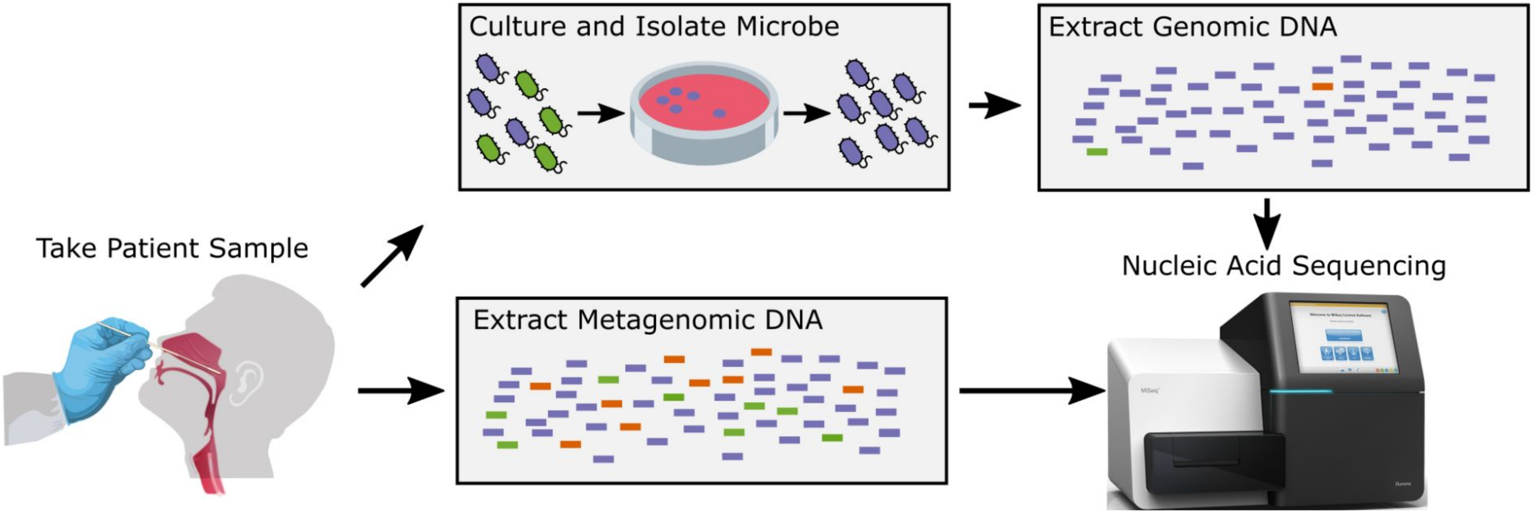
Sequencing Pathogen Genomes



Sequencing Pathogen Genomes



Sequencing Pathogen Genomes



Take Patient Sample

Culture and Isolate Microbe

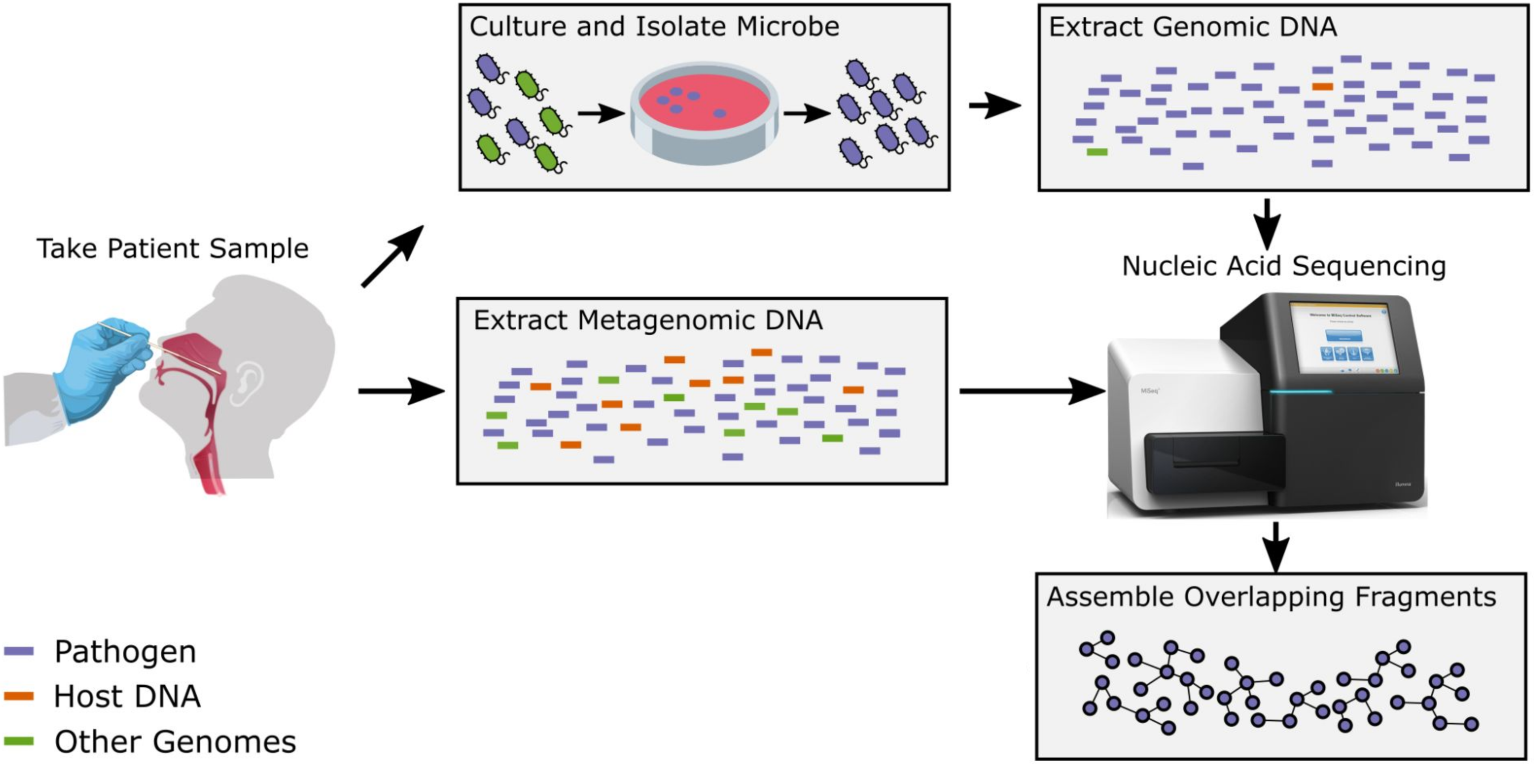
Extract Genomic DNA

Extract Metagenomic DNA

Nucleic Acid Sequencing

- 🟪 Pathogen
- 🟧 Host DNA
- 🟩 Other Genomes

Sequencing Pathogen Genomes



Take Patient Sample

Culture and Isolate Microbe

Extract Genomic DNA

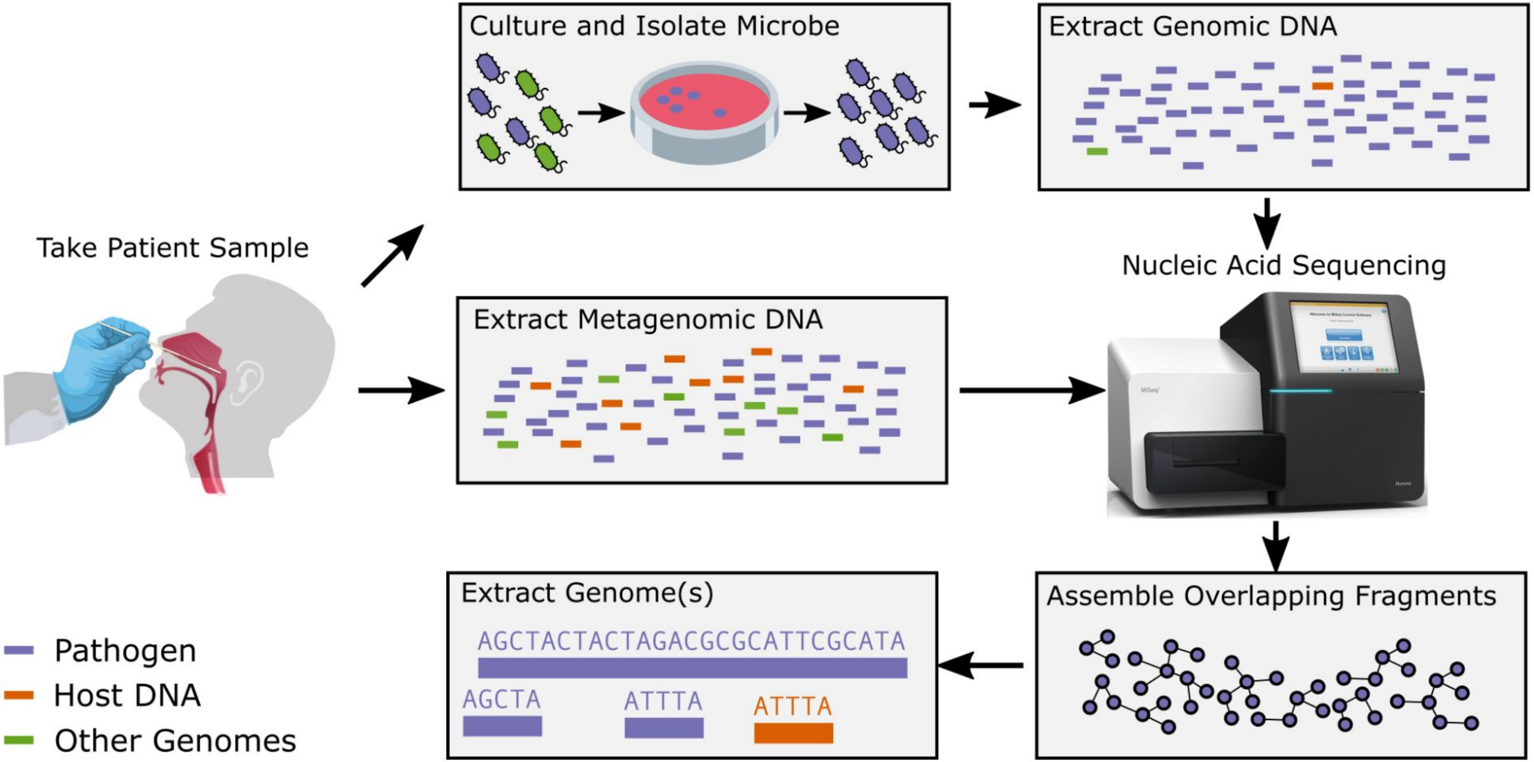
Extract Metagenomic DNA

Nucleic Acid Sequencing

Assemble Overlapping Fragments

- 🟪 Pathogen
- 🟧 Host DNA
- 🟩 Other Genomes

Sequencing Pathogen Genomes



Take Patient Sample

Culture and Isolate Microbe

Extract Genomic DNA

Extract Metagenomic DNA

Nucleic Acid Sequencing

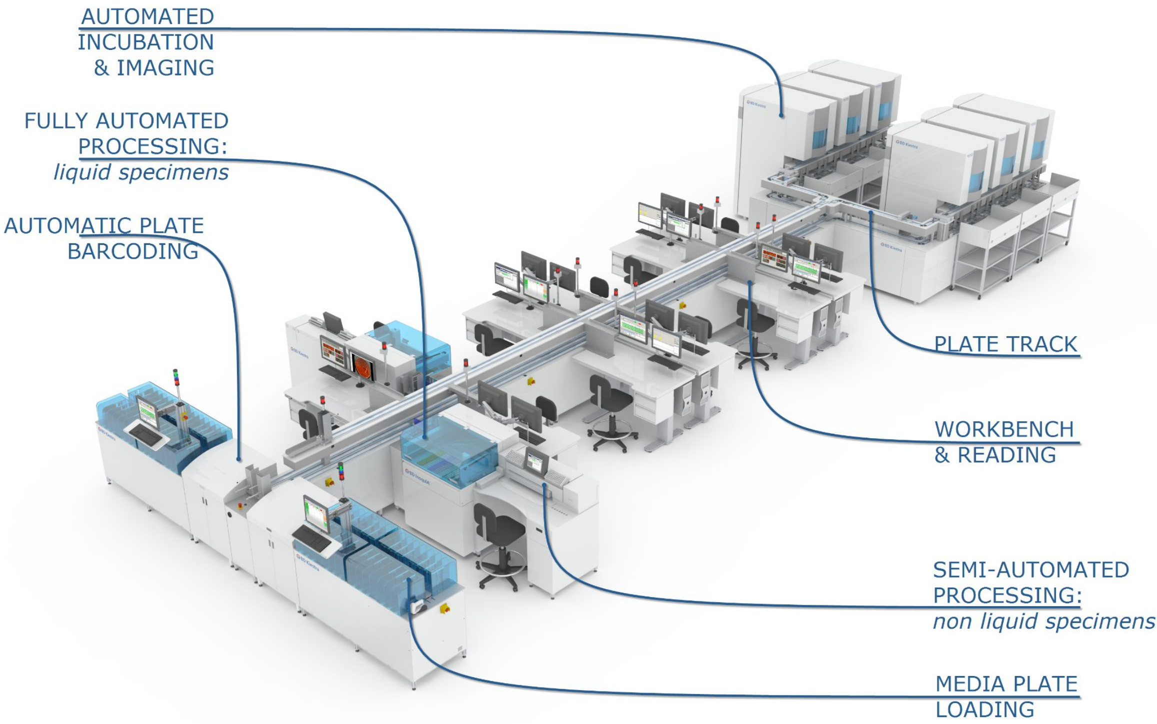
Assemble Overlapping Fragments

Extract Genome(s)

- 🟪 Pathogen
- 🟧 Host DNA
- 🟩 Other Genomes

How do we do this at scale?

Automation of labour intensive steps



AUTOMATED INCUBATION & IMAGING

FULLY AUTOMATED PROCESSING: *liquid specimens*

AUTOMATIC PLATE BARCODING

PLATE TRACK

WORKBENCH & READING

SEMI-AUTOMATED PROCESSING: *non liquid specimens*

MEDIA PLATE LOADING

https://www.bd.com/scripts/europe/labautomation/productsdrilldown.asp?CatID=455&SubID=1836&siteID=20309&d=&s=europe%2Flabautomation&sTitle=Lab+Automation&metaTitle=Total+Lab+Automation&dc=europe&dcTitle=Europe

Dr. Vermeiren

Automation of labour intensive steps

https://www.bd.com/scripts/europe/labautomation/productsdrilldown.asp?CatID=455&SubID=1836&siteID=20309&d=&s=europe%2Flabautomation&sTitle=Lab+Automation&metaTitle=Total+Lab+Automation&dc=europe&dcTitle=Europe

Sequencing technology has rapidly changed and improved

~1972-1977

First generation

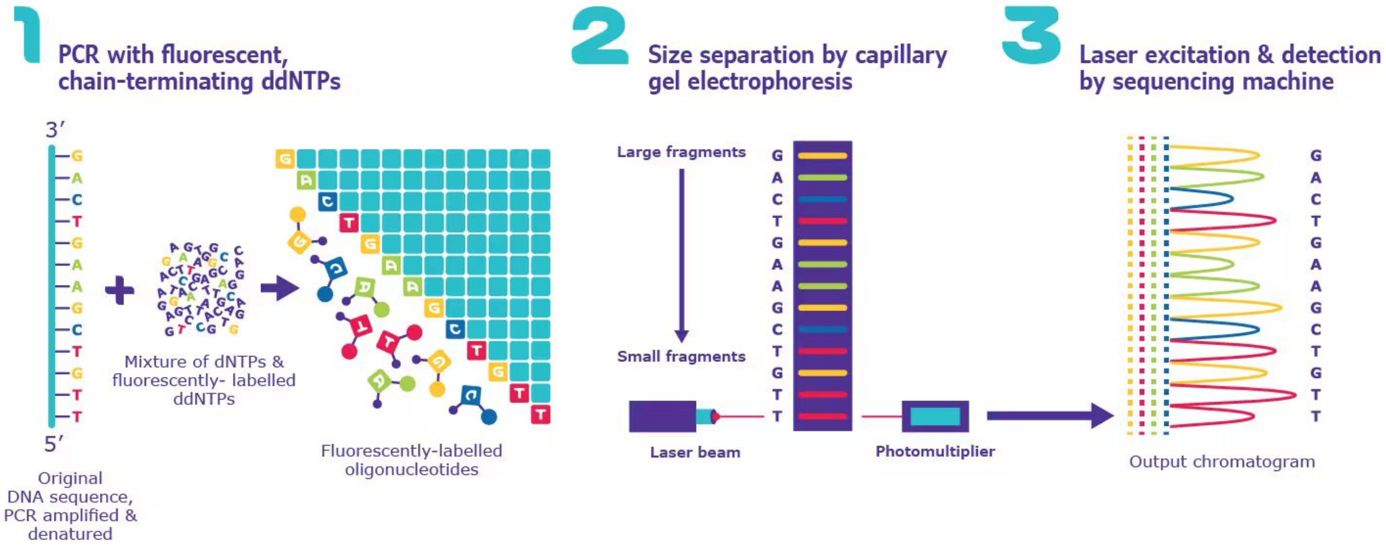
Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs,
then visualize with electrophoresis

500–1,000 bp fragments

https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

Sanger Sequencing

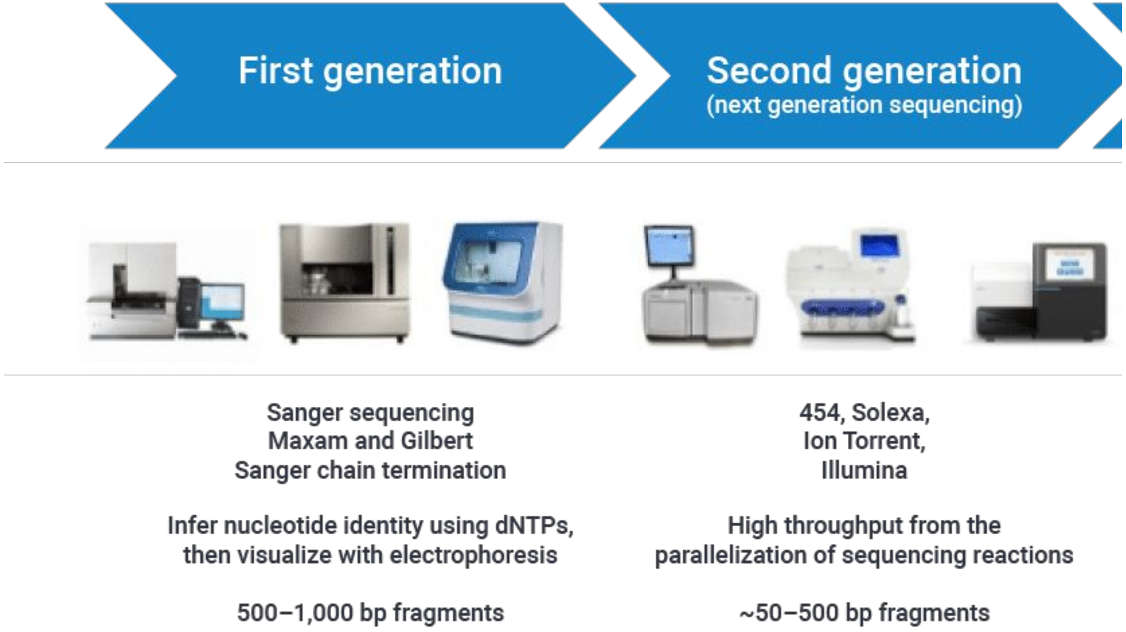


https://www.sigmaaldrich.com/CA/en/technical-documents/protocol/genomics/sequencing/sanger-sequencing

Sequencing technology has rapidly changed and improved

~1972-1977

~2001-2004



First generation

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

Second generation
(next generation sequencing)

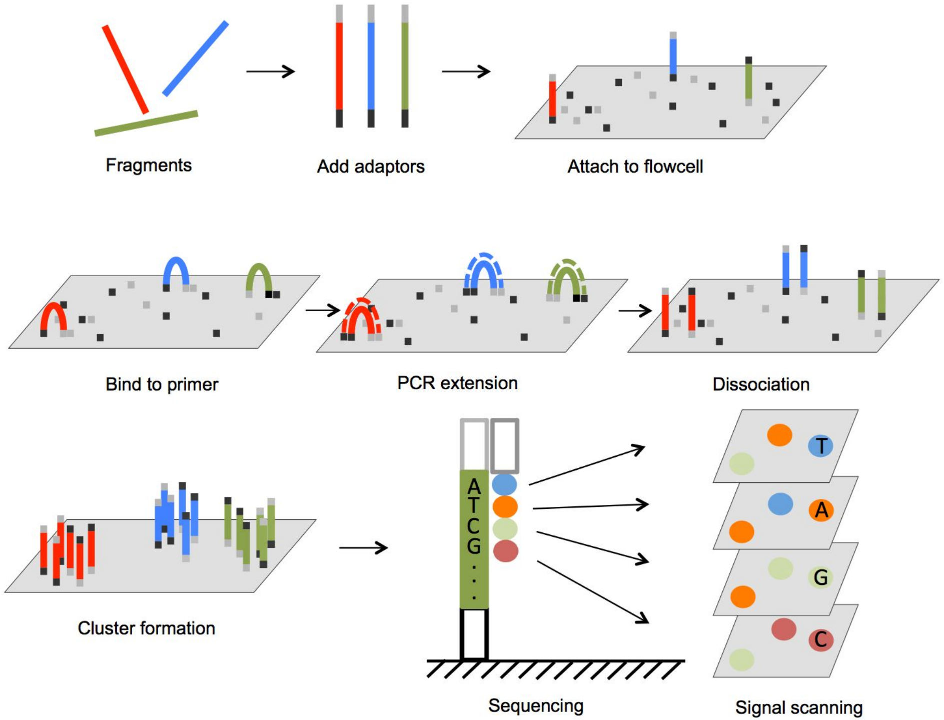
454, Solexa,
Ion Torrent,
Illumina

High throughput from the parallelization of sequencing reactions

~50–500 bp fragments

https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

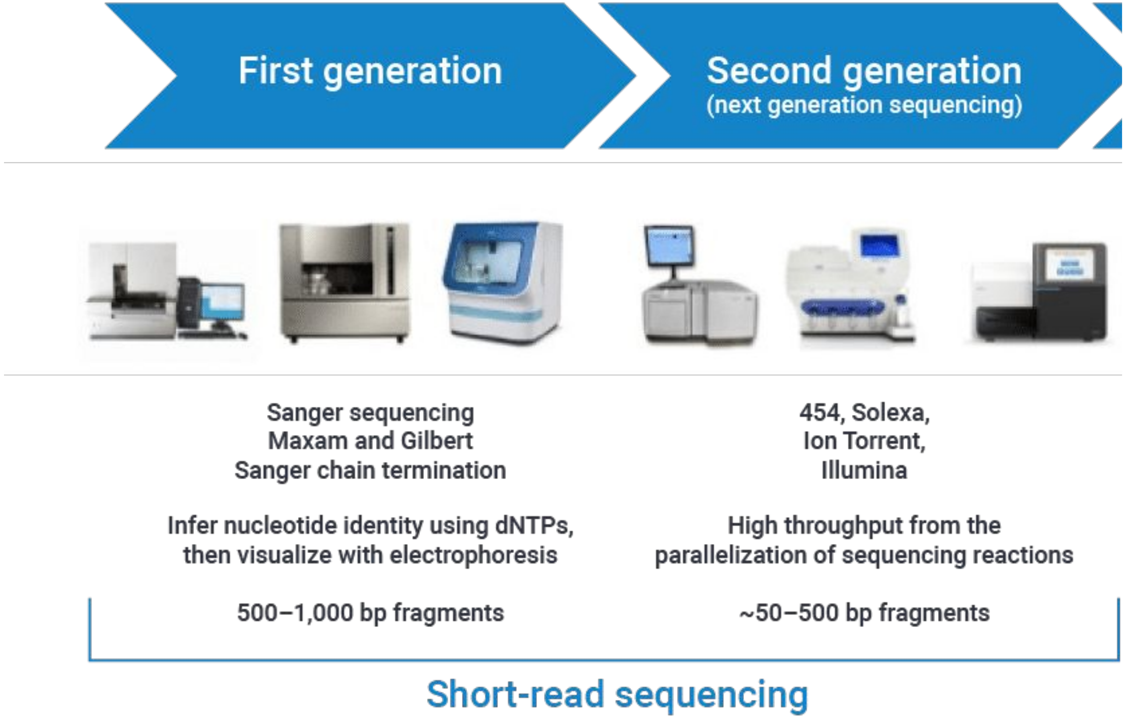
Sequencing by Synthesis



Lu, Yuan, et al. "Next generation sequencing in aquatic models." Next Generation Sequencing-Advances, Applications and Challenges 1 (2016): 13.

Sequencing technology has rapidly changed and improved

~1972-1977 ~2001-2004



First generation

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

Second generation
(next generation sequencing)

454, Solexa,
Ion Torrent,
Illumina

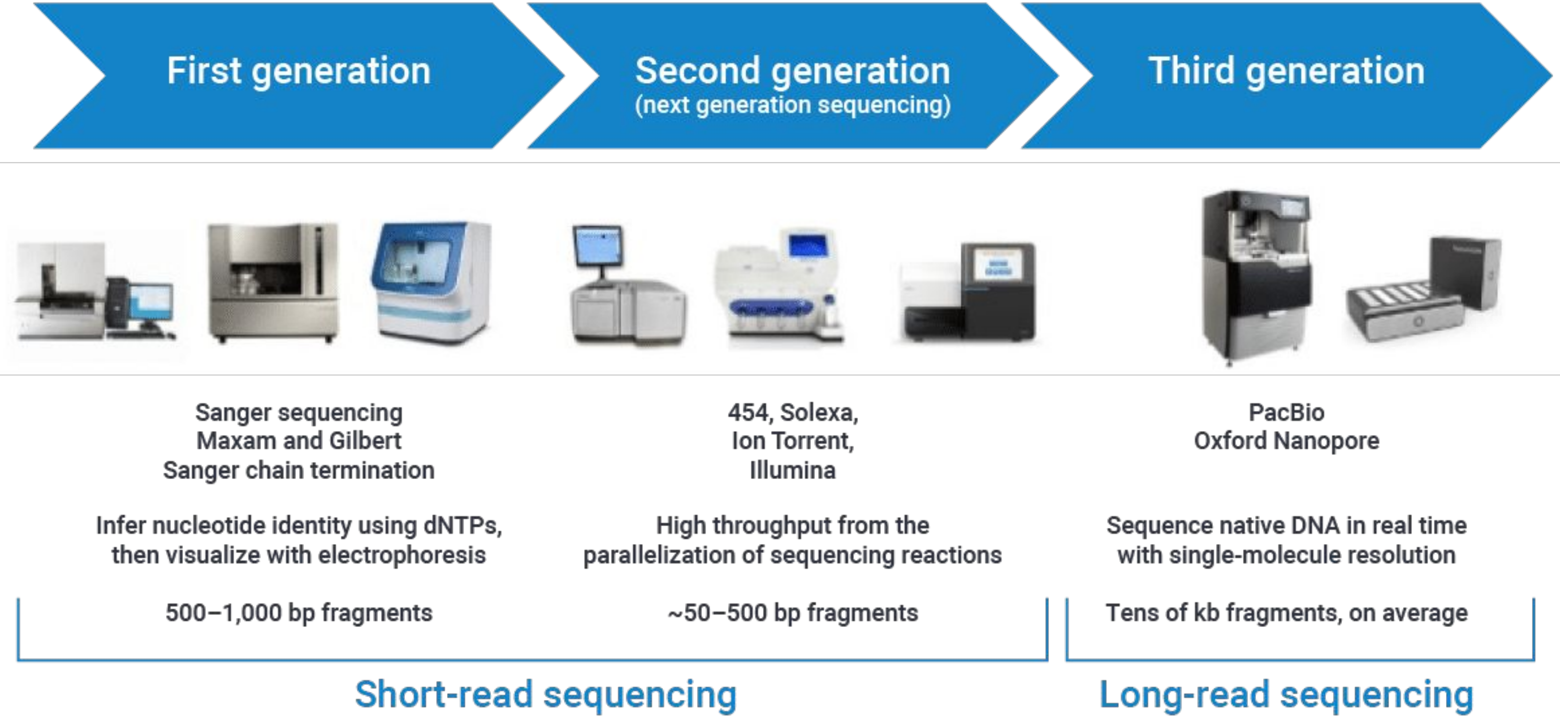
High throughput from the parallelization of sequencing reactions

~50–500 bp fragments

Short-read sequencing

https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

Sequencing technology has rapidly changed and improved



~1972-1977

First generation

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

~2001-2004

Second generation
(next generation sequencing)

454, Solexa,
Ion Torrent,
Illumina

High throughput from the parallelization of sequencing reactions

~50–500 bp fragments

~2011-2015

Third generation

PacBio
Oxford Nanopore

Sequence native DNA in real time with single-molecule resolution

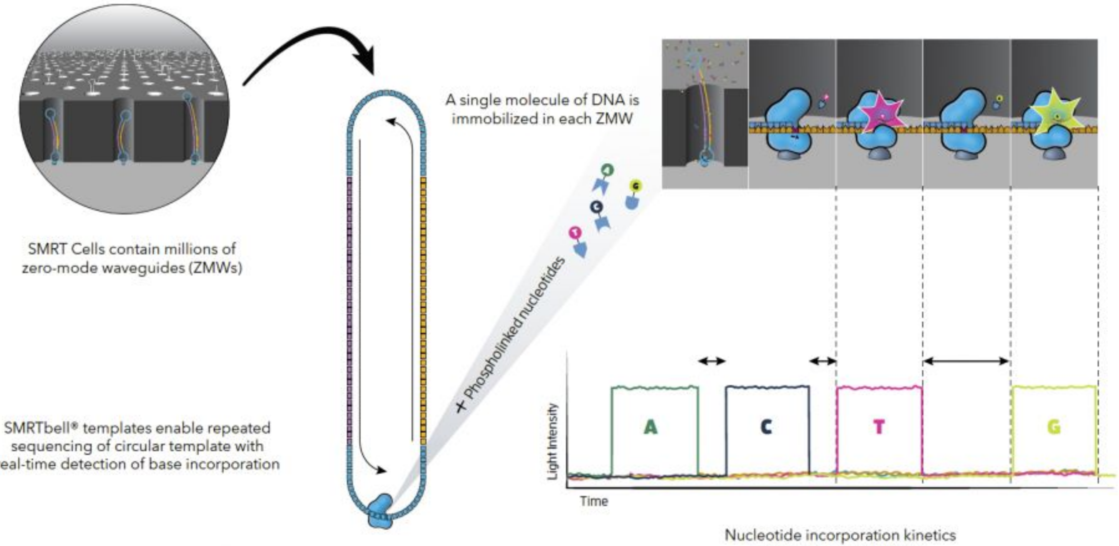
Tens of kb fragments, on average

Short-read sequencing

Long-read sequencing

https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

PacBio Sequencing



SMRT Cells contain millions of
zero-mode waveguides (ZMWs)

SMRTbell® templates enable repeated
sequencing of circular template with
real-time detection of base incorporation

A single molecule of DNA is
immobilized in each ZMW

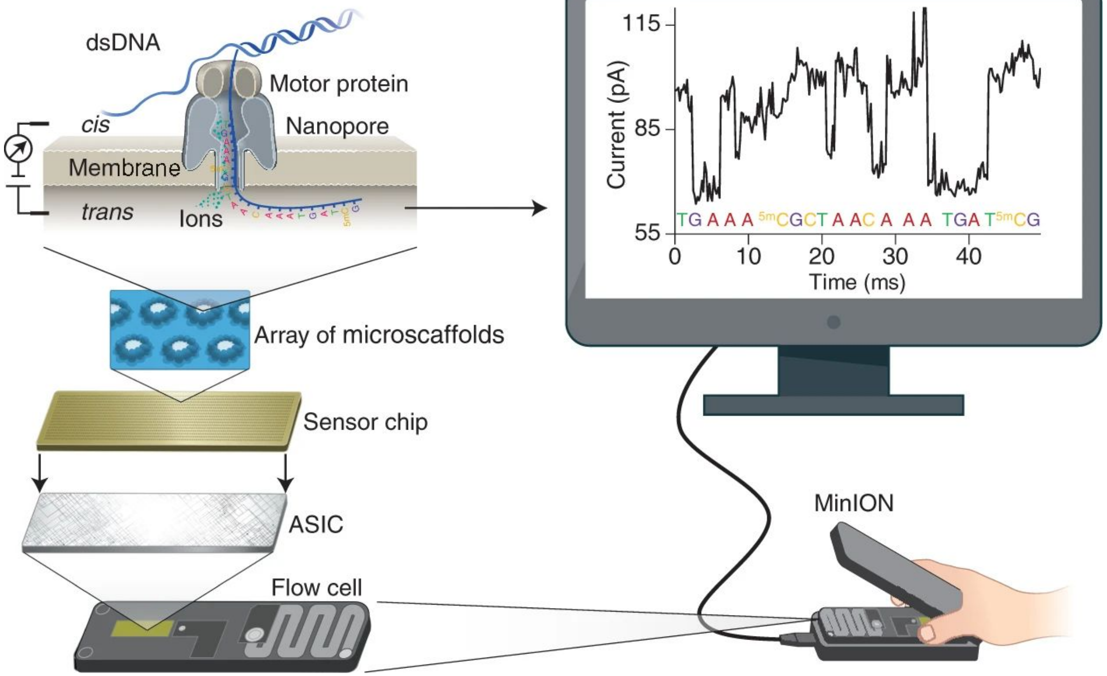
As anchored
polymerases
incorporate
labeled bases,
light is emitted

Directly detect
DNA modifications
during sequencing

Nucleotide incorporation kinetics
are measured in real time

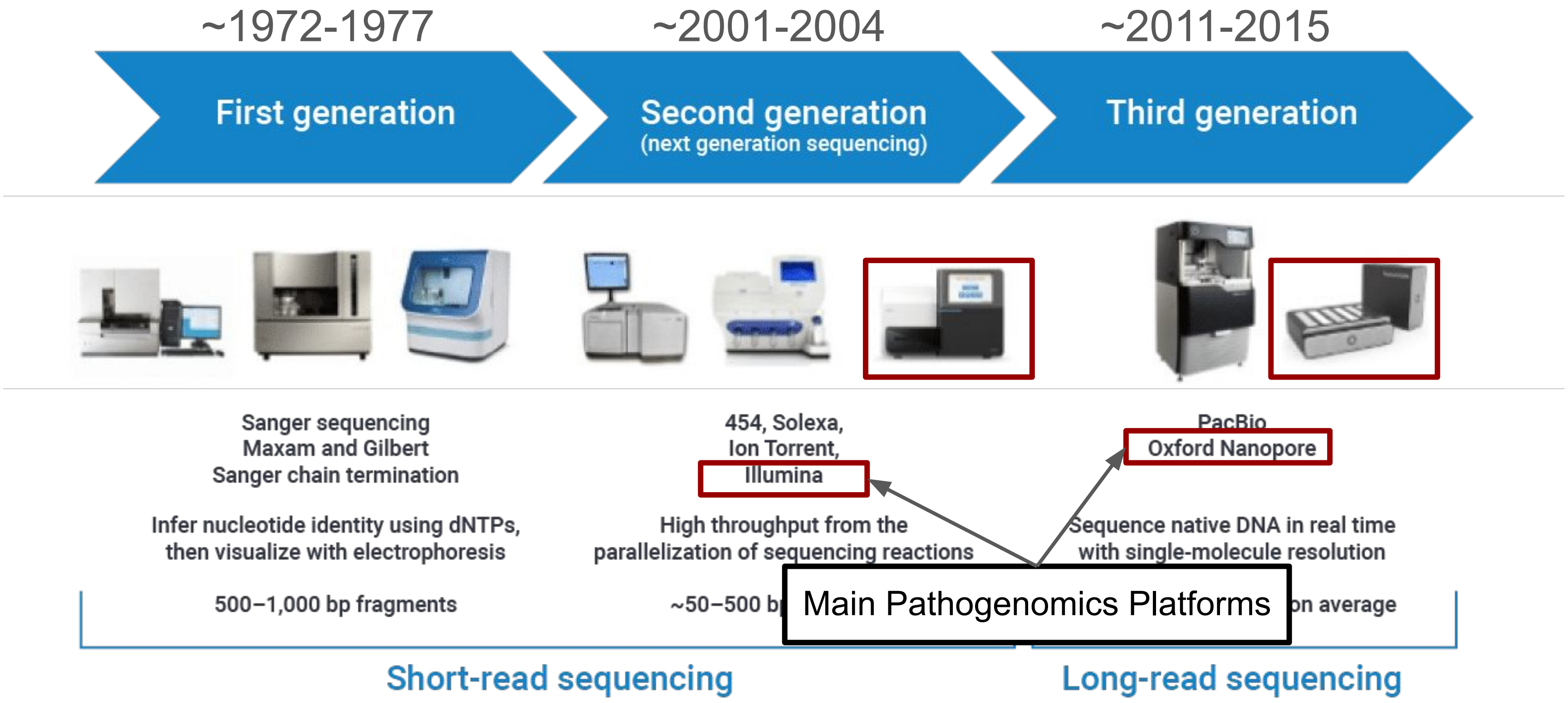
https://www.pacb.com/wp-content/uploads/SMRT-Sequencing-Brochure-Delivering-highly-accurate-long-reads-to-drive-discovery-in-life-science.pdf

Nanopore Sequencing



Wang, Yunhao, et al. "Nanopore sequencing technology, bioinformatics and applications." Nature biotechnology 39.11 (2021): 1348-1365.

Sequencing technology has rapidly changed and improved



~1972-1977

First generation

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

~2001-2004

Second generation
(next generation sequencing)

454, Solexa,
Ion Torrent,
Illumina

High throughput from the parallelization of sequencing reactions

~50–500 b

~2011-2015

Third generation

PacBio
Oxford Nanopore

Sequence native DNA in real time with single-molecule resolution

on average

Main Pathogenomics Platforms

Short-read sequencing

Long-read sequencing

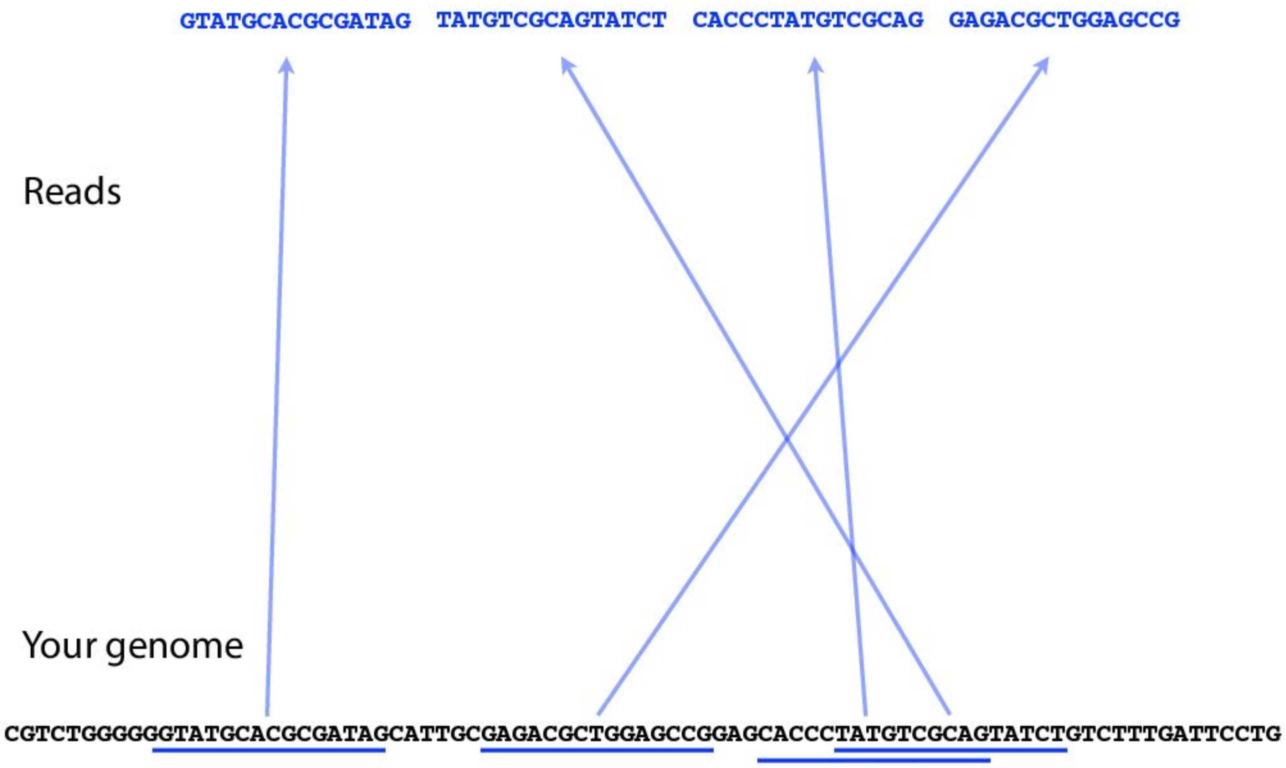
https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

Mobile sequencing lab in a suitcase

https://crain-platform-genomeweb-prod.s3.amazonaws.com/s3fs-public/styles/1200x630/public/lab_in_a_suitcase.jpeg

See lecture reading: Loman & Gardy 2017

Reads are randomly(-ish) sampled from the DNA



GTATGCACGCGATAG TATGTCGCAGTATCT CACCCTATGTCGCAG GAGACGCTGGAGCCG

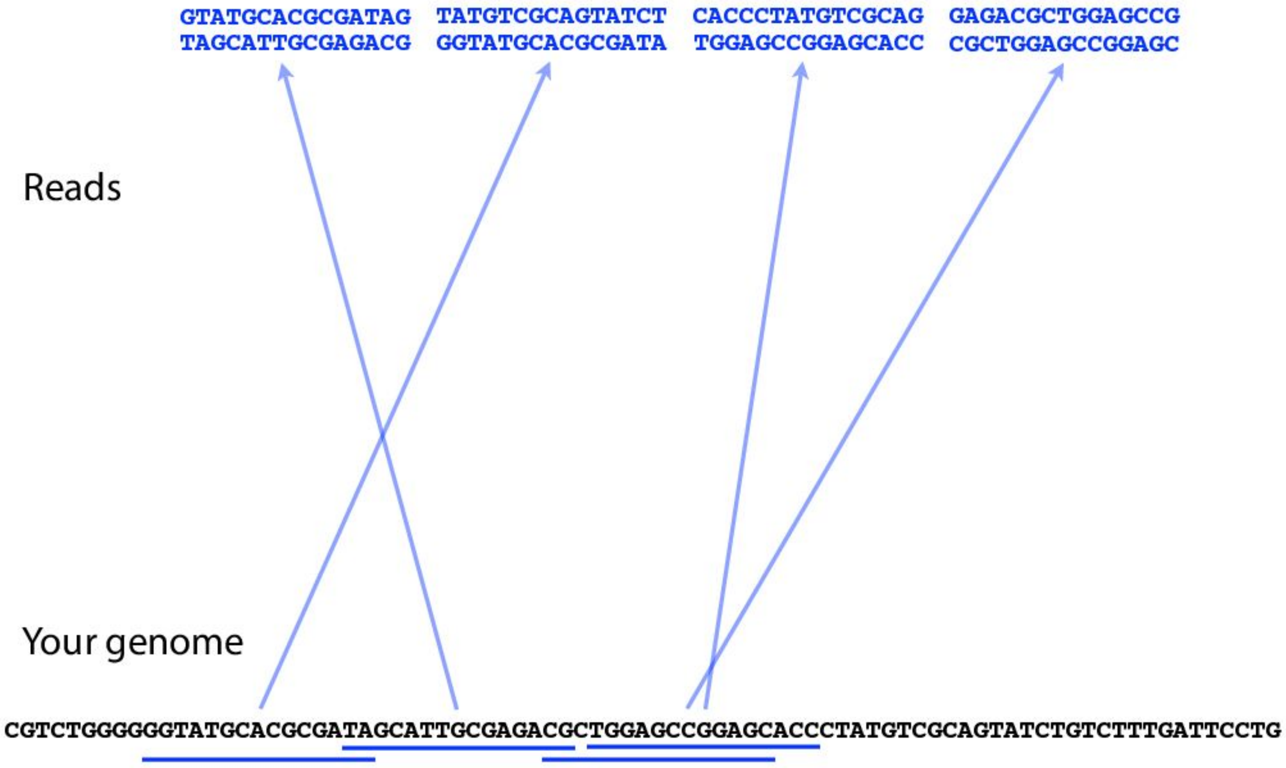
Reads

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG

www.langmead-lab.org/teaching-materials

Reads are randomly(-ish) sampled from the DNA



GTATGCACGCGATAG TATGTCGCAGTATCT CACCCTATGTCGCAG GAGACGCTGGAGCCG
TAGCATTGCGAGACG GGTATGCACGCGATA TGGAGCCGGAGCACC CGCTGGAGCCGGAGC

Reads

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG

www.langmead-lab.org/teaching-materials

Reads are randomly(-ish) sampled from the DNA

GTATGCACGCGATAG TATGTCGCAGTATCT CACCCTATGTCGCAG GAGACGCTGGAGCCG
TAGCATTGCGAGACG GGTATGCACGCGATA TGGAGCCGGAGCACC CGCTGGAGCCGGAGC
TGTCTTTGATTCCTG CGCGATAGCATTGCG GCATTGCGAGACGCT CCTATGTCGCAGTAT

Reads

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG

www.langmead-lab.org/teaching-materials

Reads are randomly(-ish) sampled from the DNA

Reads

GTATGCACGCGATAG
TAGCATTGCGAGACG
TGTCTTTGATTCCTG
GACGCTGGAGCCGGA
TATCGCACCTACGTT
CACGGGAGCTCTCCA
GTATGCACGCGATAG
GCGAGACGCTGGAGC
CCTACGTTCAATATT
GACGCTGGAGCCGGA
TATCGCACCTACGTT
CACGGGAGCTCTCCA

TATGTCGCAGTATCT
GGTATGCACGCGATA
CGCGATAGCATTGCG
GCACCCTATGTCGCA
CAATATTCGATCATG
TGCATTTGGTATTTT
ACCTACGTTCAATAT
CTATCACCCTATTAA
GCACCTACGTTCAAT
GCACCCTATGTCGCA
CAATATTCGATCATG
TGCATTTGGTATTTT

CACCCTATGTCGCAG
TGGAGCCGGAGCACC
GCATTGCGAGACGCT
GTATCTGTCTTTGAT
GATCACAGGTCTATC
CGTCTGGGGGGTATG
TATTTATCGCACCTA
CTGTCTTTGATTCCT
GTCTGGGGGGTATGC
GTATCTGTCTTTGAT
GATCACAGGTCTATC
CGTCTGGGGGGTATG

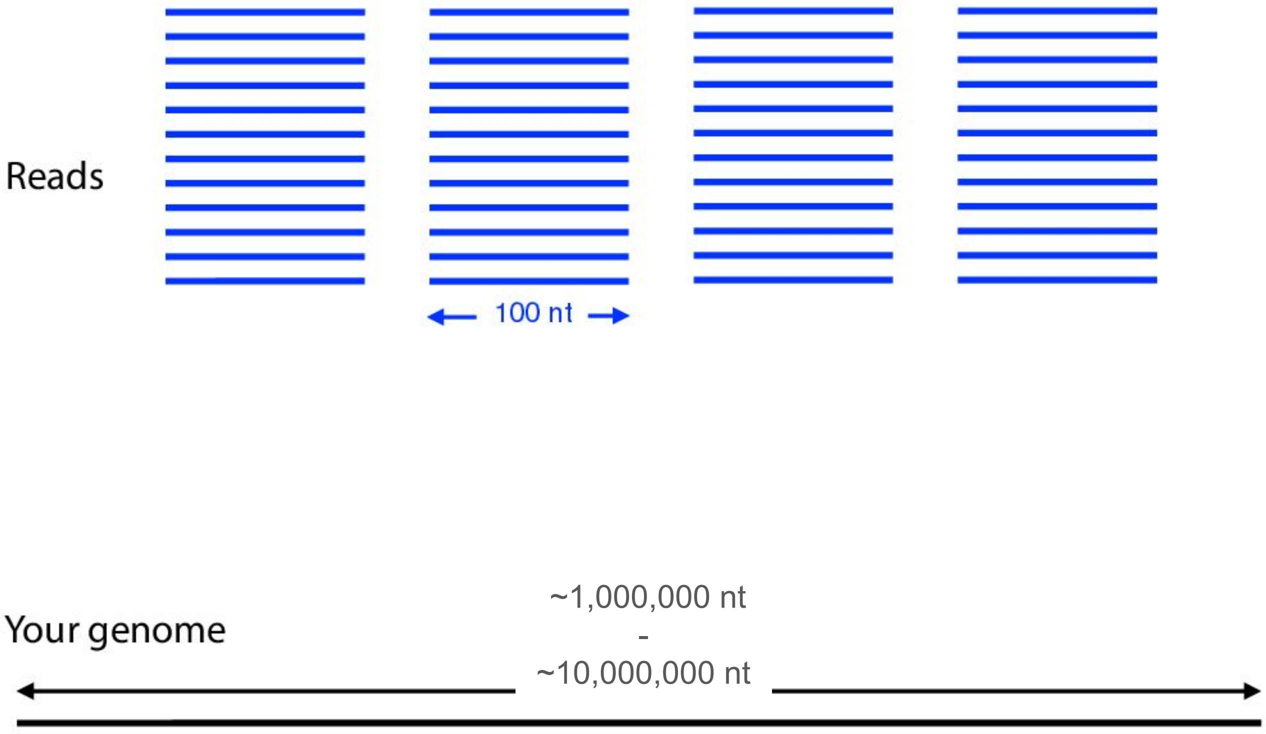
GAGACGCTGGAGCCG
CGCTGGAGCCGGAGC
CCTATGTCGCAGTAT
CCTCATCCTATTATT
ACCCTATTAACCACT
CACGCGATAGCATTG
CCACTCACGGGAGCT
ACTCACGGGAGCTCT
AGCCGGAGCACCCTA
CCTCATCCTATTATT
ACCCTATTAACCACT
CACGCGATAGCATTG

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG

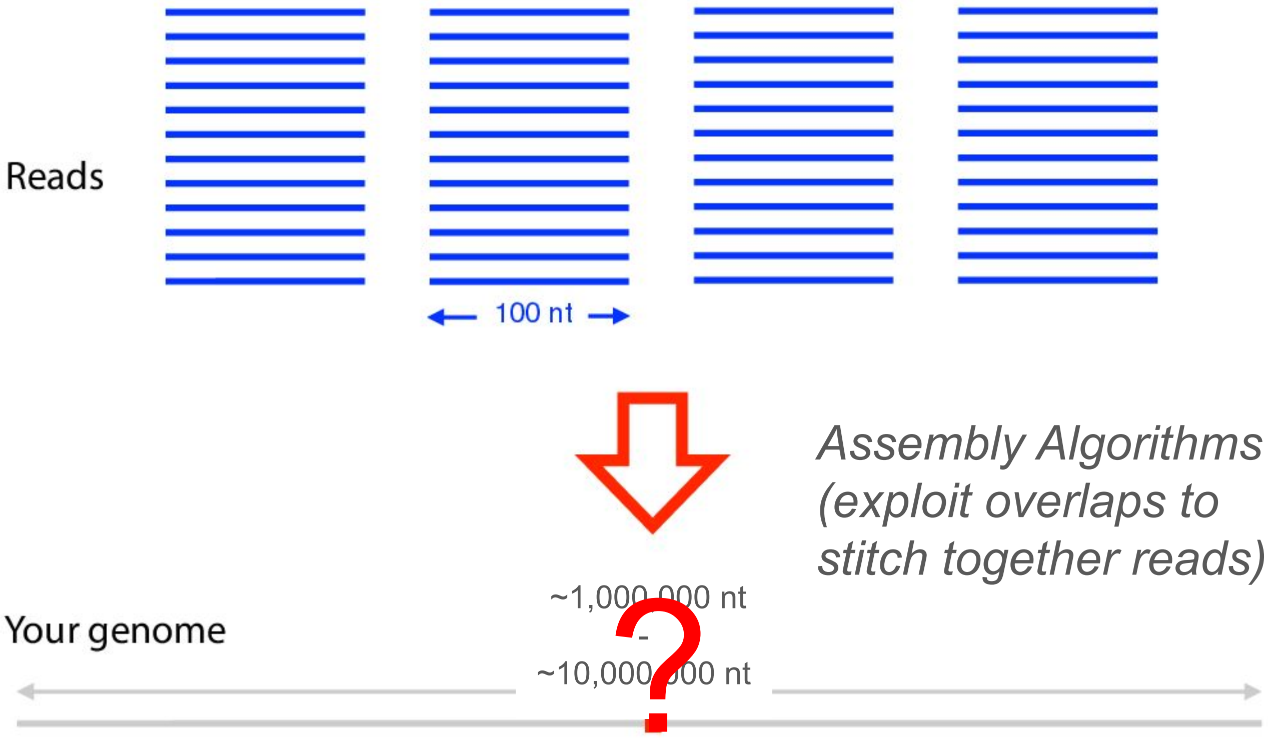
www.langmead-lab.org/teaching-materials

Reads are randomly(-ish) sampled from the DNA



www.langmead-lab.org/teaching-materials

Reads are randomly(-ish) sampled from the DNA



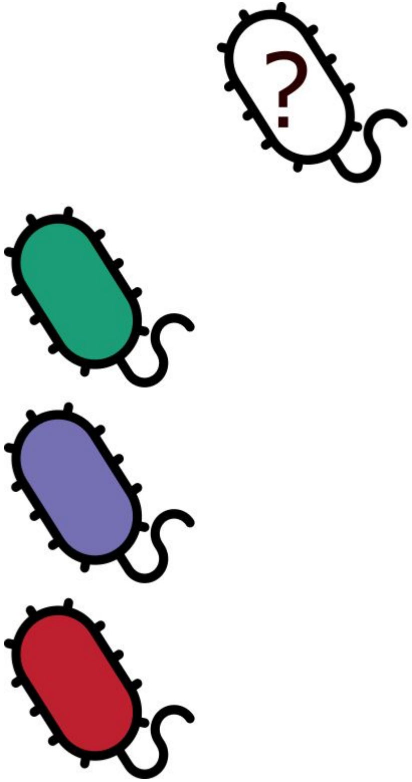
Assembly Algorithms (exploit overlaps to stitch together reads)

www.langmead-lab.org/teaching-materials

Got a genome, now what?

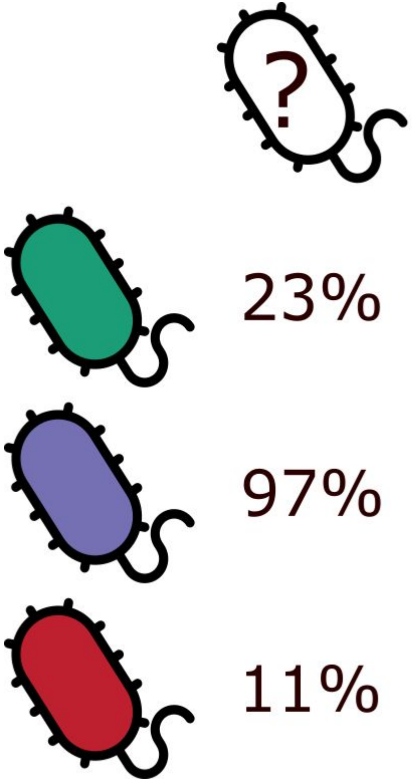
Genomic Diagnostics: What is the pathogen?

Genomic Diagnostics: What is the pathogen?



- Compare to genomes in database from known organisms

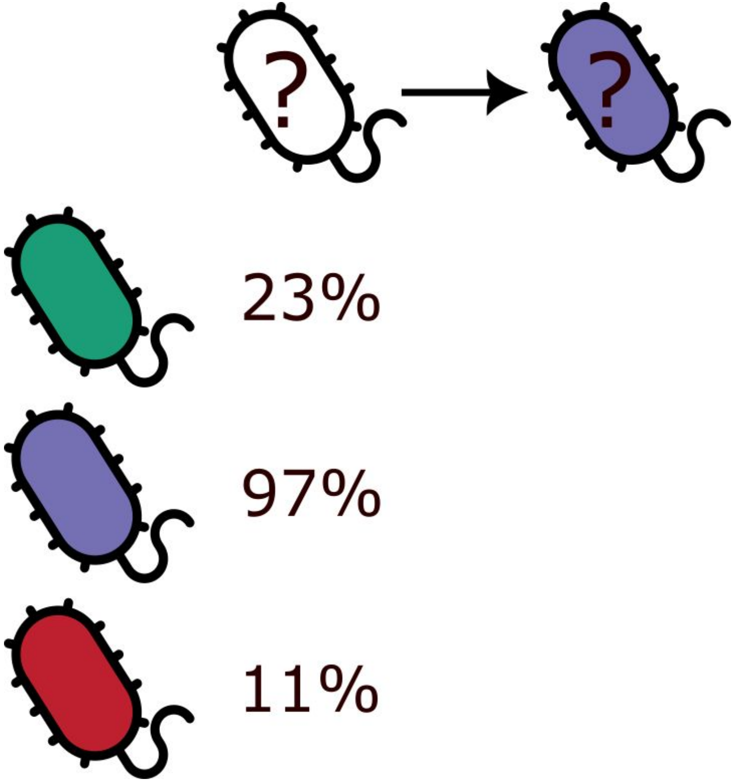
Genomic Diagnostics: What is the pathogen?



- Compare to genomes in database from known organisms
- Average Nucleotide Identity (ANI) is an example of a similarity metric

$$\mathrm{gANI}_{G1 \rightarrow G2} = \frac{\sum_{bbh} (\text{Percent Identity} * \text{Alignment length})}{\text{lengths of BBH genes}}$$

Genomic Diagnostics: What is the pathogen?



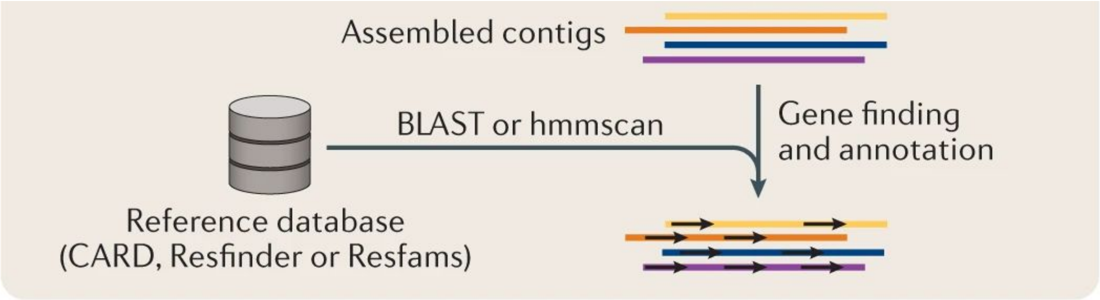
- Compare to genomes in database from known organisms
- Average Nucleotide Identity (ANI) is an example of a similarity metric

$$\mathrm{gANI}_{G1 \rightarrow G2} = \frac{\sum_{bbh}(\text{Percent Identity} * \text{Alignment length})}{\text{lengths of BBH genes}}$$

- Identify pathogen as closest reference genome taxa
- Use identity to drive treatment (if x then treat by y)

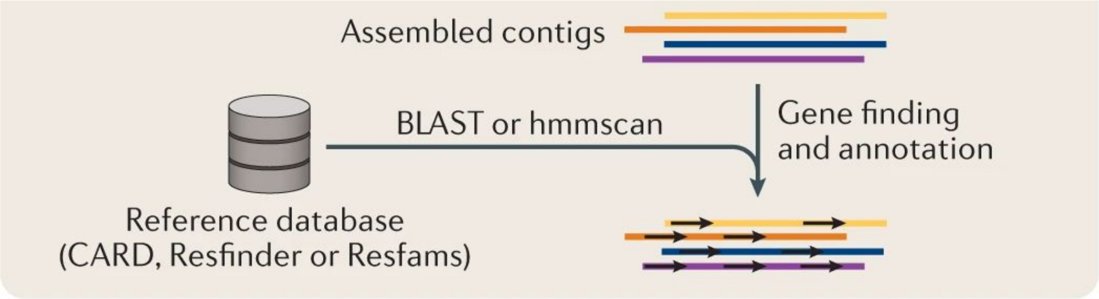
- Typing for outbreak investigation linkage

Genomic Diagnostics: What drugs will work?



10.1038/s41576-019-0108-4

Genomic Diagnostics: What drugs will work?

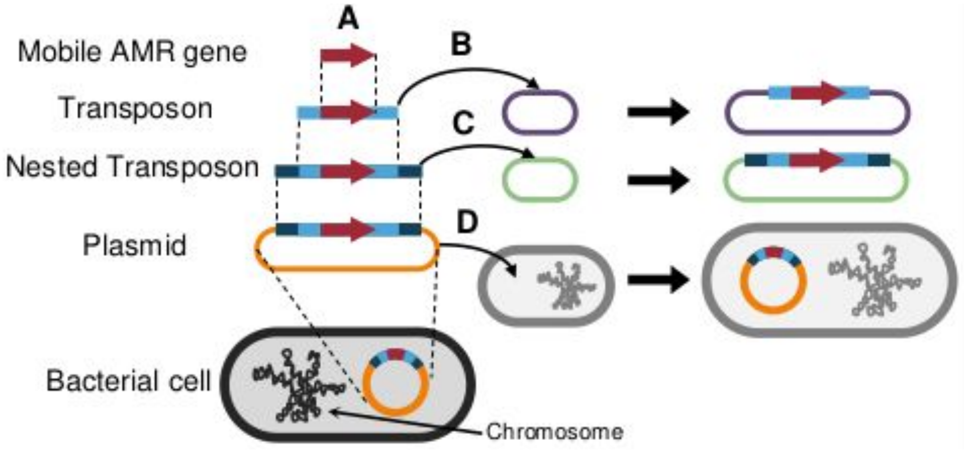


10.1038/s41576-019-0108-4

- Detect NDM-1 carbapenemase gene
- Pathogen protein that destroys many antibiotics (beta-lactams)
- => Treat with alternative class of antibiotics (e.g., colistin)

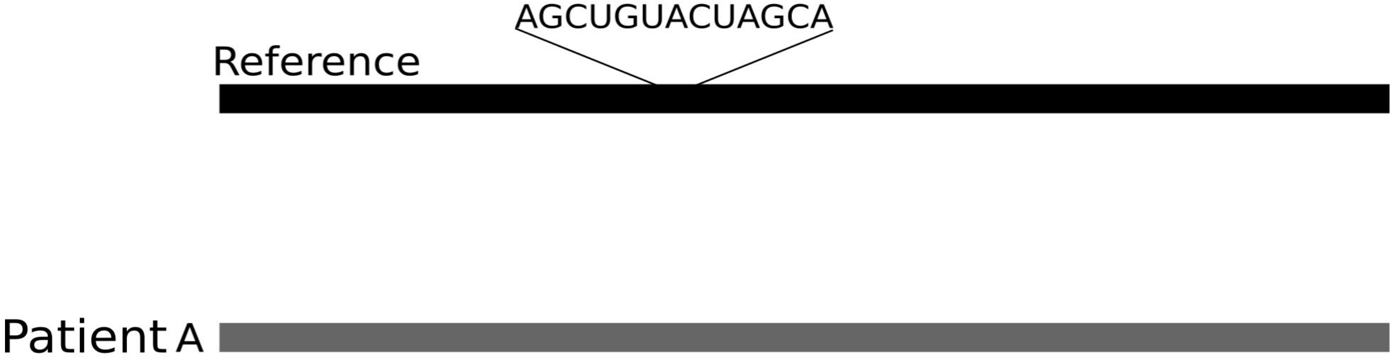
- Hours vs weeks for some pathogens (TB)

Genomic Diagnostics: identifying mobile genetic elements

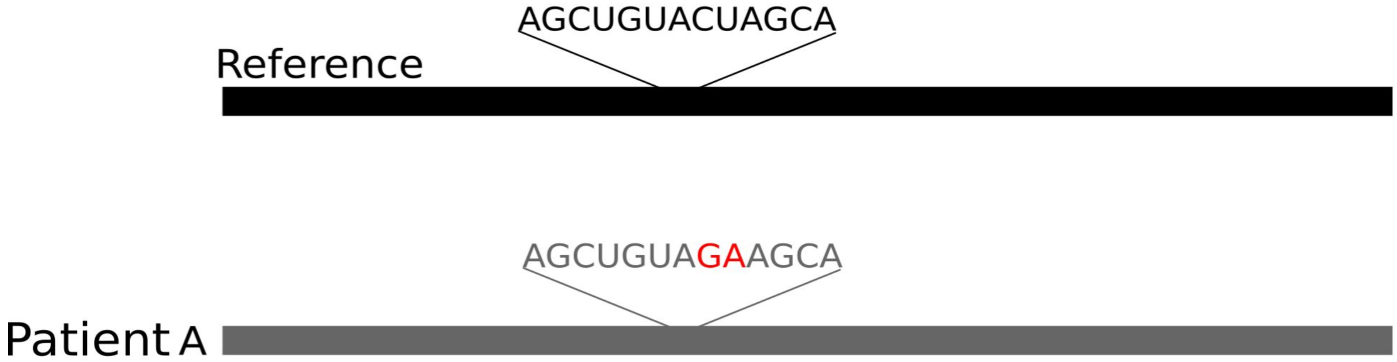


OK, but what else can we do with genomes?

Detection mutations relative to reference



Detection mutations relative to reference



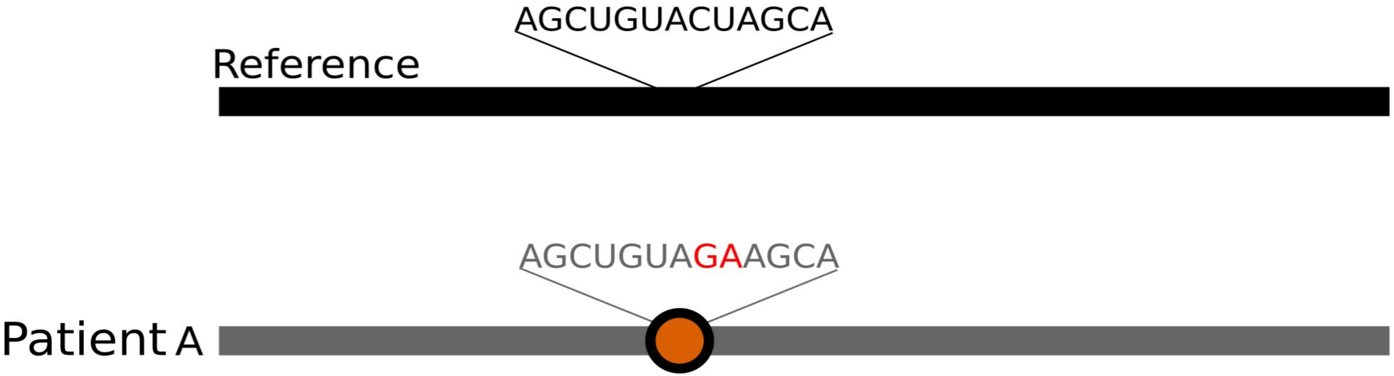
AGCUGUACUAGCA

Reference

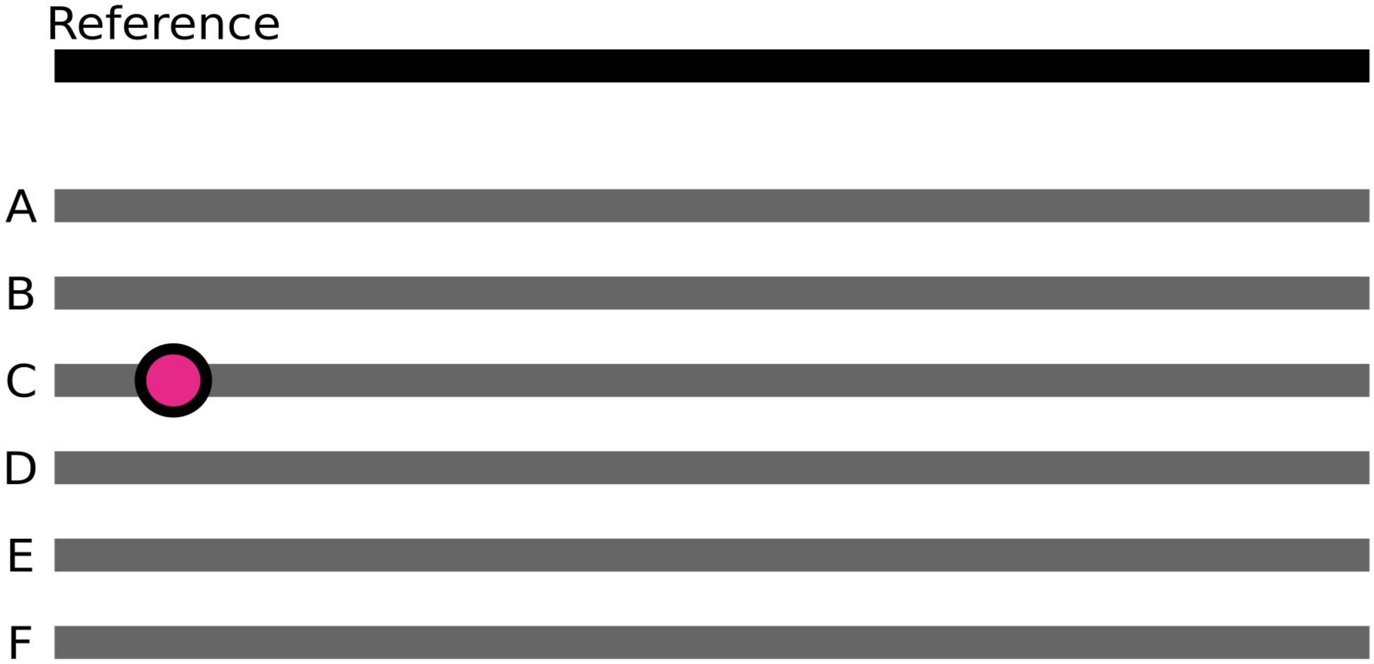
AGCUGUAGAAGCA

Patient A

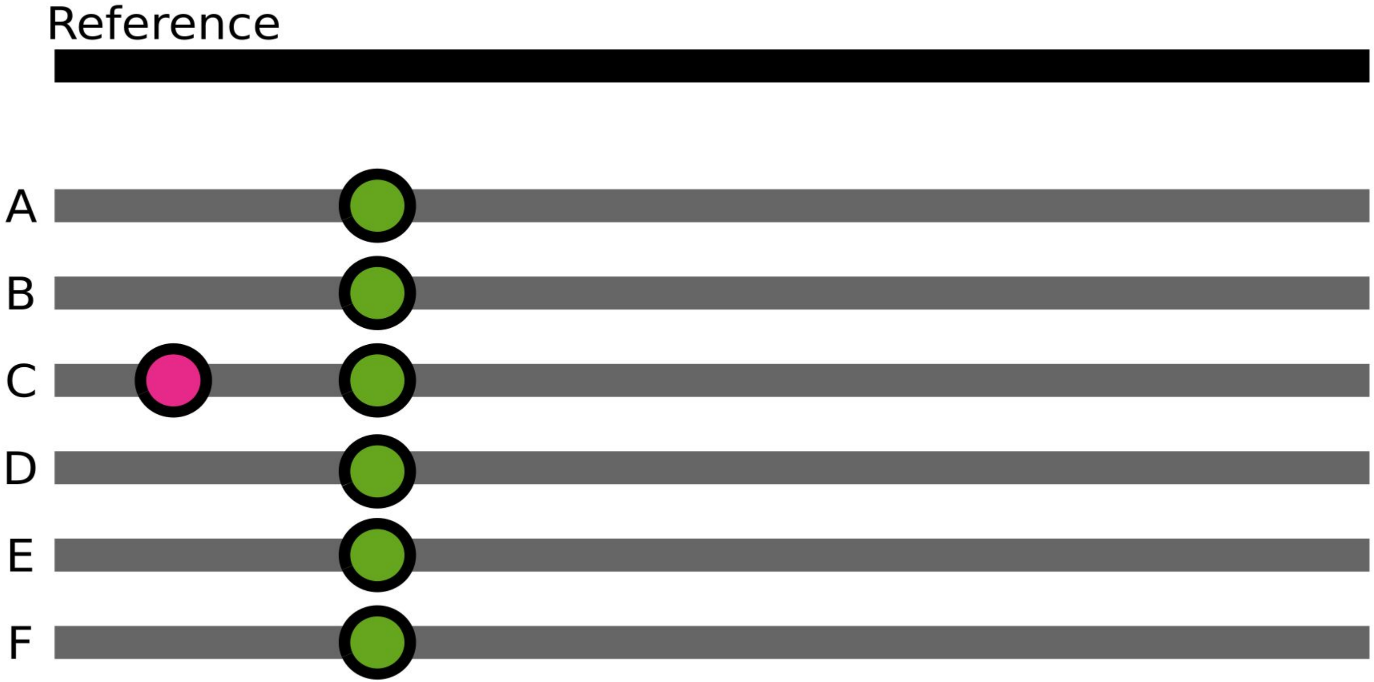
Detection mutations relative to reference



Compare mutations across patients

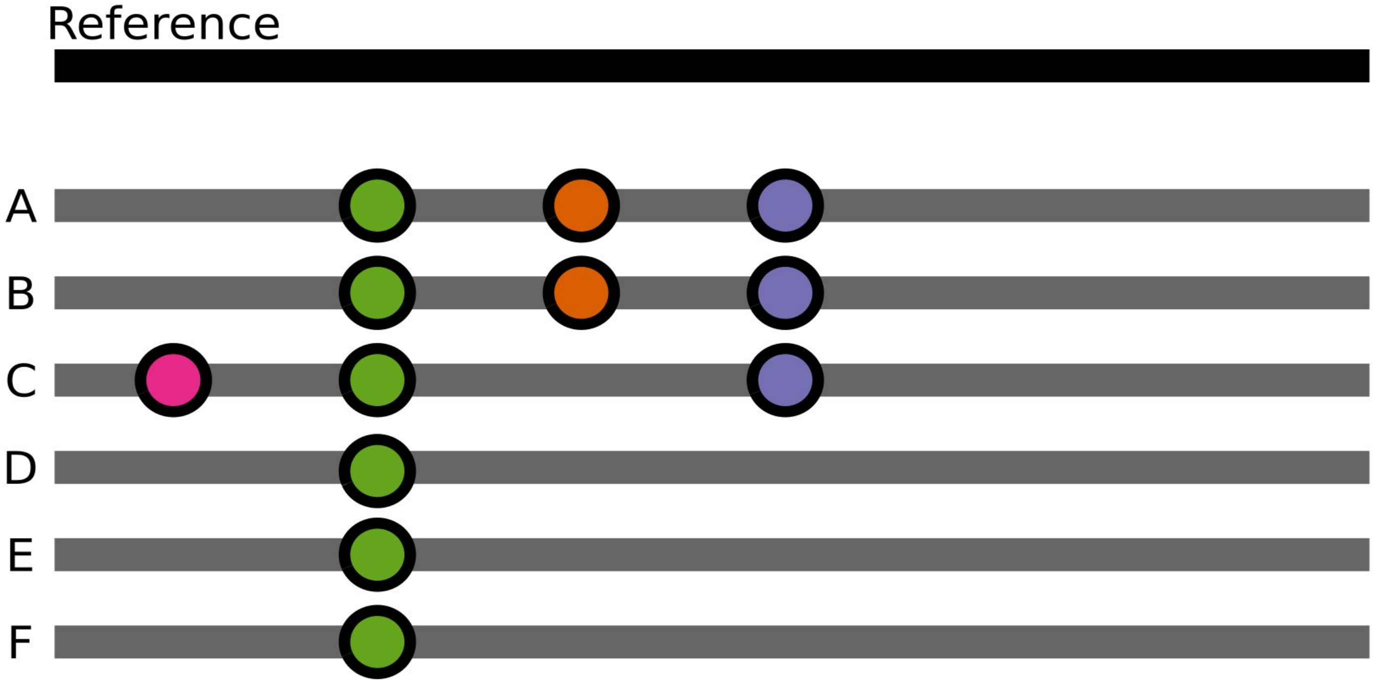


Compare mutations across patients

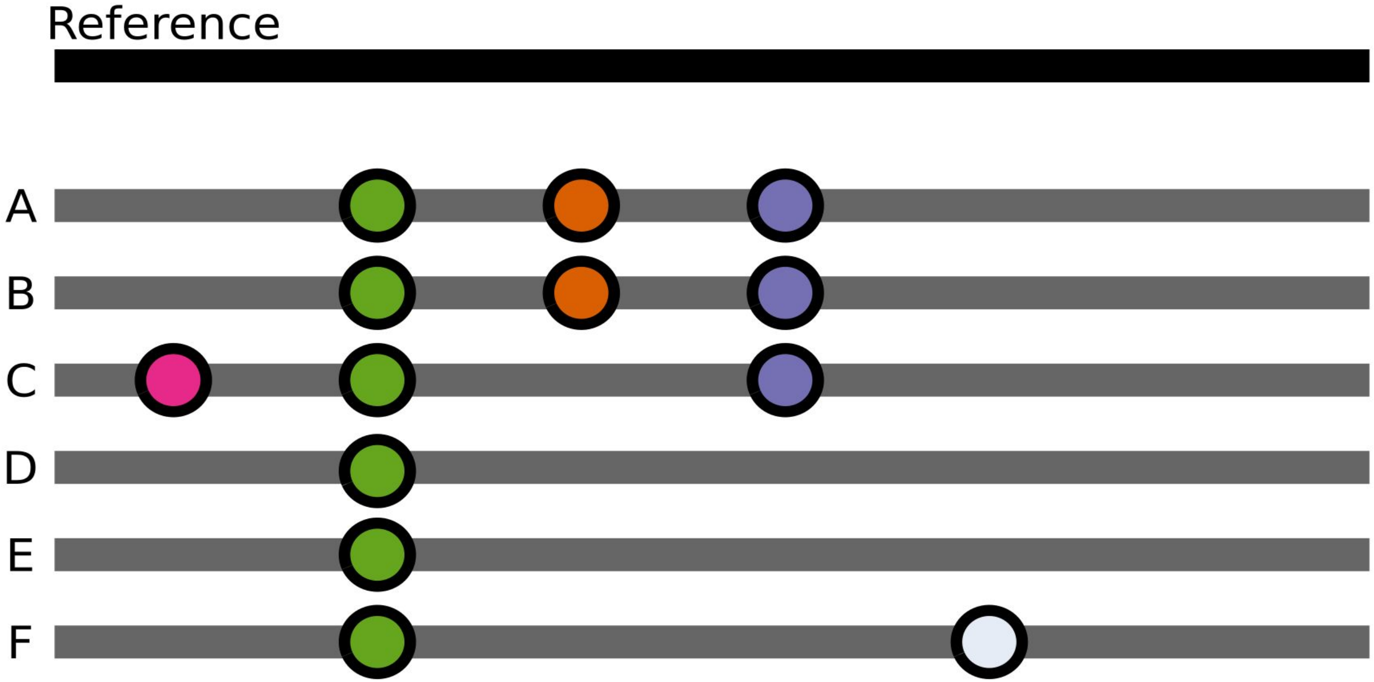


Compare mutations across patients

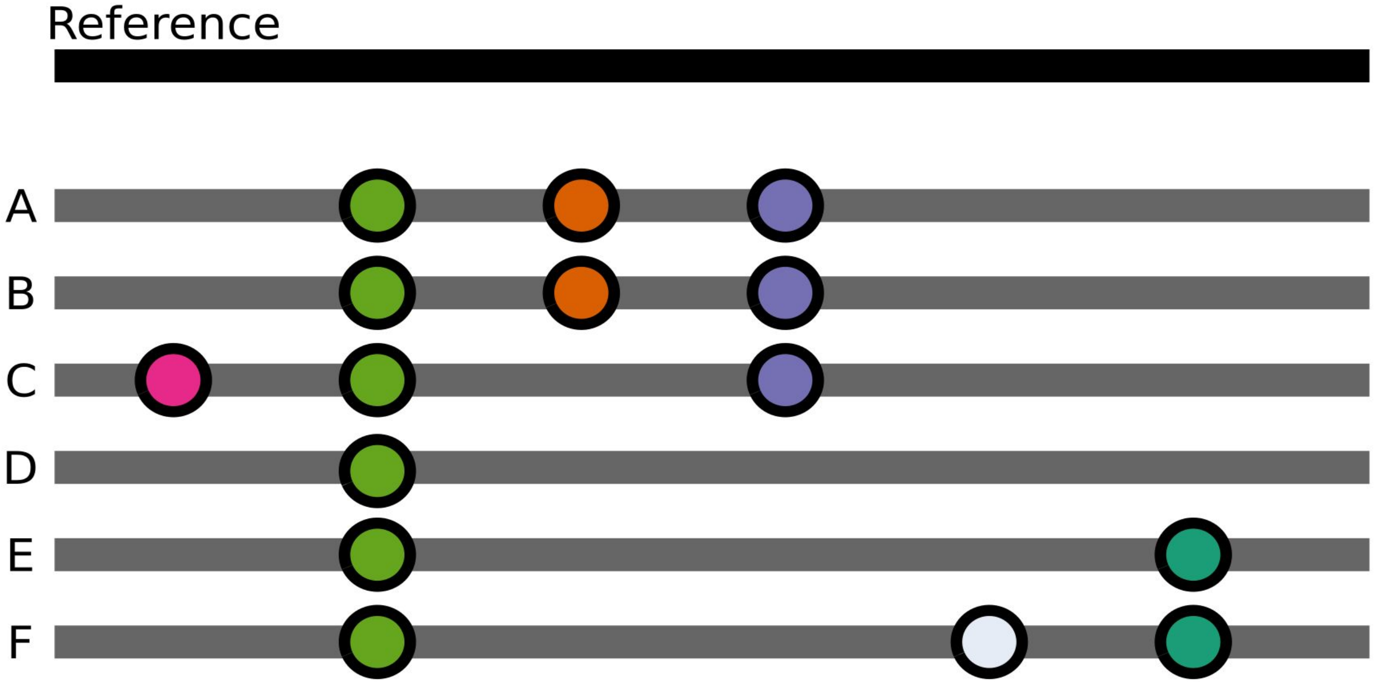
Compare mutations across patients



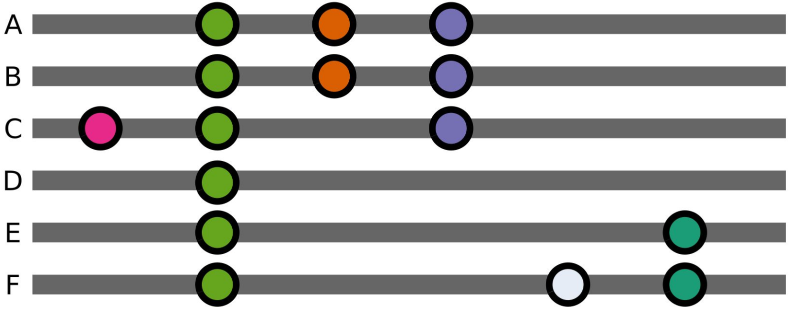
Compare mutations across patients



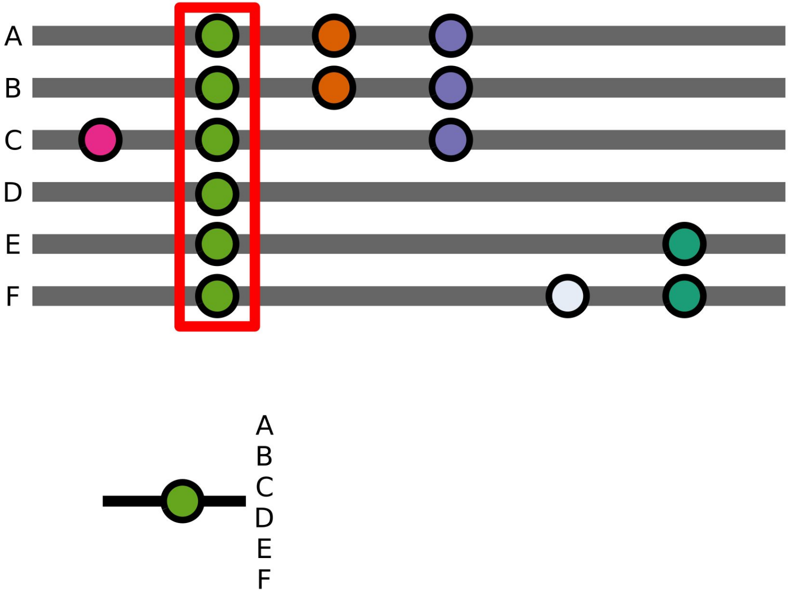
Compare mutations across patients



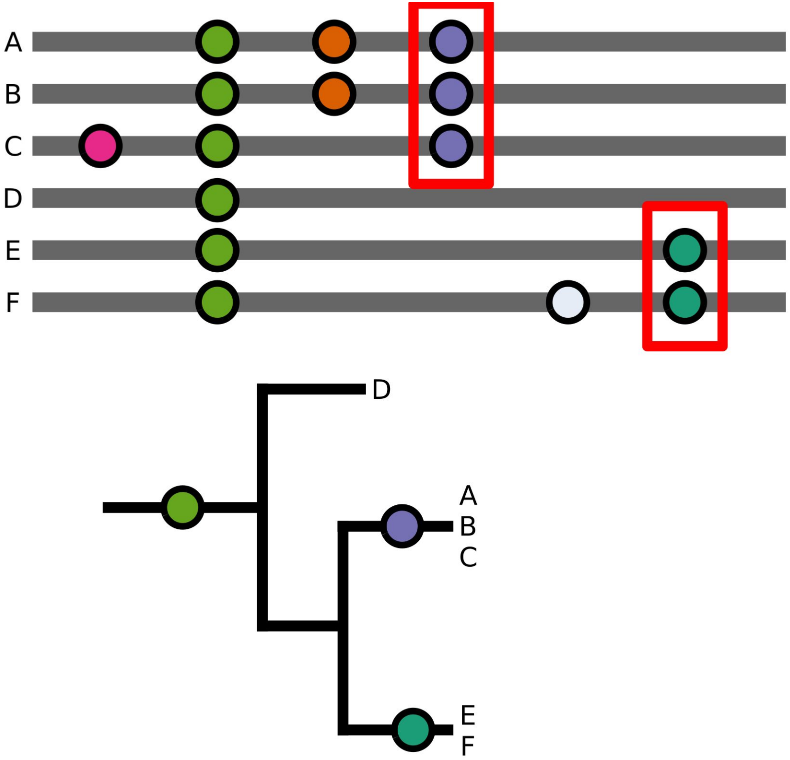
Using pattern of mutations to infer relationships



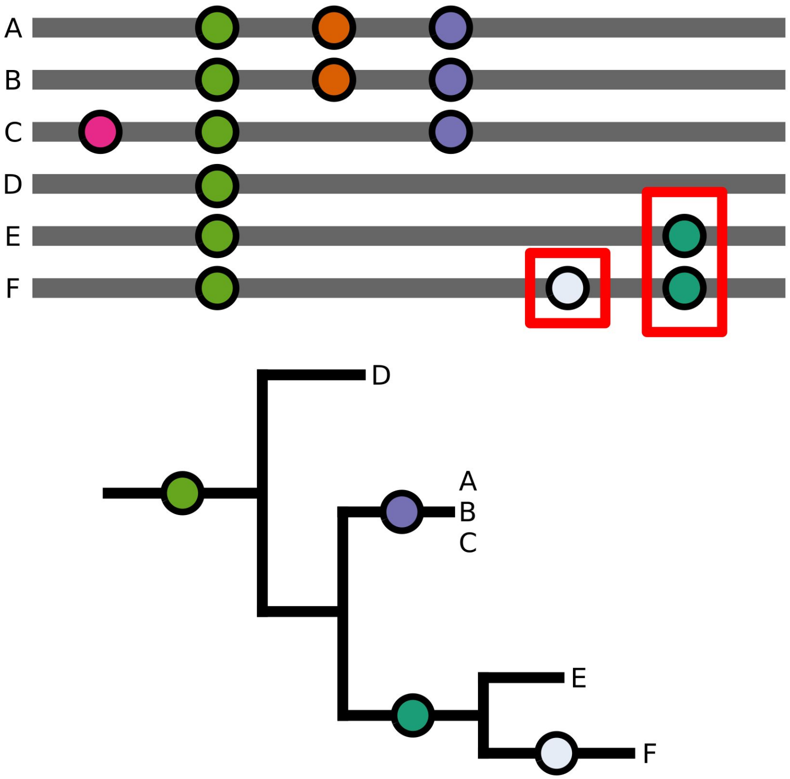
Using pattern of mutations to infer relationships



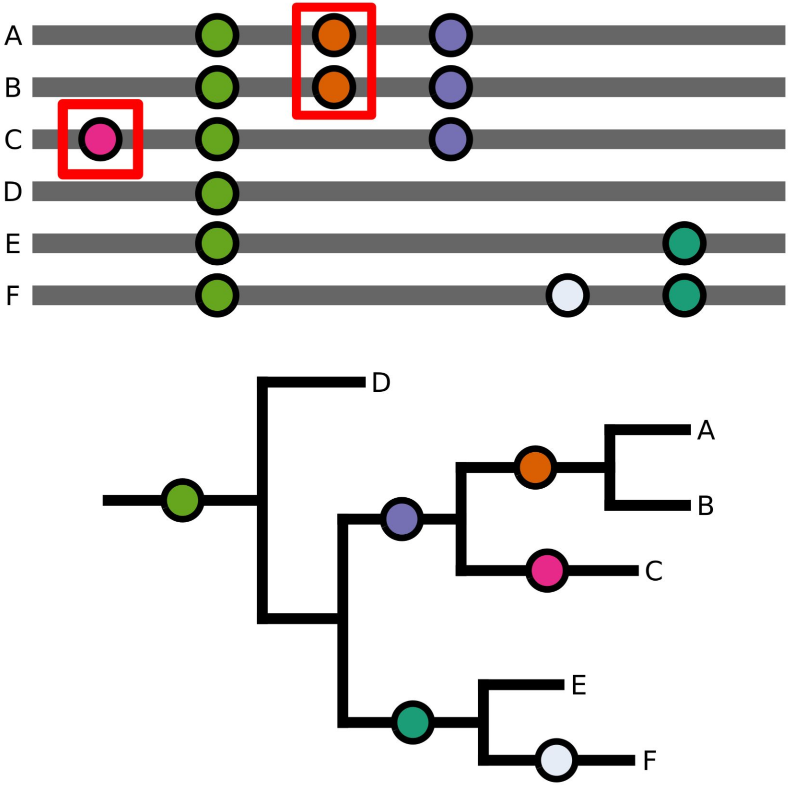
Using pattern of mutations to infer relationships



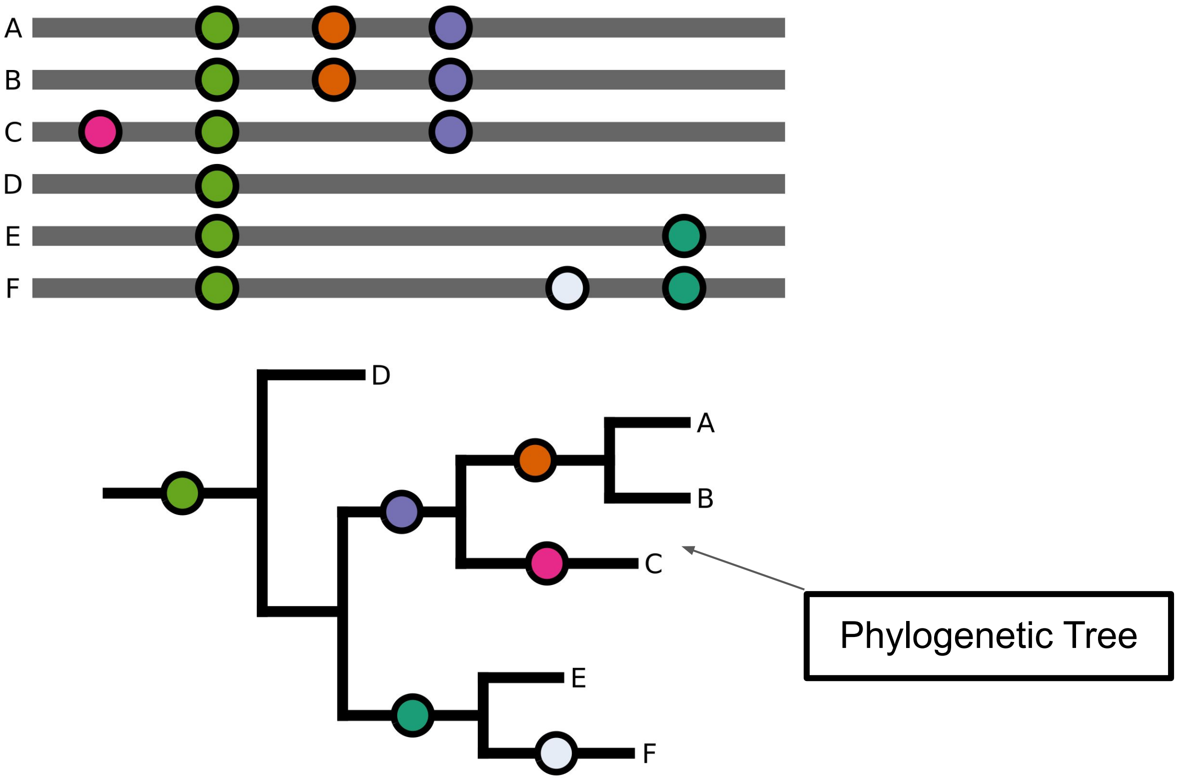
Using pattern of mutations to infer relationships



Using pattern of mutations to infer relationships



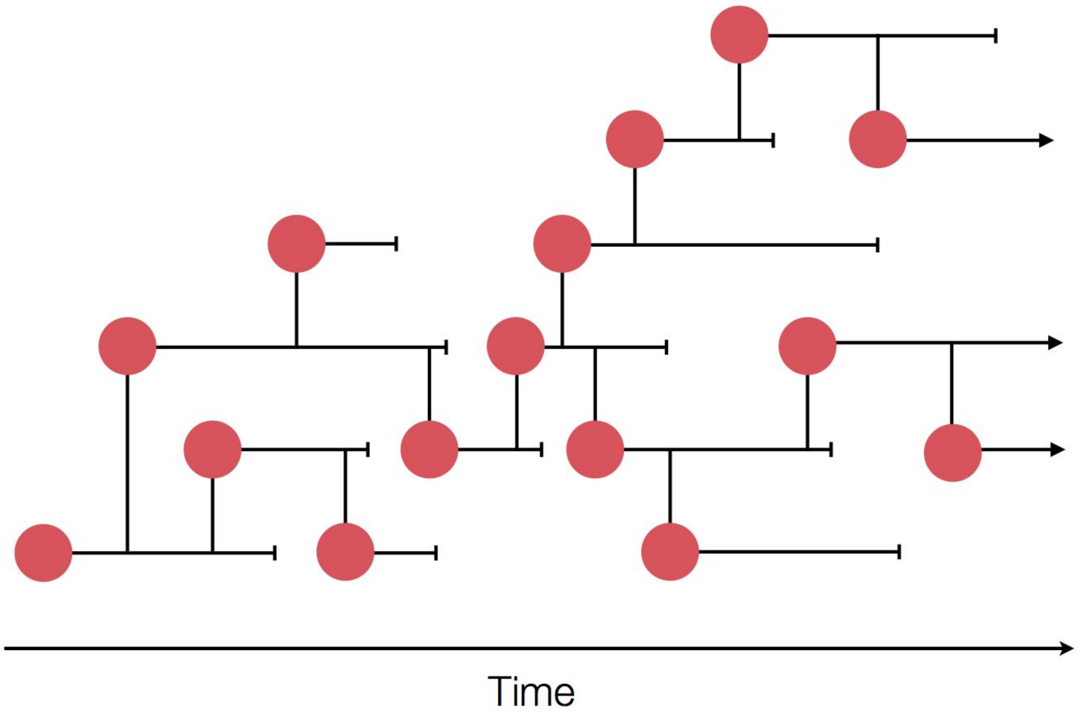
Using pattern of mutations to infer relationships



Phylogenetic Tree

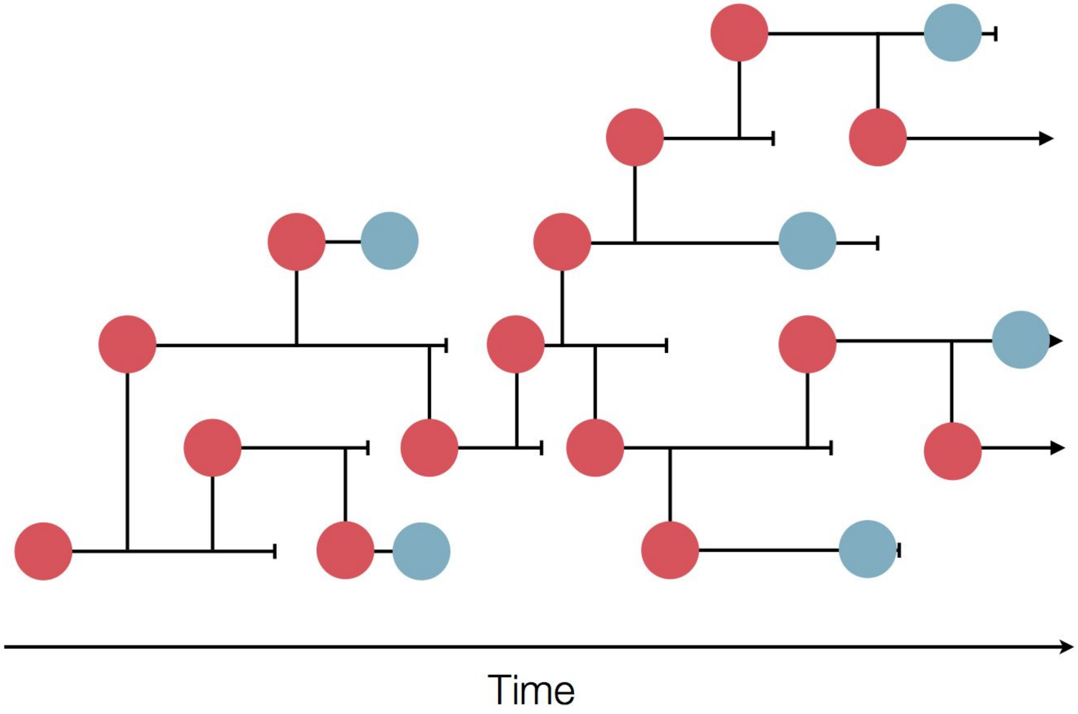
What does this tree actually represent?

Sampling & partially reconstructing underlying epidemic process



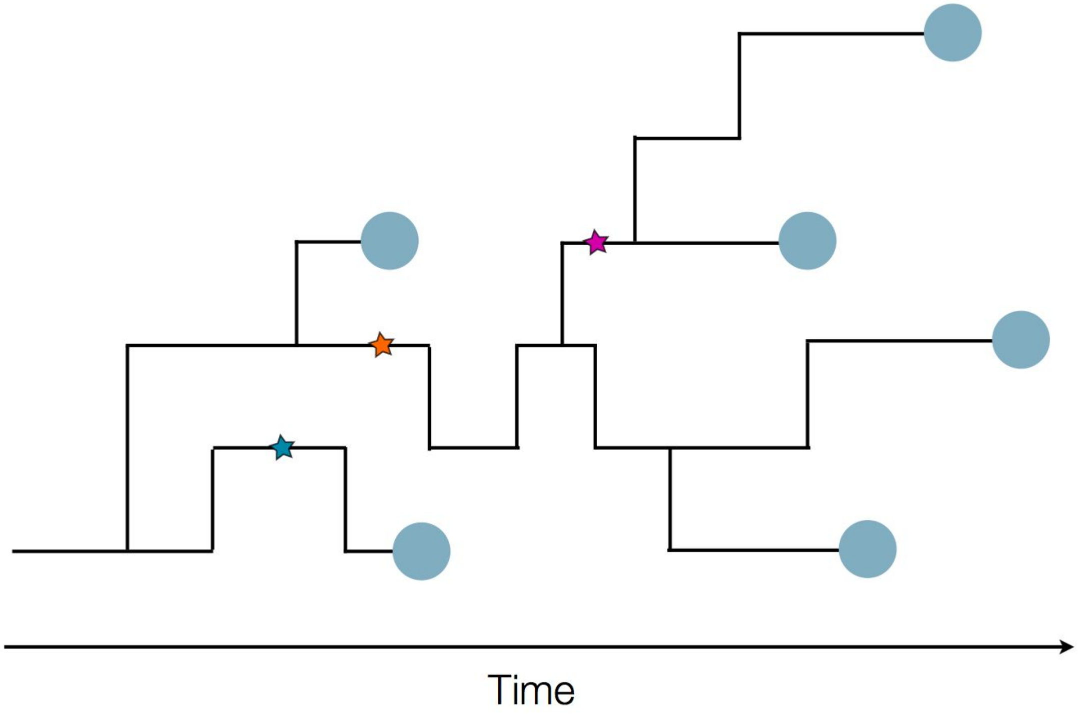
https://github.com/trvrb/phylodynamics-lecture

Sampling & partially reconstructing underlying epidemic process



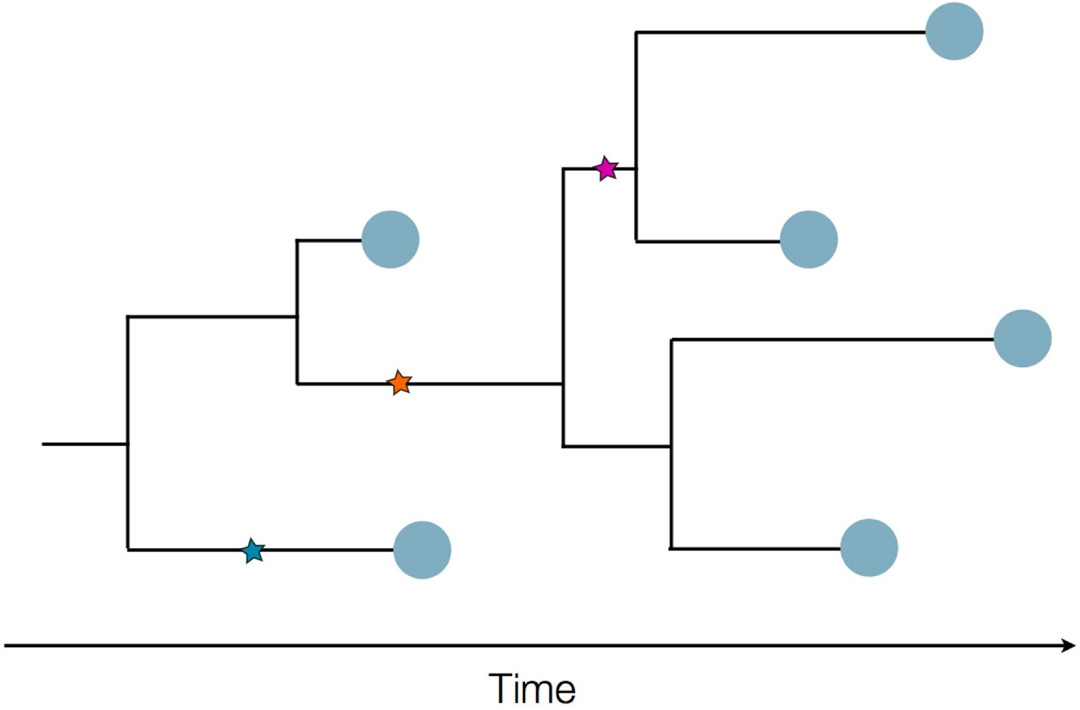
https://github.com/trvrb/phylodynamics-lecture

Sampling & partially reconstructing underlying epidemic process



https://github.com/trvrb/phylodynamics-lecture

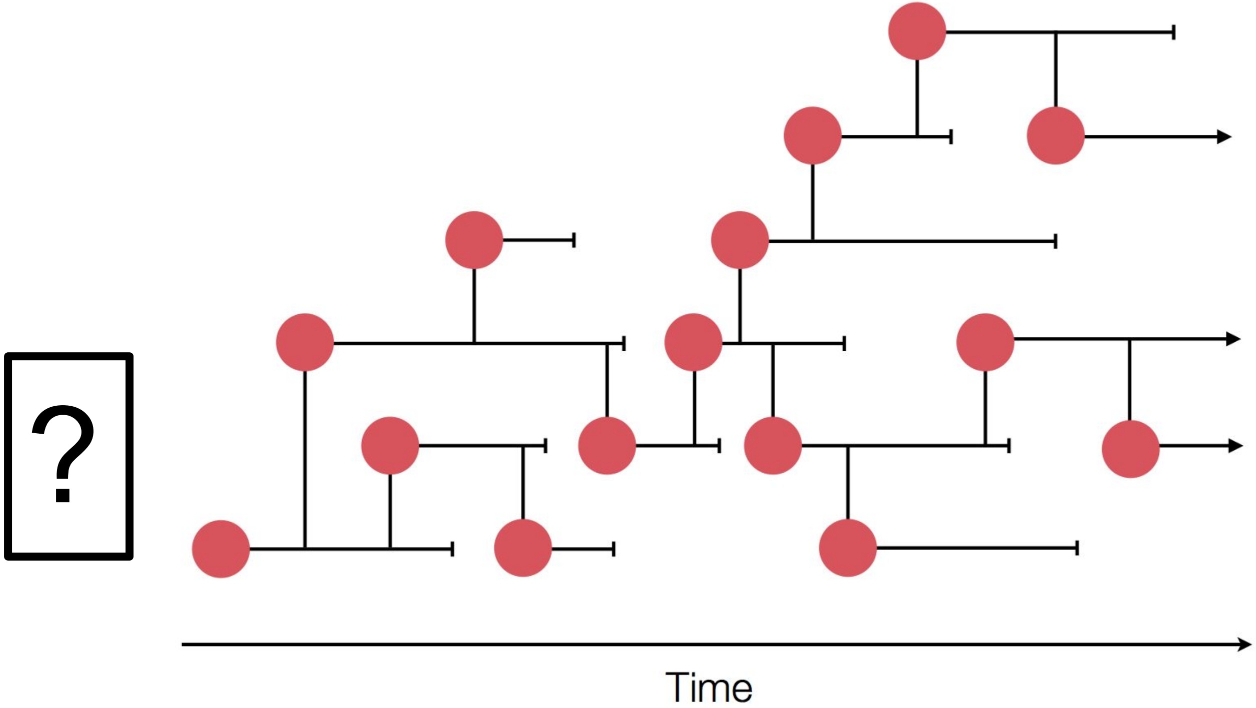
Sampling & partially reconstructing underlying epidemic process



https://github.com/trvrb/phylodynamics-lecture

What determines underlying process?

Many forces shaping underlying process



https://github.com/trvrb/phylodynamics-lecture

Many forces shaping underlying process

Within-host Evolution

Between-host Evolution

Host Migration

Vaccination

Host Immune Response

Infectivity

Virulence

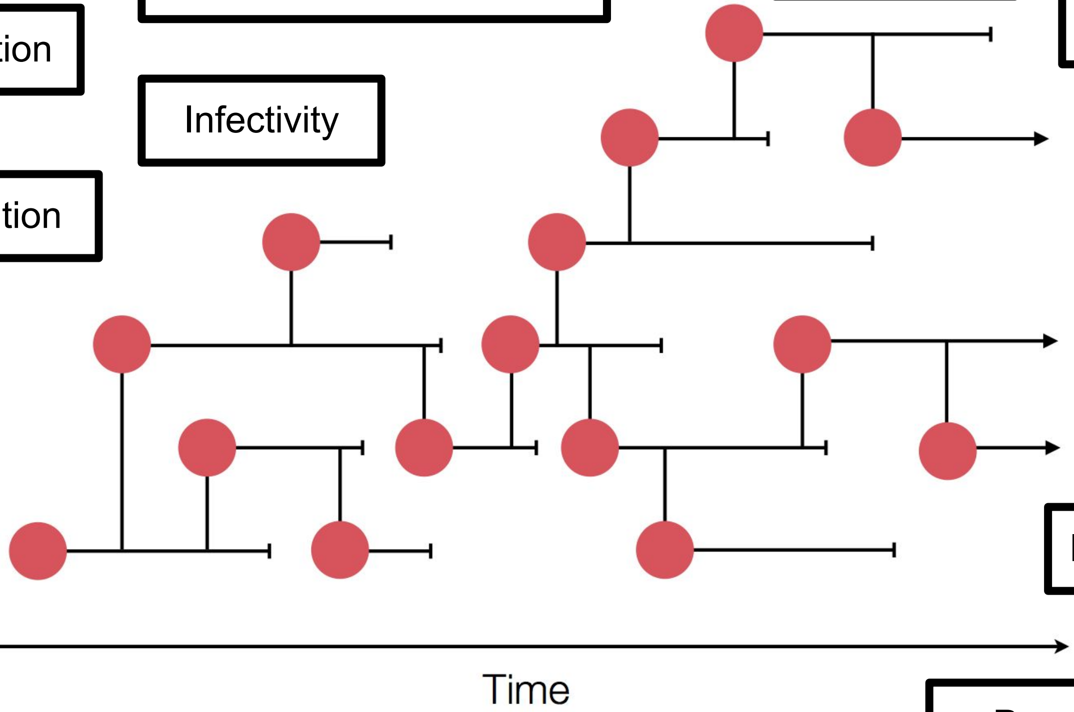
Non-Pharmaceutical Interventions

Mutation Rate

Generation Time

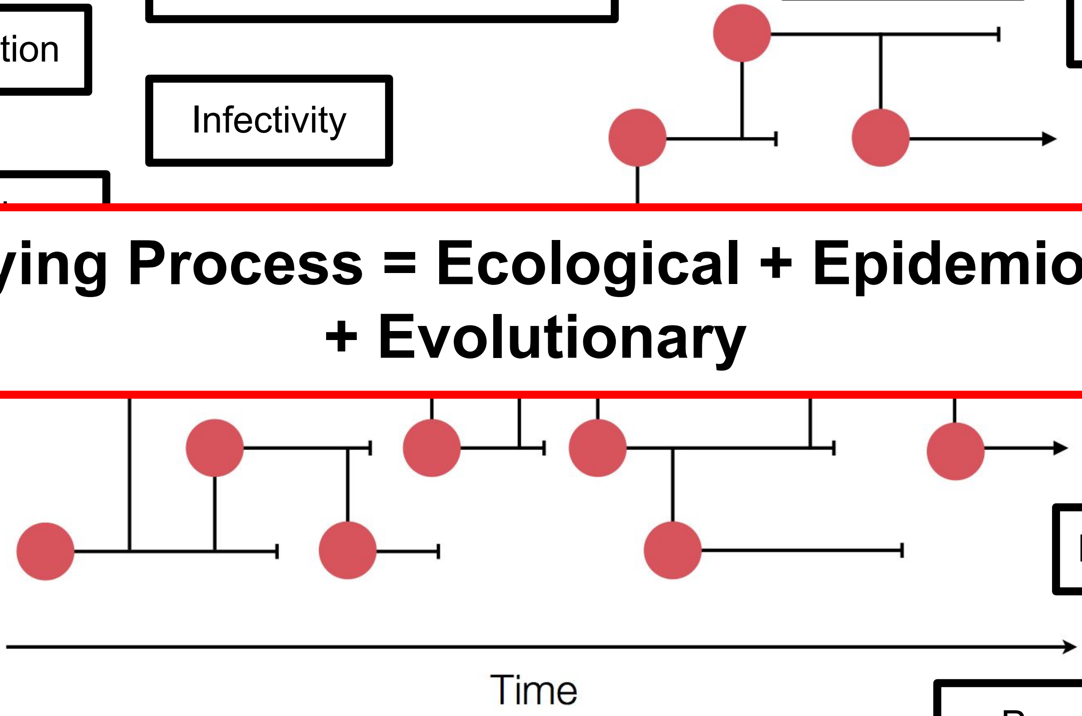
Immune Memory

Population Structure



https://github.com/trvrb/phylodynamics-lecture

Many forces shaping underlying process



Within-host Evolution

Host Immune Response

Infectivity

Virulence

Non-Pharmaceutical Interventions

Betwe

Rate

Underlying Process = Ecological + Epidemiological + Evolutionary

Host Migration

Generation Time

Vaccination

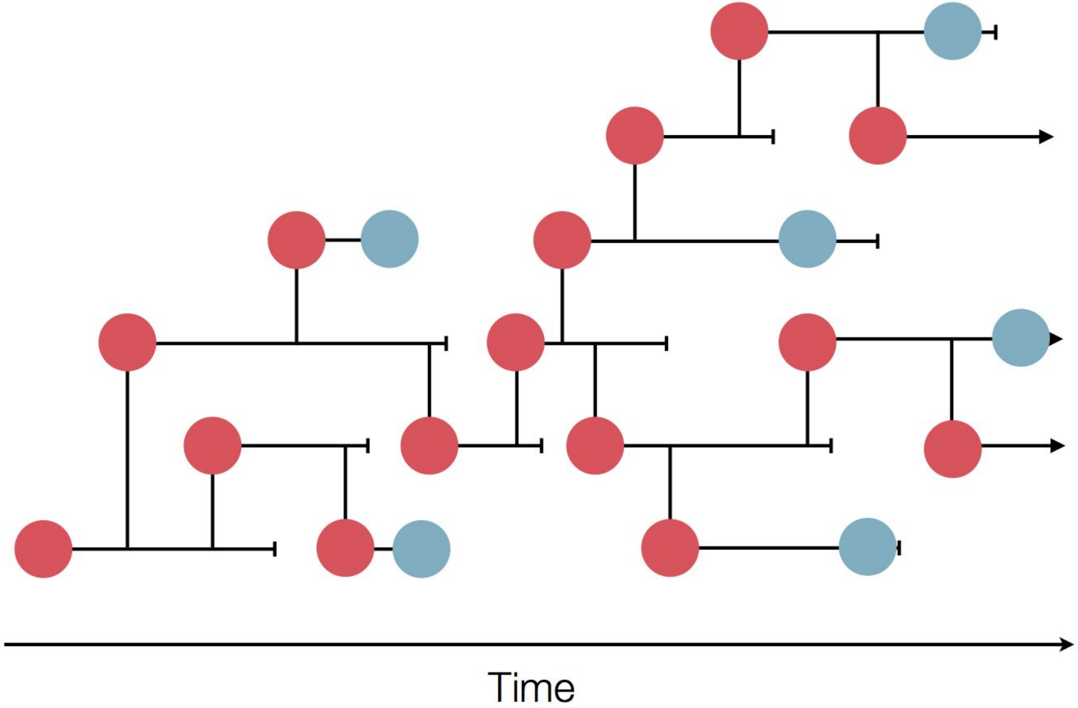
Immune Memory

Population Structure

https://github.com/trvrb/phylodynamics-lecture

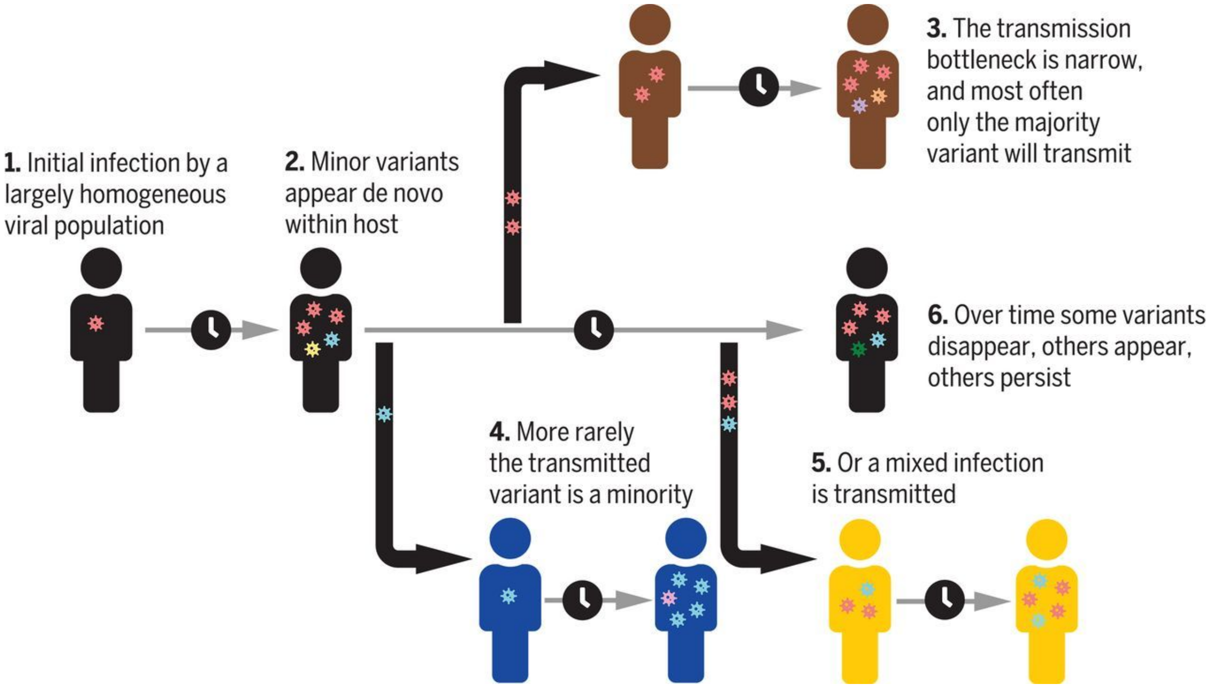
Let’s start with the “simple” reconstruction of the transmission network

Complicated sampling of a (within-host) population of a (between host) population



https://github.com/trvrb/phylodynamics-lecture

Complicated sampling of a (within-host) population of a (between host) population



1. Initial infection by a largely homogeneous viral population

2. Minor variants appear de novo within host

3. The transmission bottleneck is narrow, and most often only the majority variant will transmit

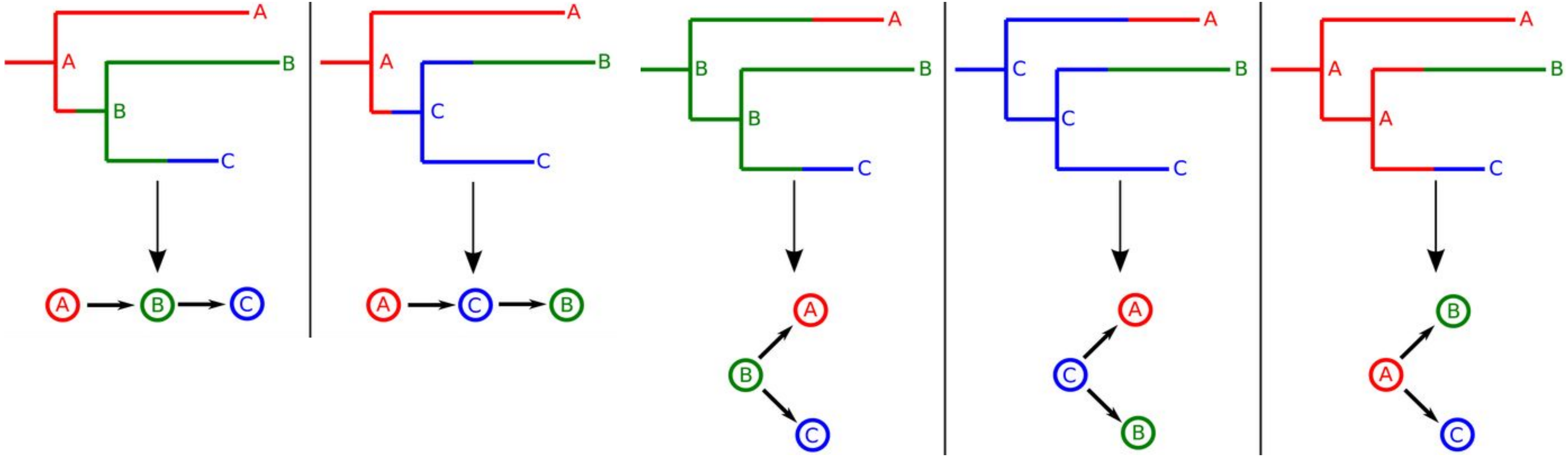
4. More rarely the transmitted variant is a minority

5. Or a mixed infection is transmitted

6. Over time some variants disappear, others appear, others persist

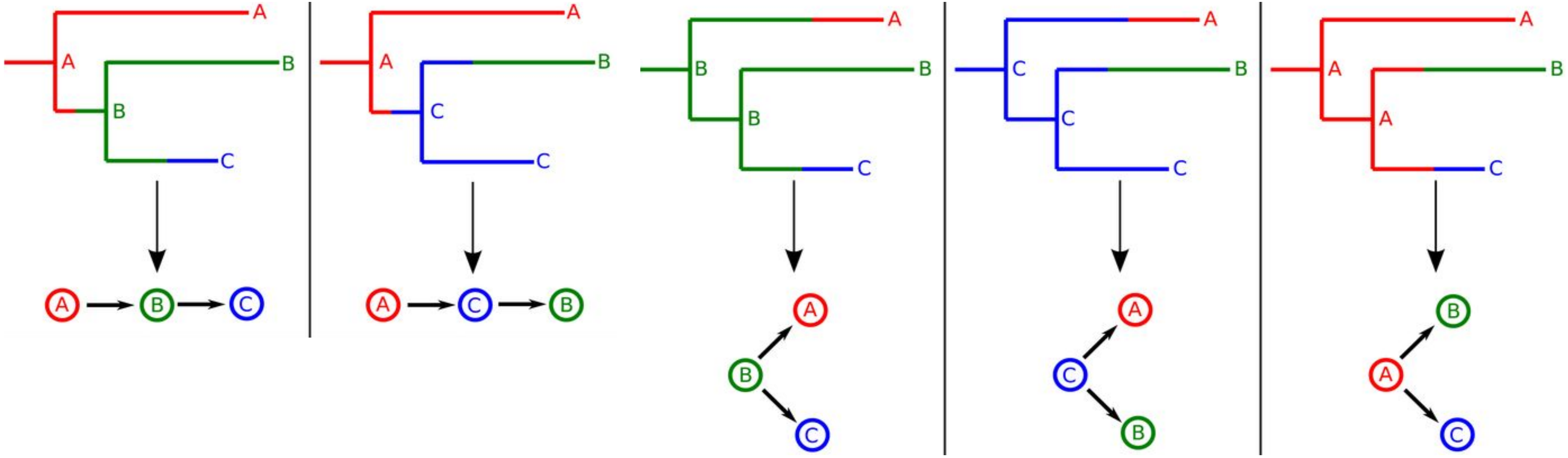
https://www.science.org/doi/10.1126/science.abg0821

Same tree can be consistent with different scenarios



10.1371/journal.pcbi.1004613

Same tree can be consistent with different scenarios



10.1371/journal.pcbi.1004613

=> Probabilistic inference!

How can we can model phylodynamic processes probabilistically?

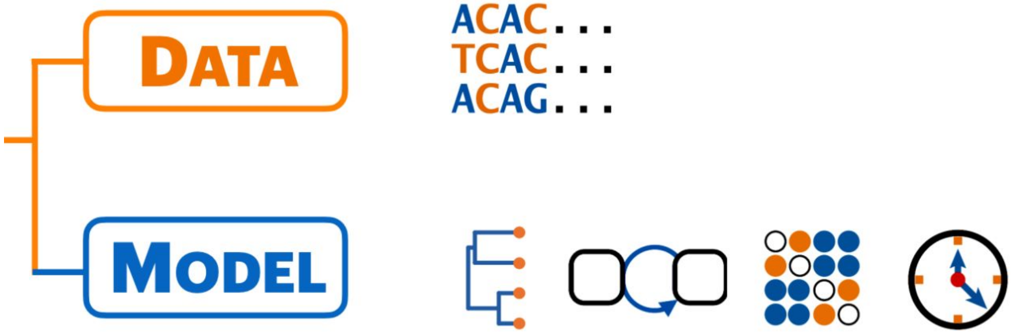
Bayesian inference is a key tool in genomic epidemiology

DATA

ACAC...
TCAC...
ACAG...

https://github.com/Taming-the-BEAST/Taming-the-BEAST-2021-Online-Lectures/blob/master/Day1_First_Steps_-_Ceci_Valenzuela.pdf

Bayesian inference is a key tool in genomic epidemiology



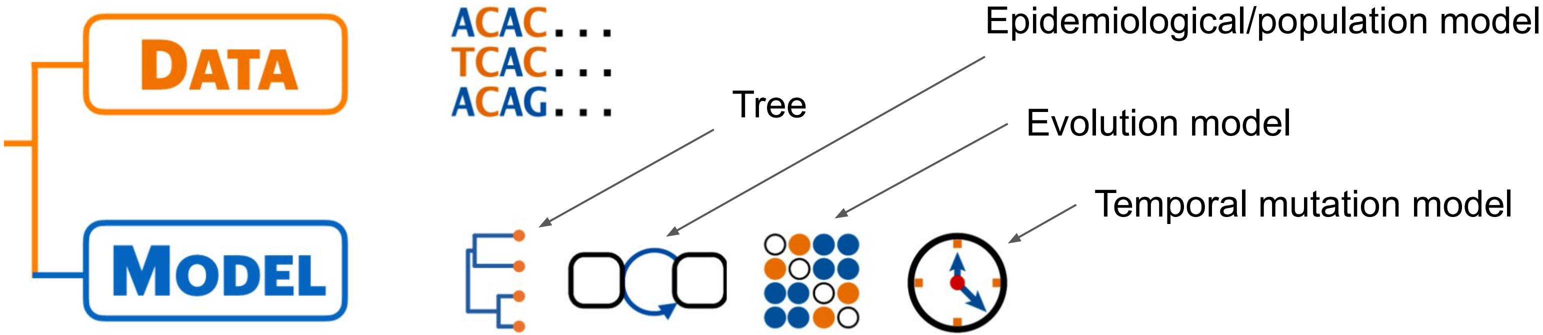
DATA

ACAC...
TCAC...
ACAG...

MODEL

https://github.com/Taming-the-BEAST/Taming-the-BEAST-2021-Online-Lectures/blob/master/Day1_First_Steps_-_Ceci_Valenzuela.pdf

Bayesian inference is a key tool in genomic epidemiology



Epidemiological/population model

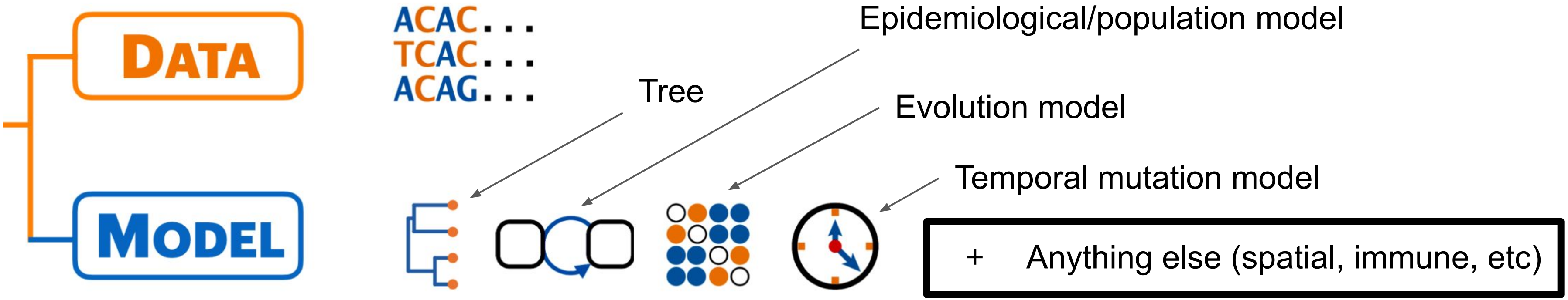
Tree

Evolution model

Temporal mutation model

https://github.com/Taming-the-BEAST/Taming-the-BEAST-2021-Online-Lectures/blob/master/Day1_First_Steps_-_Ceci_Valenzuela.pdf

Bayesian inference is a key tool in genomic epidemiology



DATA

ACAC...
TCAC...
ACAG...

MODEL

Tree

Epidemiological/population model

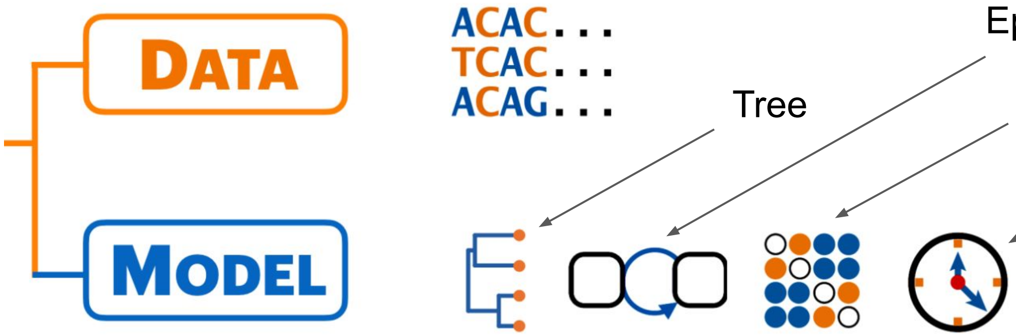
Evolution model

Temporal mutation model

\+ Anything else (spatial, immune, etc)

https://github.com/Taming-the-BEAST/Taming-the-BEAST-2021-Online-Lectures/blob/master/Day1_First_Steps_-_Ceci_Valenzuela.pdf

Bayesian inference is a key tool in genomic epidemiology



Tree

Epidemiological/population model

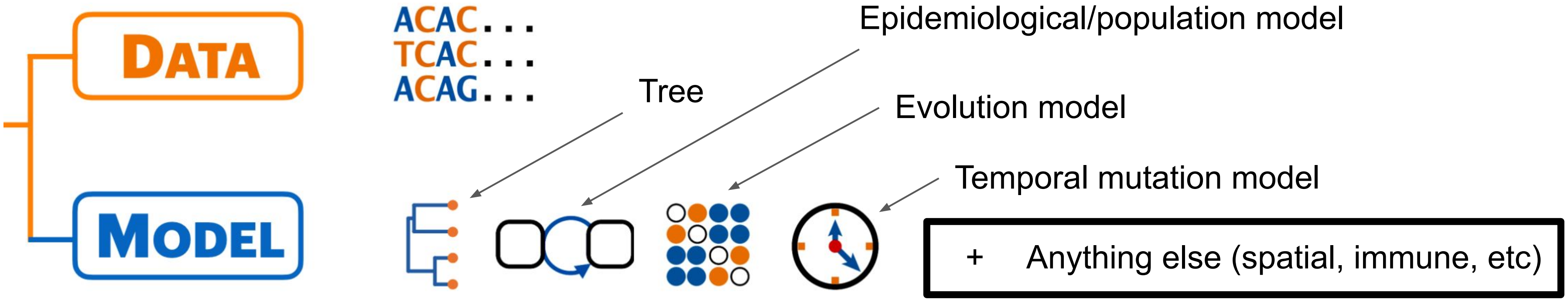
Evolution model

Temporal mutation model

\+ Anything else (spatial, immune, etc)

https://github.com/Taming-the-BEAST/Taming-the-BEAST-2021-Online-Lectures/blob/master/Day1_First_Steps_-_Ceci_Valenzuela.pdf

Bayesian inference is a key tool in genomic epidemiology



DATA

ACAC...
TCAC...
ACAG...

MODEL

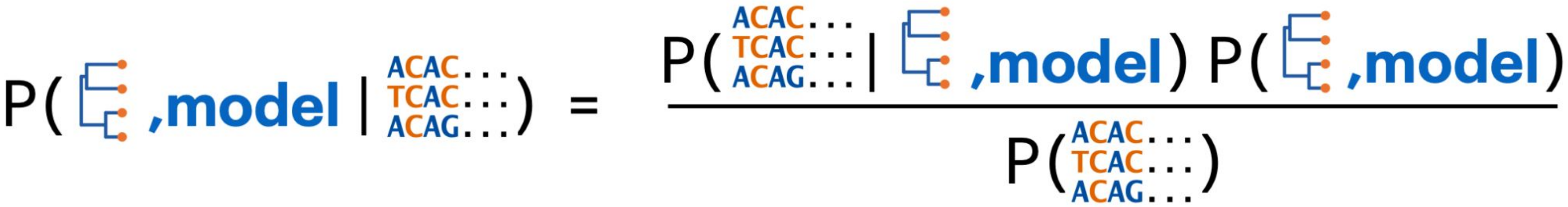
Tree

Epidemiological/population model

Evolution model

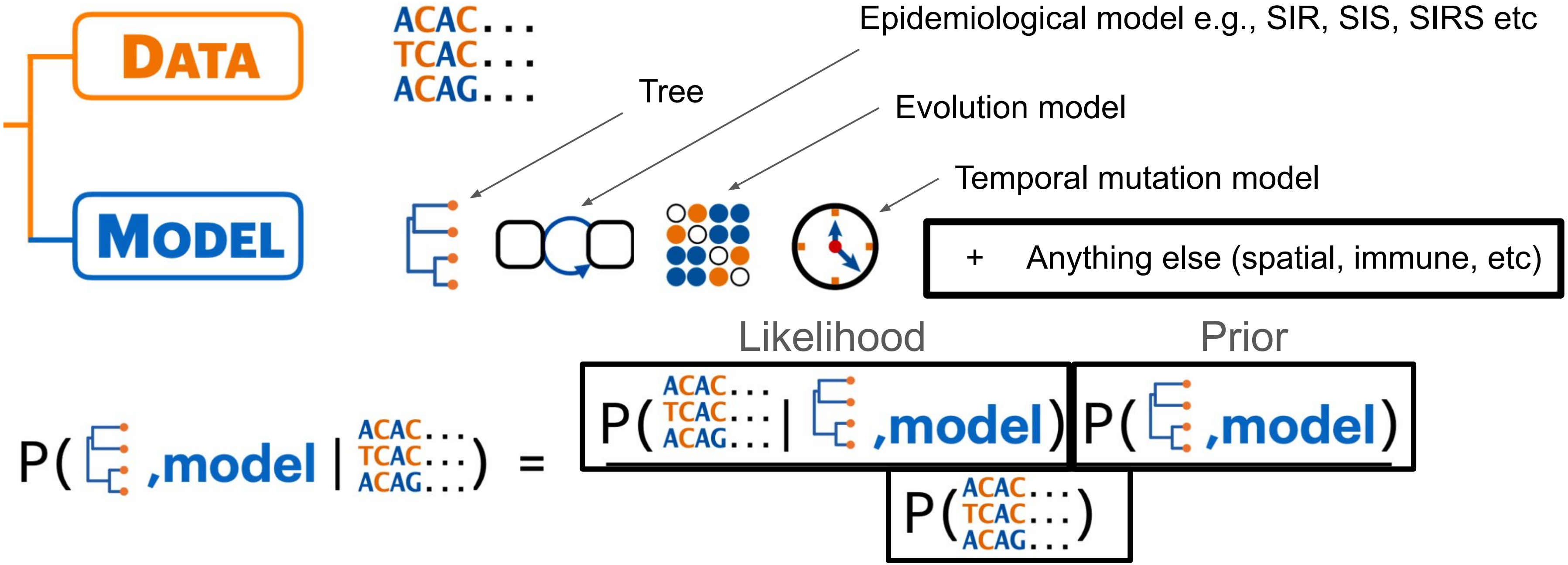
Temporal mutation model

\+ Anything else (spatial, immune, etc)



https://github.com/Taming-the-BEAST/Taming-the-BEAST-2021-Online-Lectures/blob/master/Day1_First_Steps_-_Ceci_Valenzuela.pdf

Bayesian inference is a key tool in phylodynamics



Tree

Epidemiological model e.g., SIR, SIS, SIRS etc

Evolution model

Temporal mutation model

\+ Anything else (spatial, immune, etc)

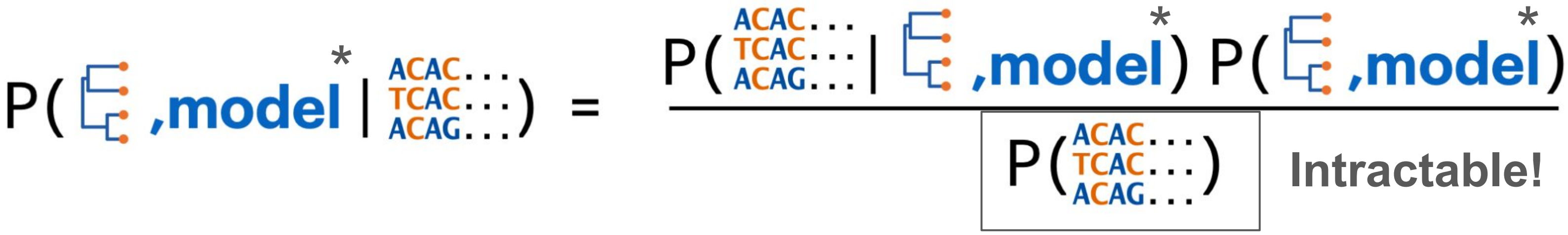
Likelihood

Prior

Marginal Probability

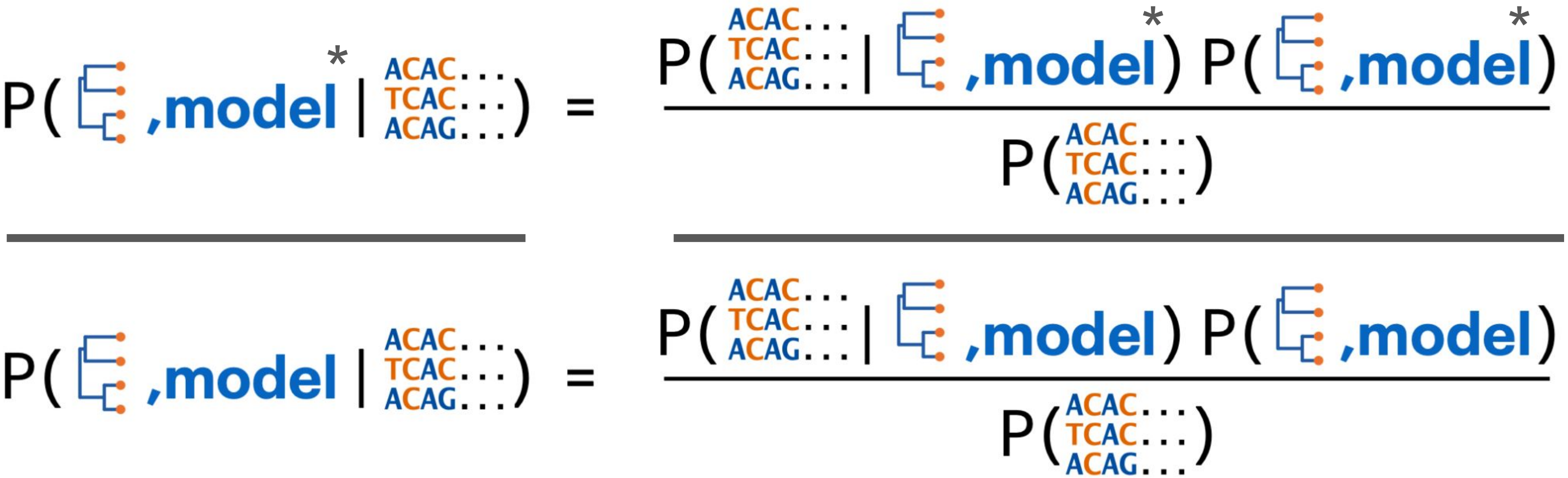
https://github.com/Taming-the-BEAST/Taming-the-BEAST-2021-Online-Lectures/blob/master/Day1_First_Steps_-_Ceci_Valenzuela.pdf

Markov Chain Monte Carlo (MCMC) Sample



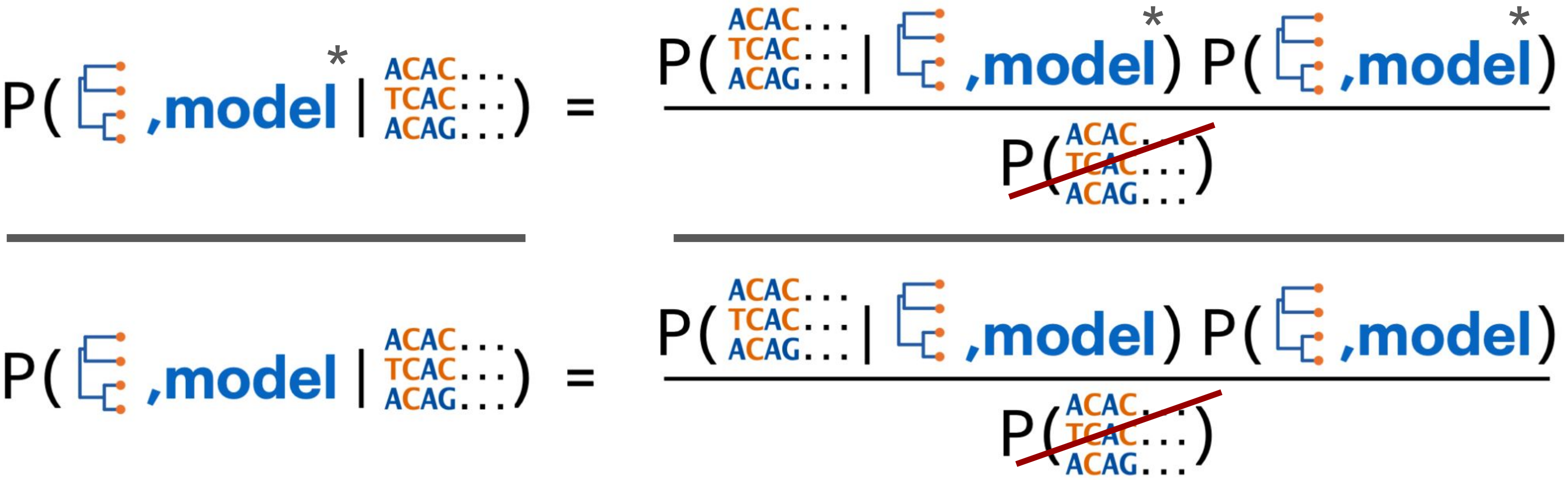
Intractable!

Markov Chain Monte Carlo (MCMC) Sample



Posterior Odds

Markov Chain Monte Carlo (MCMC) Sample



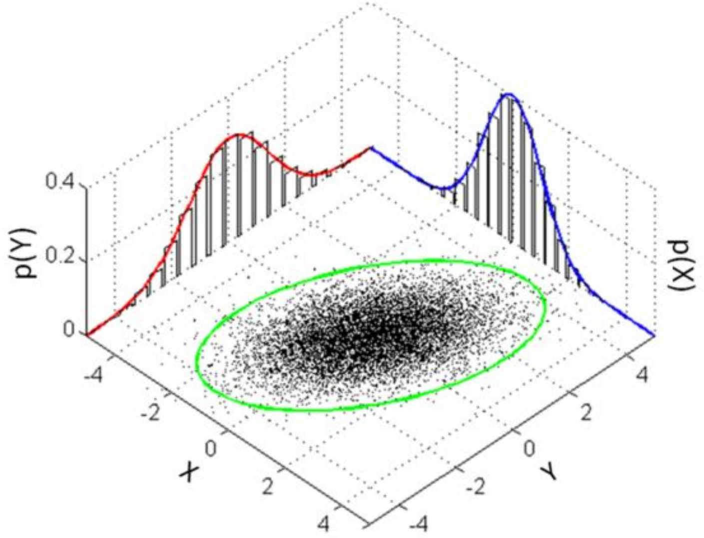
Posterior Odds

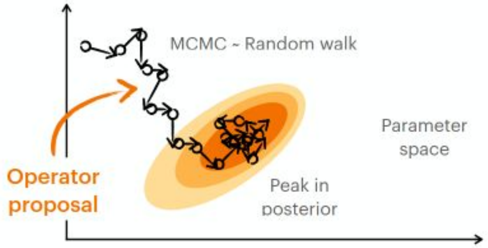
Likelihood Ratio

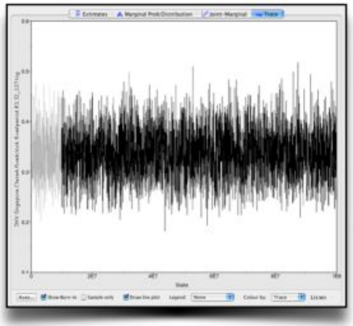
Prior Odds

Markov Chain Monte Carlo (MCMC) Sample

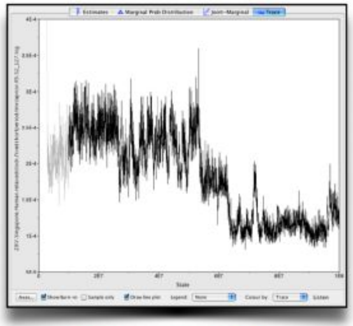
We sample from the joint posterior







Mixing well! 😄

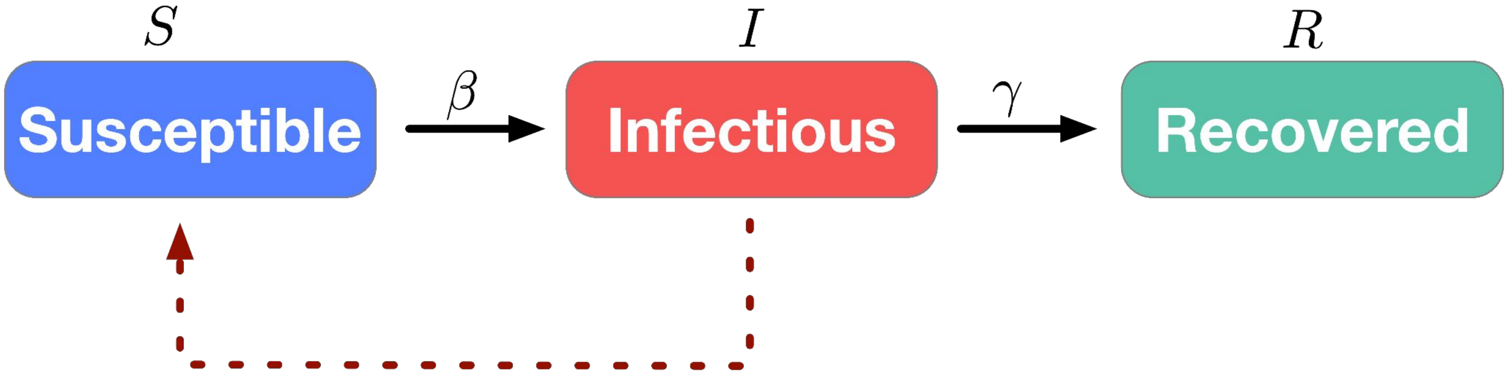


Not mixing! 😭

Live Demo - https://delphy.fathom.info/

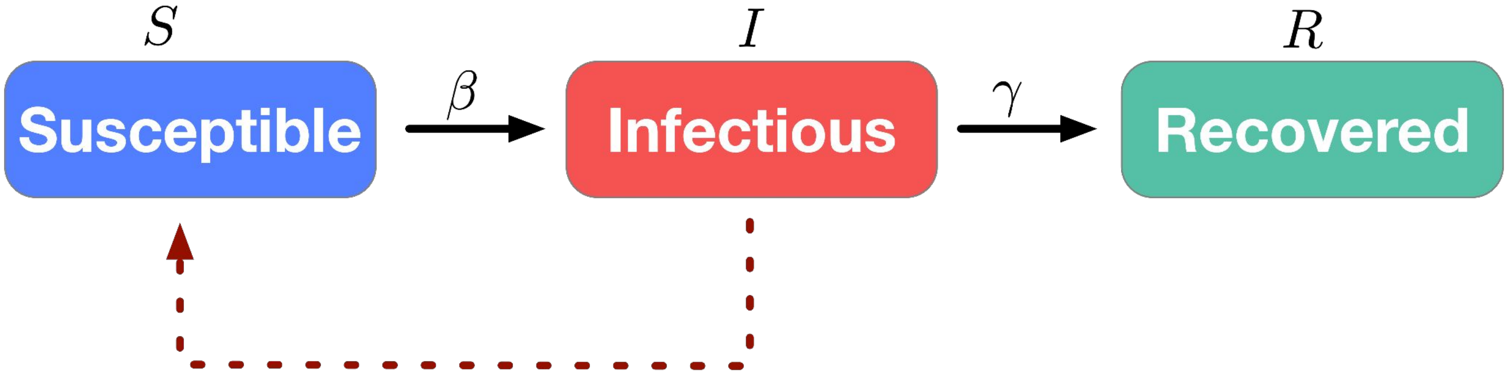
What kind of epidemic process parameters do we want to infer?

Compartmental models are used to model infections



https://covid19.uclaml.org/model.html

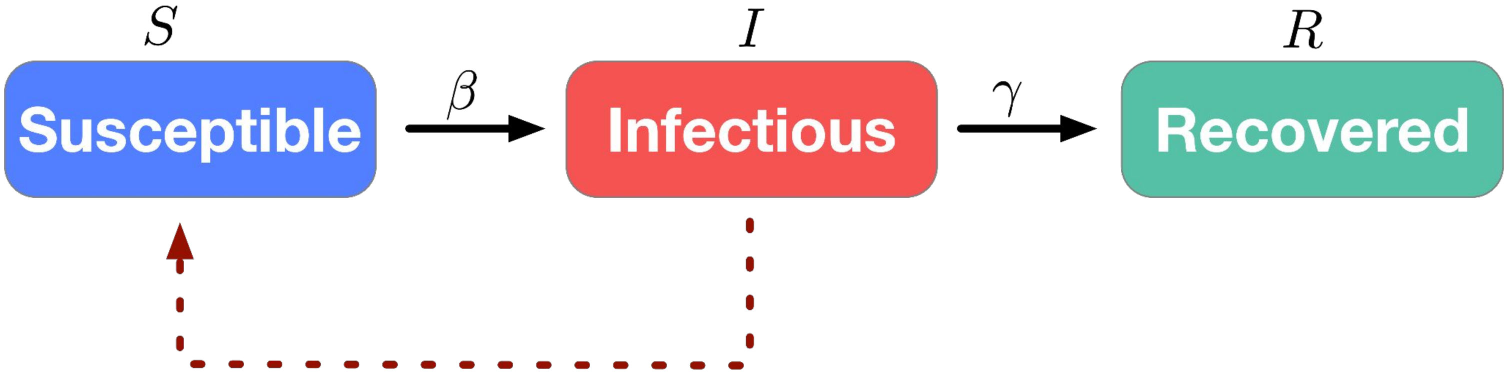
Compartmental models are used to model infections



Disclaimer: many more complex models!

https://covid19.uclaml.org/model.html

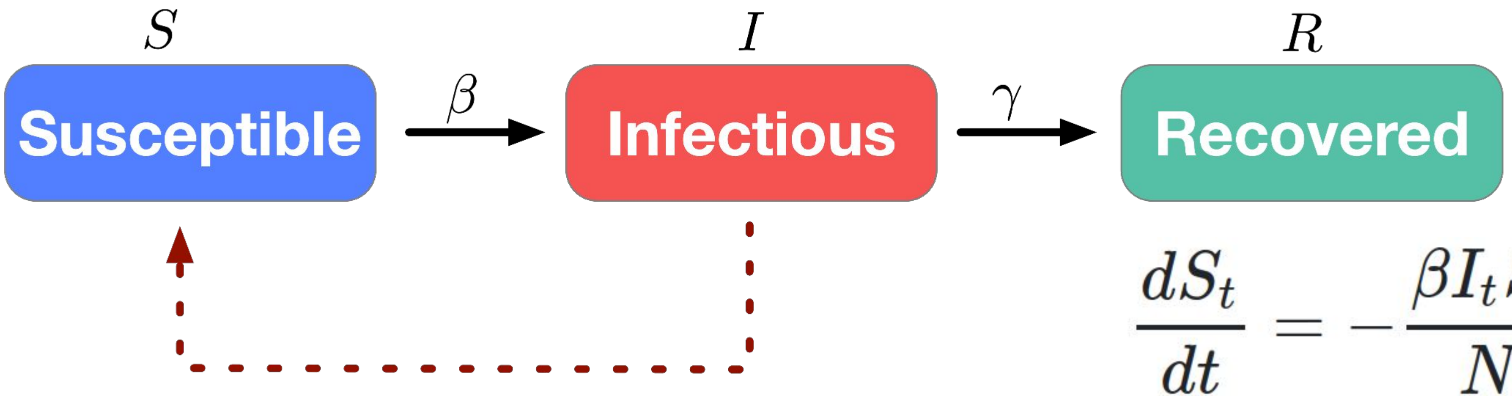
Compartmental models are used to model infections



- S_t: the number of susceptible individuals
- I_t: the number of infectious individuals
- R_t: the number of recovered/deceased/immune individuals

https://covid19.uclaml.org/model.html

Compartmental models are used to model infections



$$\frac{dS_t}{dt} = -\frac{\beta I_t S_t}{N}$$

$$\frac{dI_t}{dt} = \frac{\beta I_t S_t}{N} - \gamma I_t$$

$$\frac{dR_t}{dt} = \gamma I_t$$

- S_t: the number of susceptible individuals
- I_t: the number of infectious individuals
- R_t: the number of recovered/deceased/immune individuals

https://covid19.uclaml.org/model.html

Assuming N = fixed pop

Can calculate P(observed case counts | β=?,γ=?) with cases

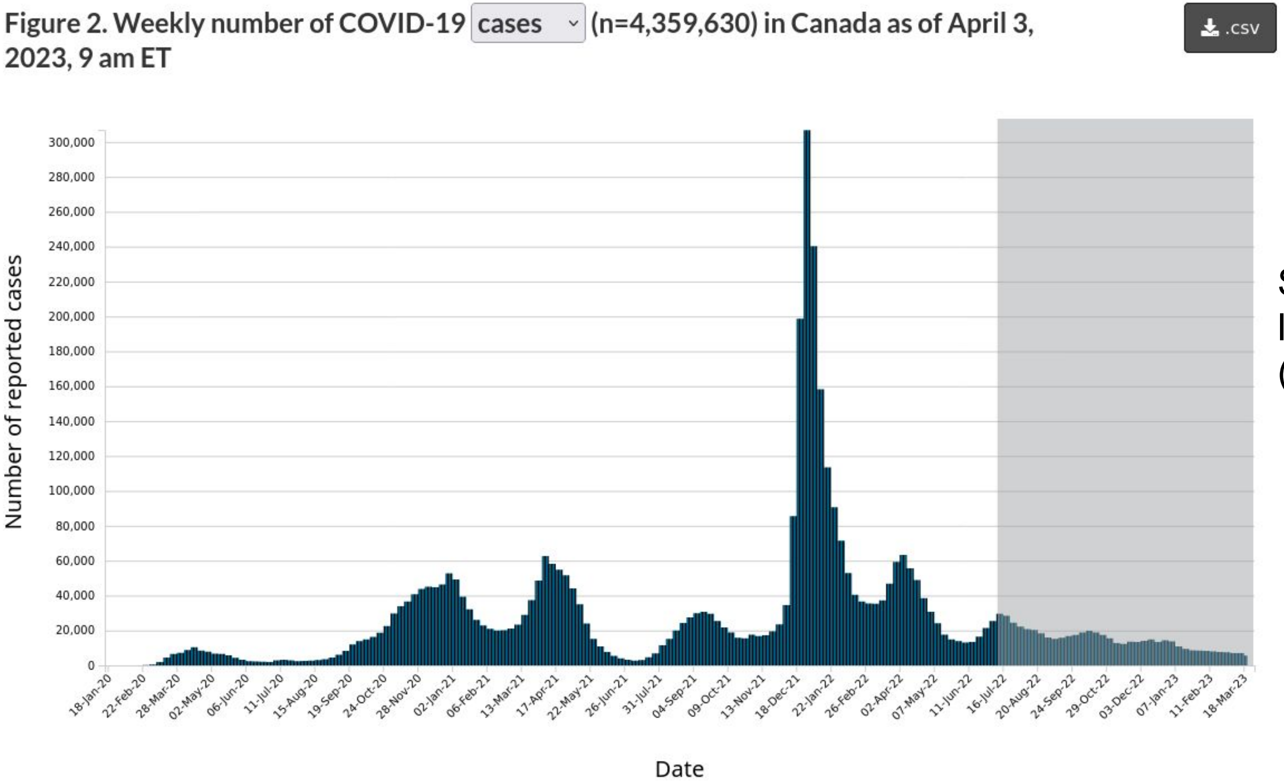


Figure 2. Weekly number of COVID-19 cases (n=4,359,630) in Canada as of April 3, 2023, 9 am ET

https://health-infobase.canada.ca/covid-19/current-situation.html

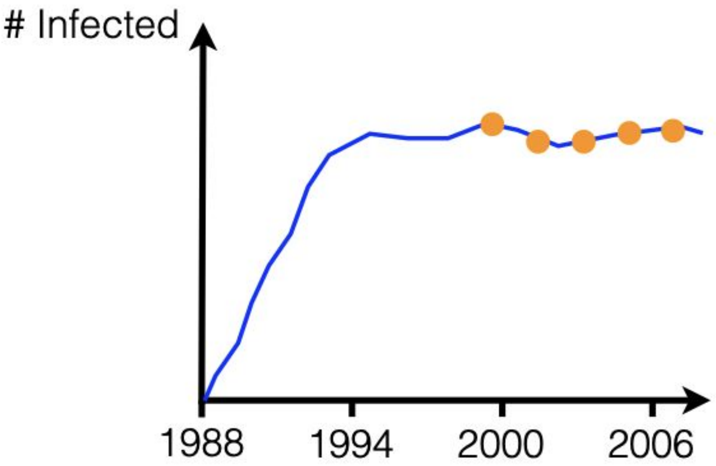
Same idea as Maximum likelihood Phylogenetics (just without any trees)

So, if we can this with case why do we bother using phylodynamics and genomic data?

Genomics can be used to infer unobserved events

If sampling in early epidemic was missed:

- **Time of epidemic outbreak?**
- **Basic reproductive number R_0?**

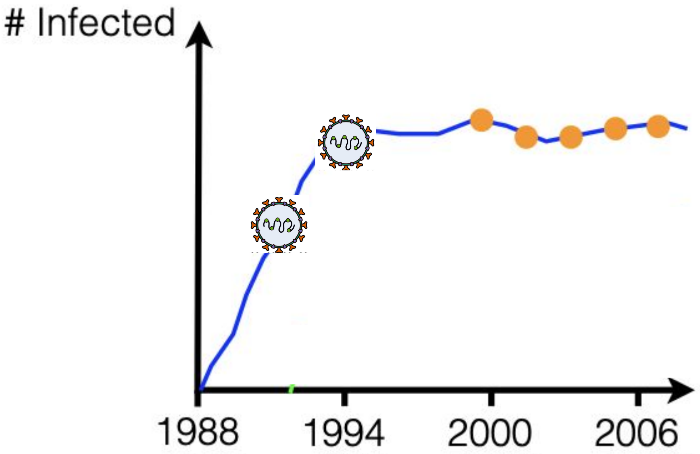


https://github.com/Taming-the-BEAST/Taming-the-BEAST-2021-Online-Lectures/raw/master/Day2_Phylodynamics_-_Tanja_Stadler.pdf

Genomics can be used to infer unobserved events

If sampling in early epidemic was missed:

- **Time of epidemic outbreak?**
- **Basic reproductive number R_0?**



https://github.com/Taming-the-BEAST/Taming-the-BEAST-2021-Online-Lectures/raw/master/Day2_Phylodynamics_-_Tanja_Stadler.pdf

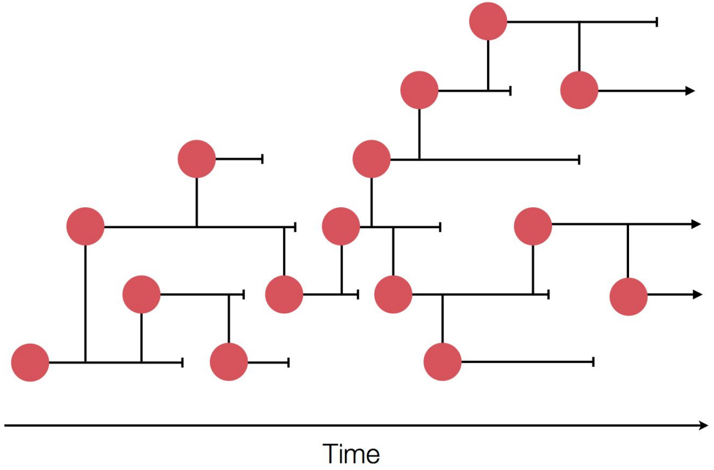
Genomics can be used to infer unobserved events

If sampling in early epidemic was missed:

- **Time of epidemic outbreak?**
- **Basic reproductive number R_0?**

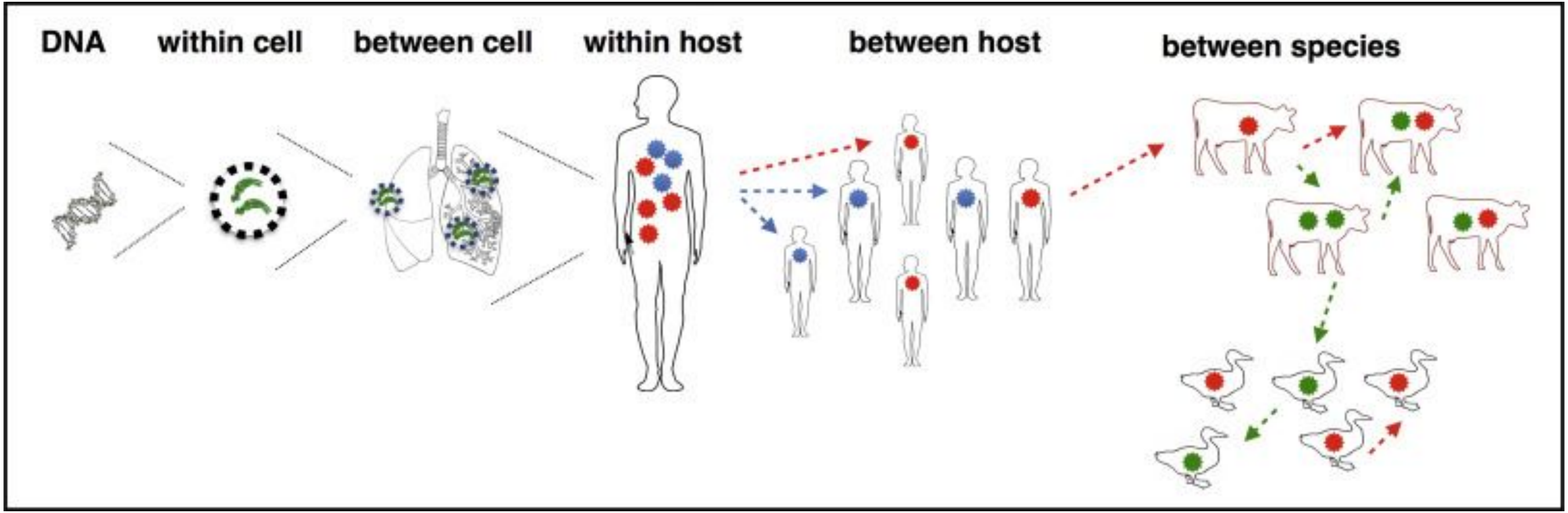
Data does not tell who infected whom:

- **Population structure?**



https://github.com/Taming-the-BEAST/Taming-the-BEAST-2021-Online-Lectures/raw/master/Day2_Phylodynamics_-_Tanja_Stadler.pdf

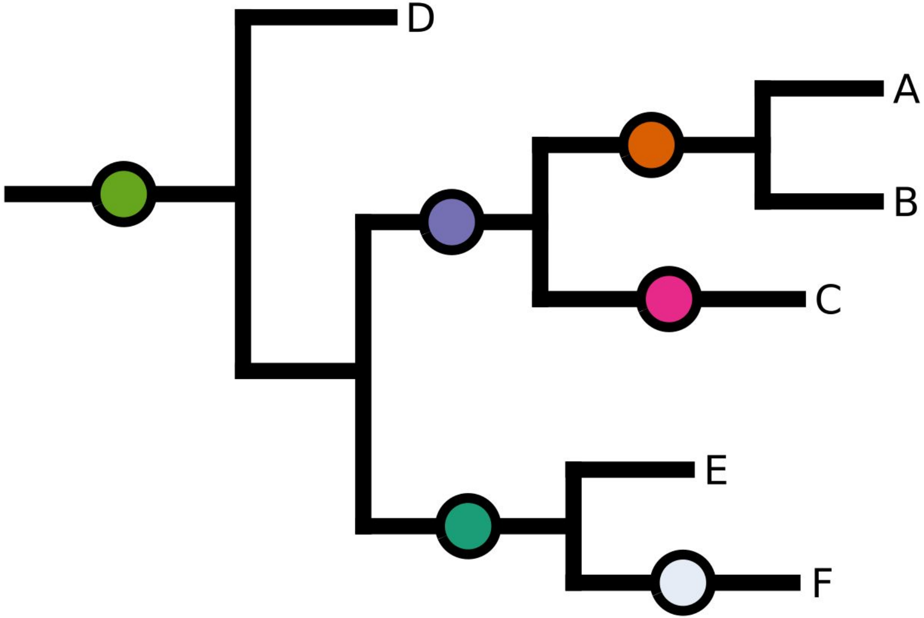
Cases don’t tell you (much) about pathogen evolution



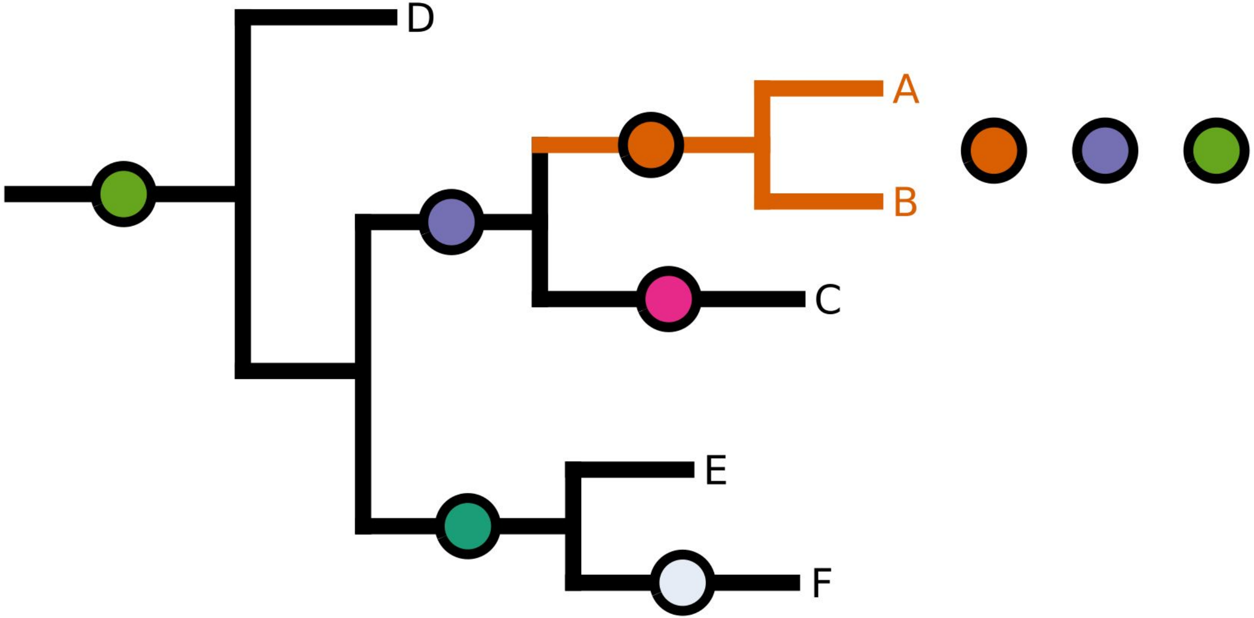
https://www.sciencedirect.com/science/article/pii/S1755436514000723

Let’s look at some specific use-cases

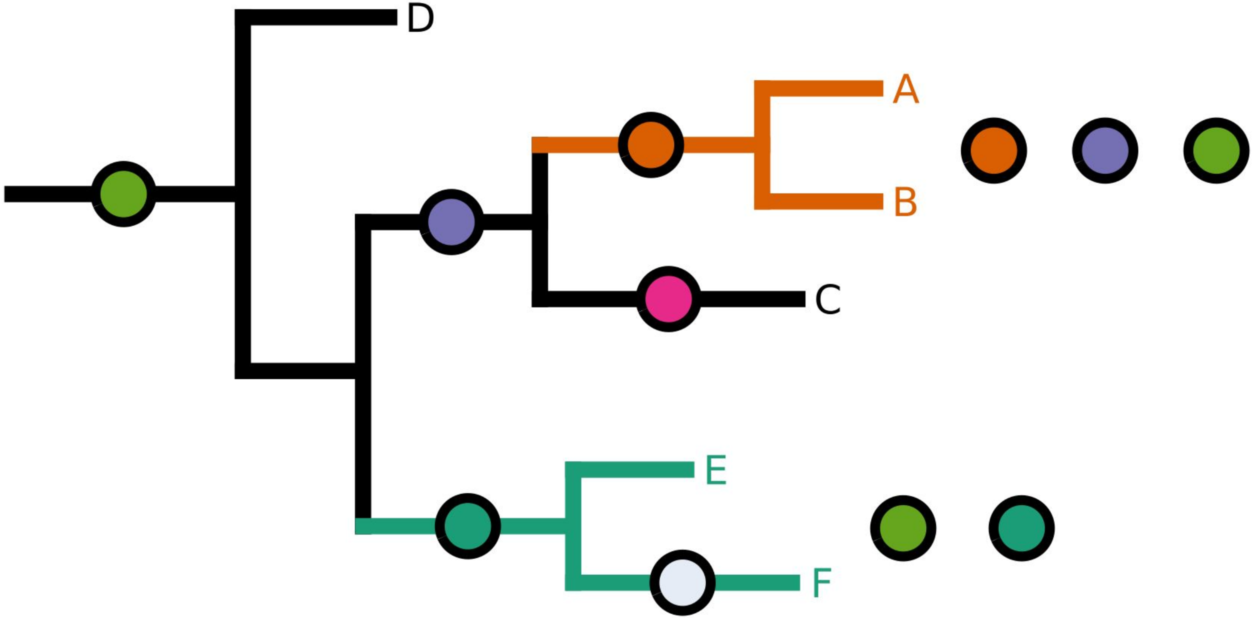
Define lineages (groups) of pathogens



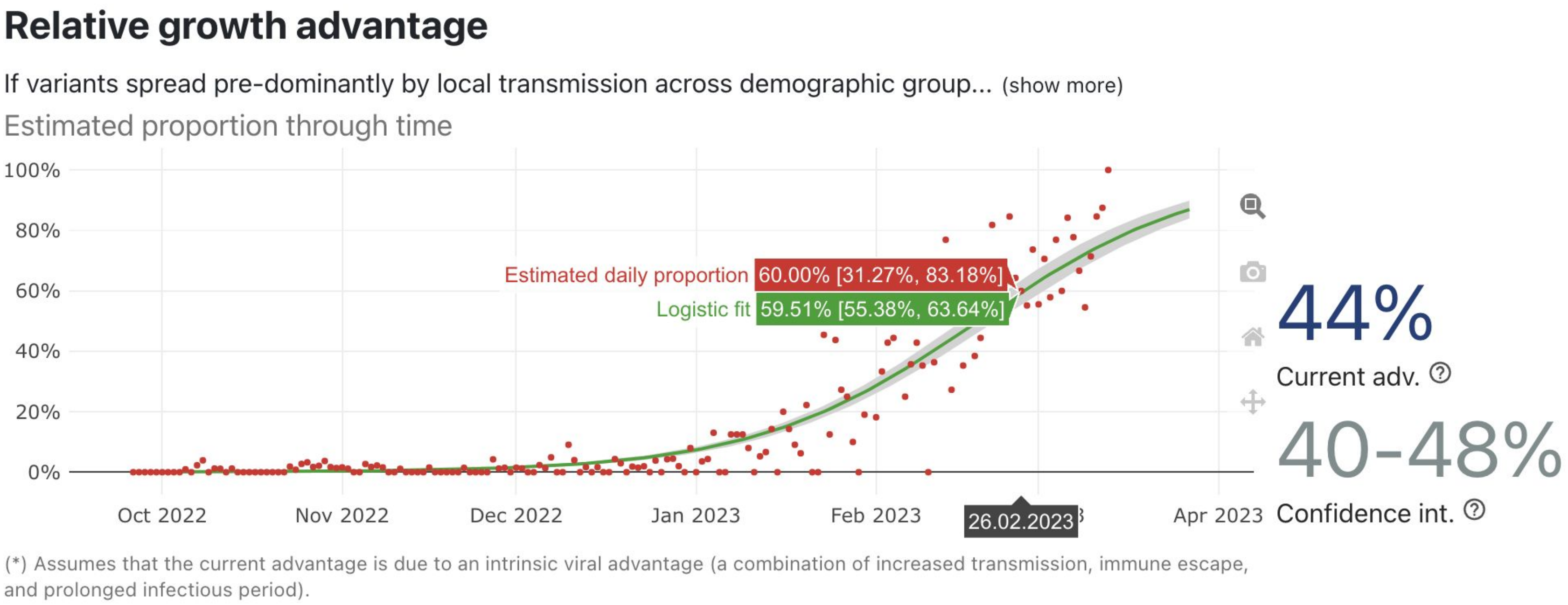
Define lineages (groups) of pathogens



Define lineages (groups) of pathogens



Early warning modelling lineage relative growth advantage



Relative growth advantage

If variants spread pre-dominantly by local transmission across demographic group... (show more)

Estimated proportion through time

44%

Current adv. ?

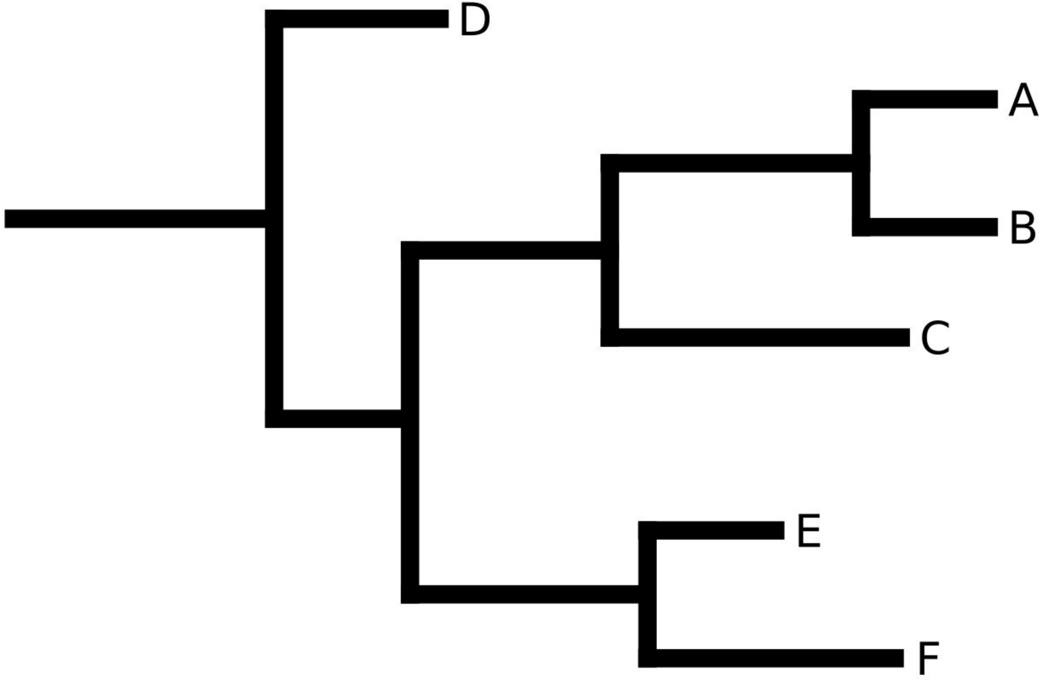
40-48%

Confidence int. ?

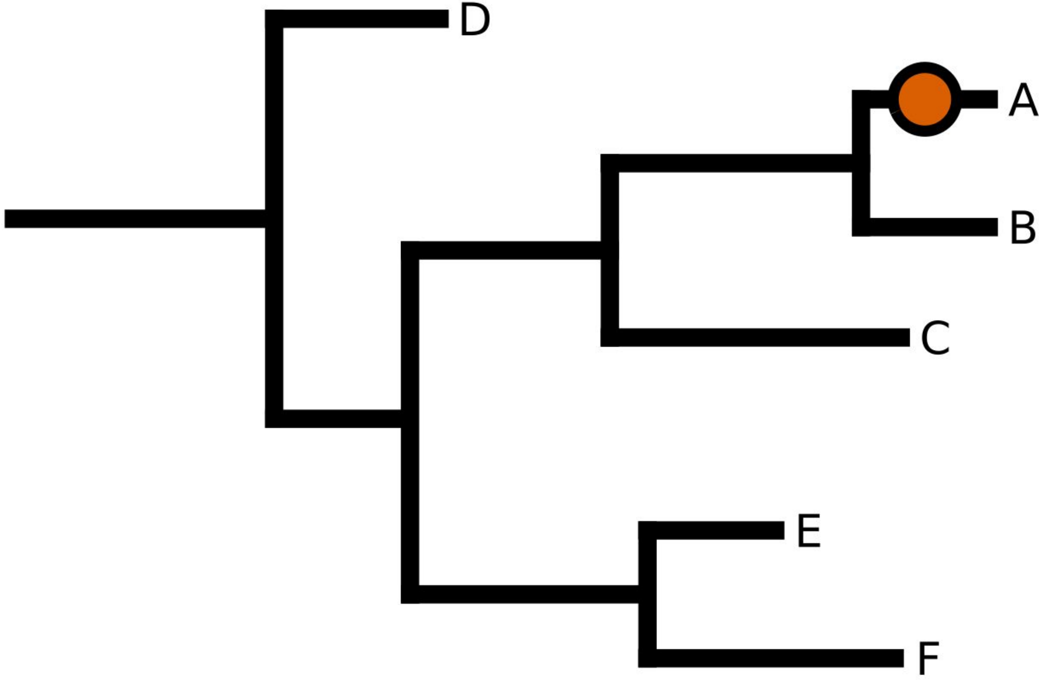
(*) Assumes that the current advantage is due to an intrinsic viral advantage (a combination of increased transmission, immune escape, and prolonged infectious period).

https://cov-spectrum.org/explore/Switzerland/Surveillance/Past6M/variants?nextcladePangoLineage=xbb*&

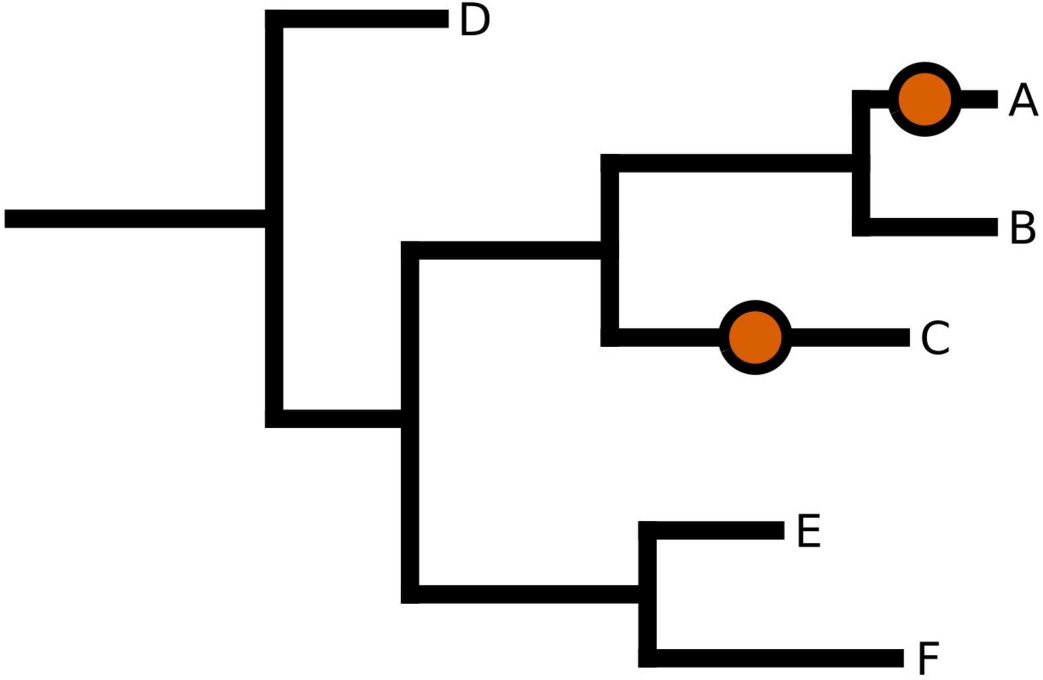
Identify clinically relevant mutations



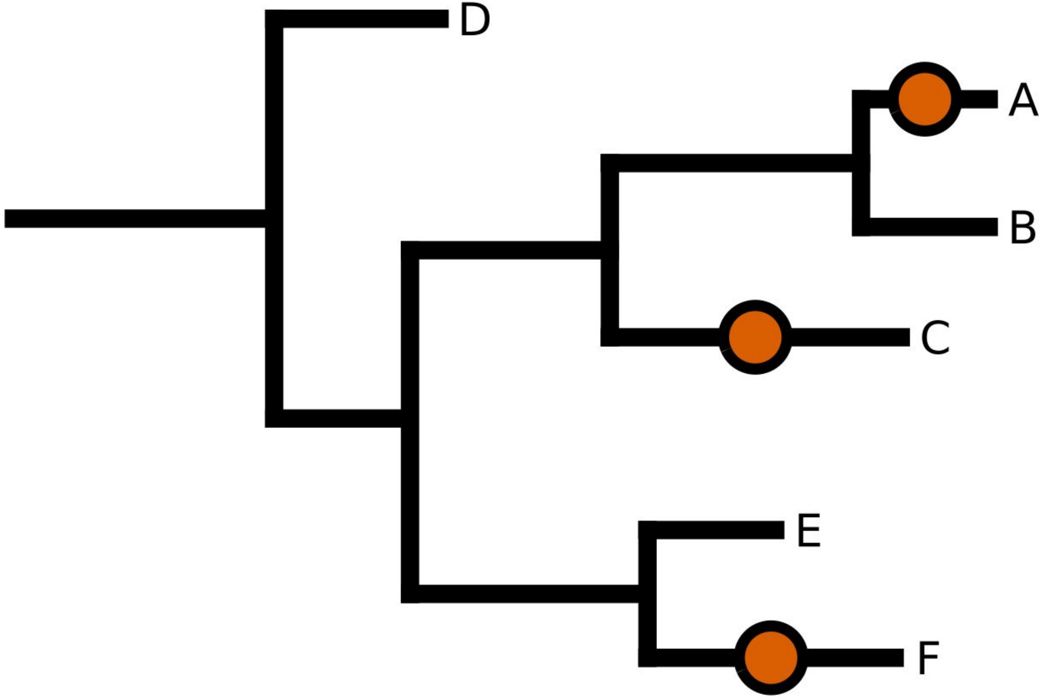
Identify clinically relevant mutations



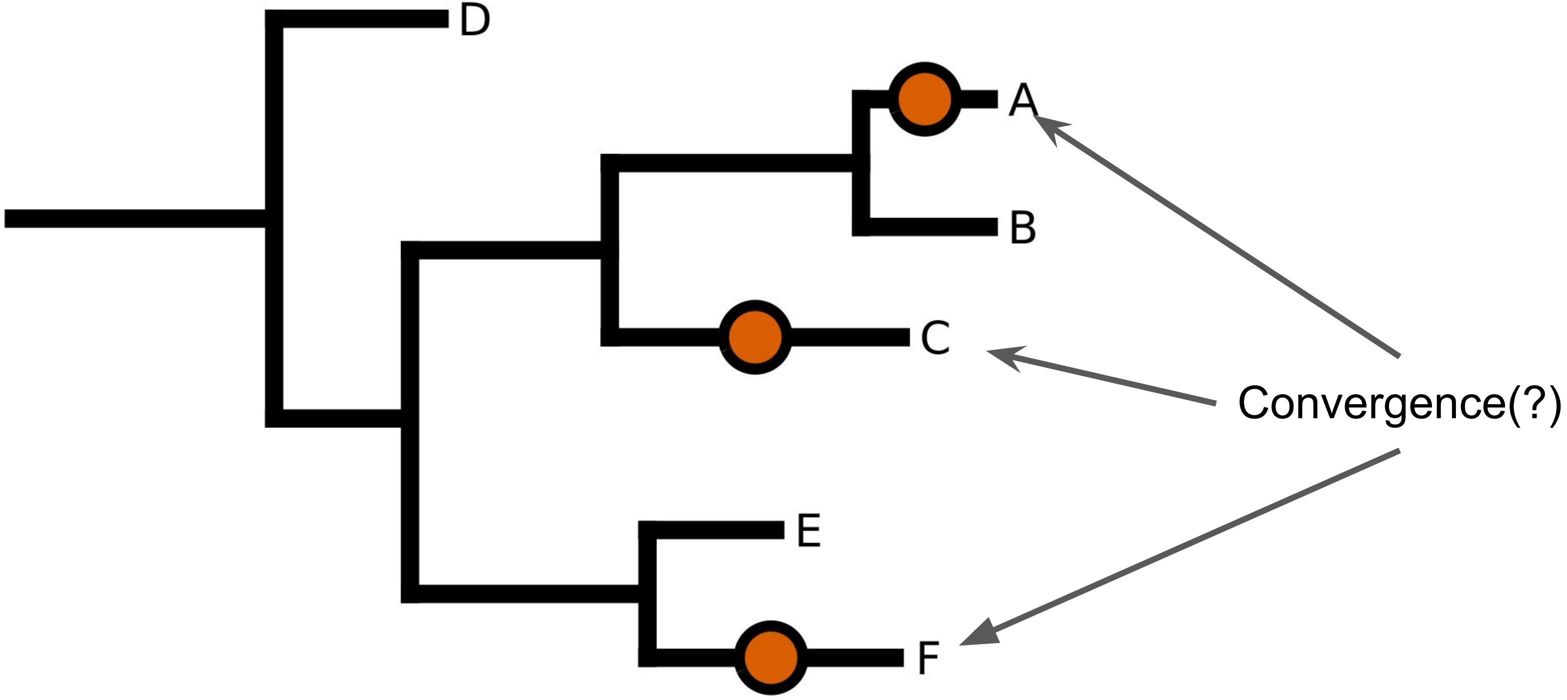
Identify clinically relevant mutations



Identify clinically relevant mutations

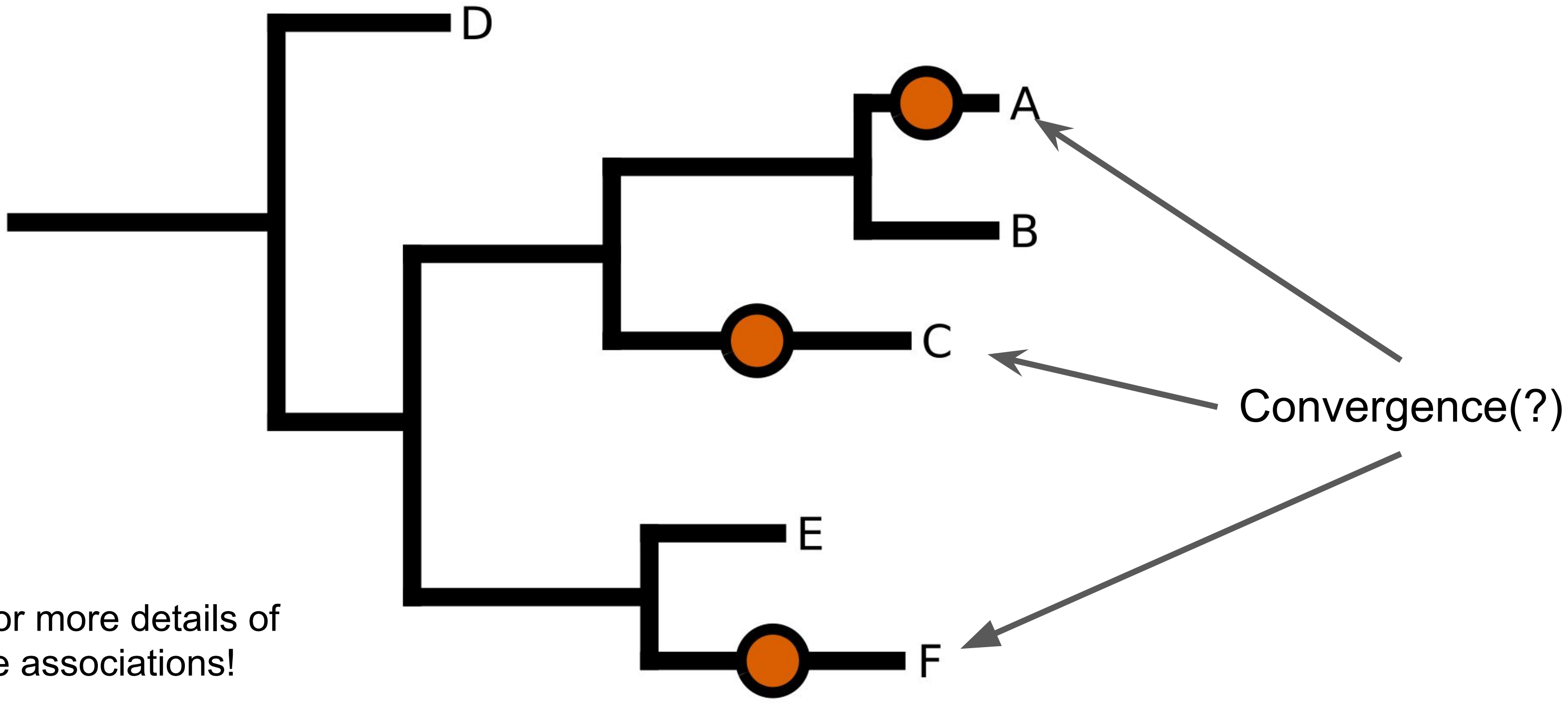


Identify clinically relevant mutations



Convergence(?)

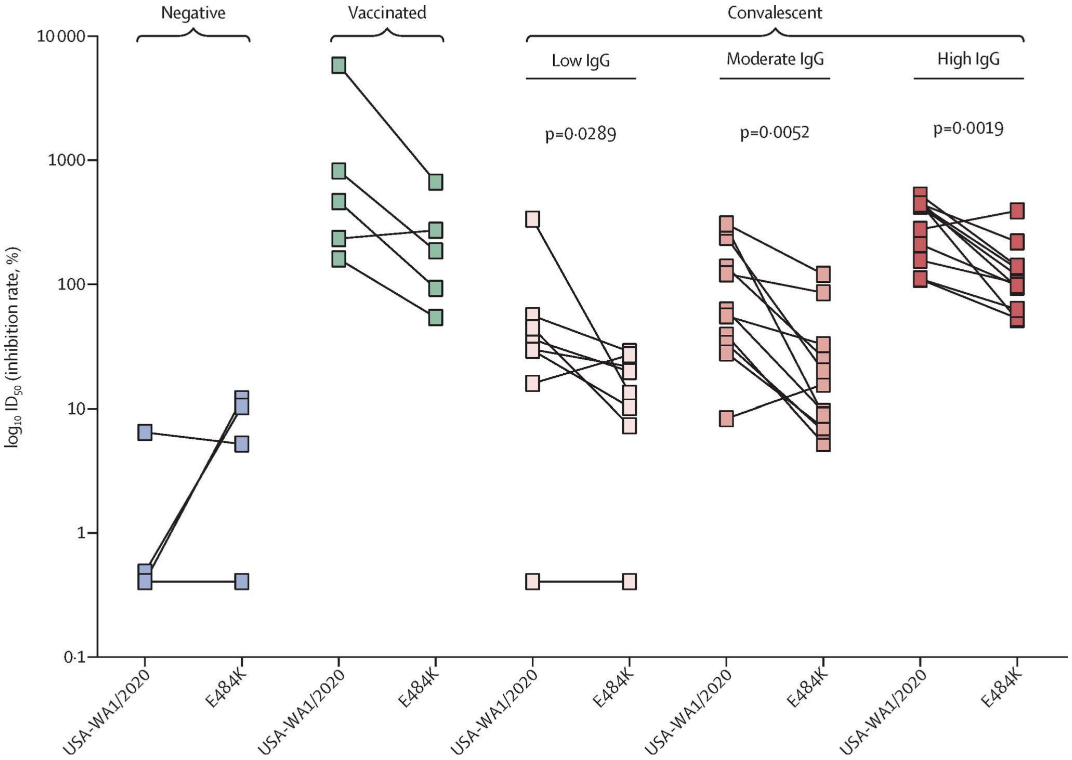
Identify clinically relevant mutations



Convergence(?)

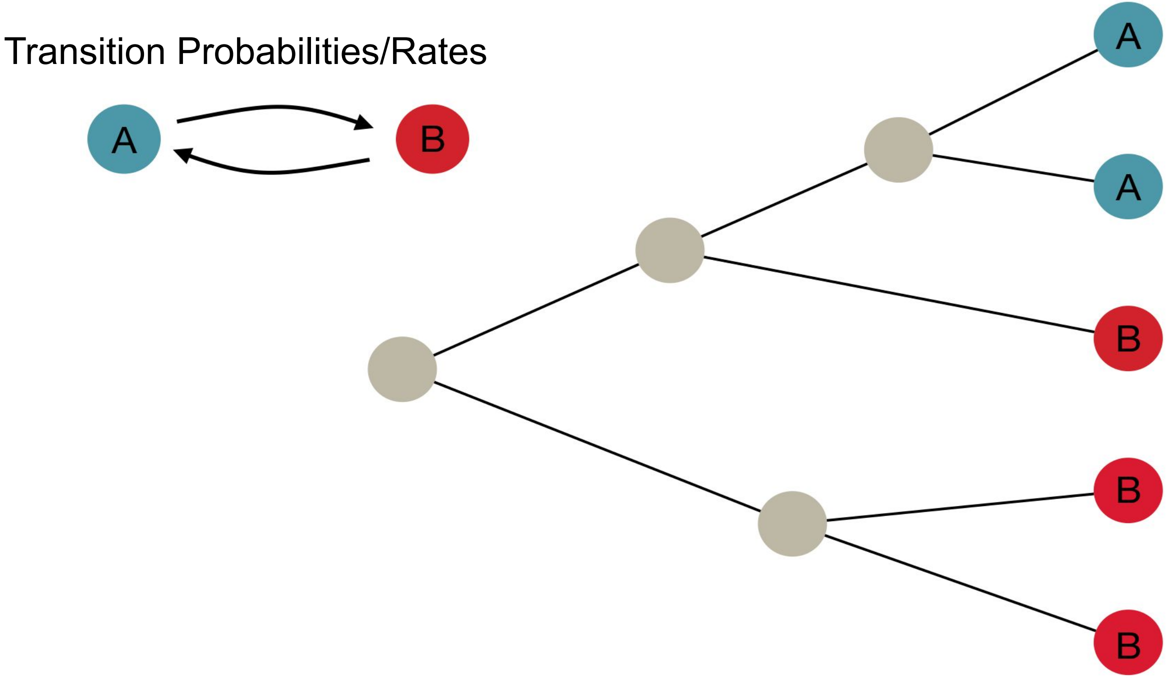
See part 2 for more details of testing these associations!

Prioritise characterisation of mutations: S:E484K



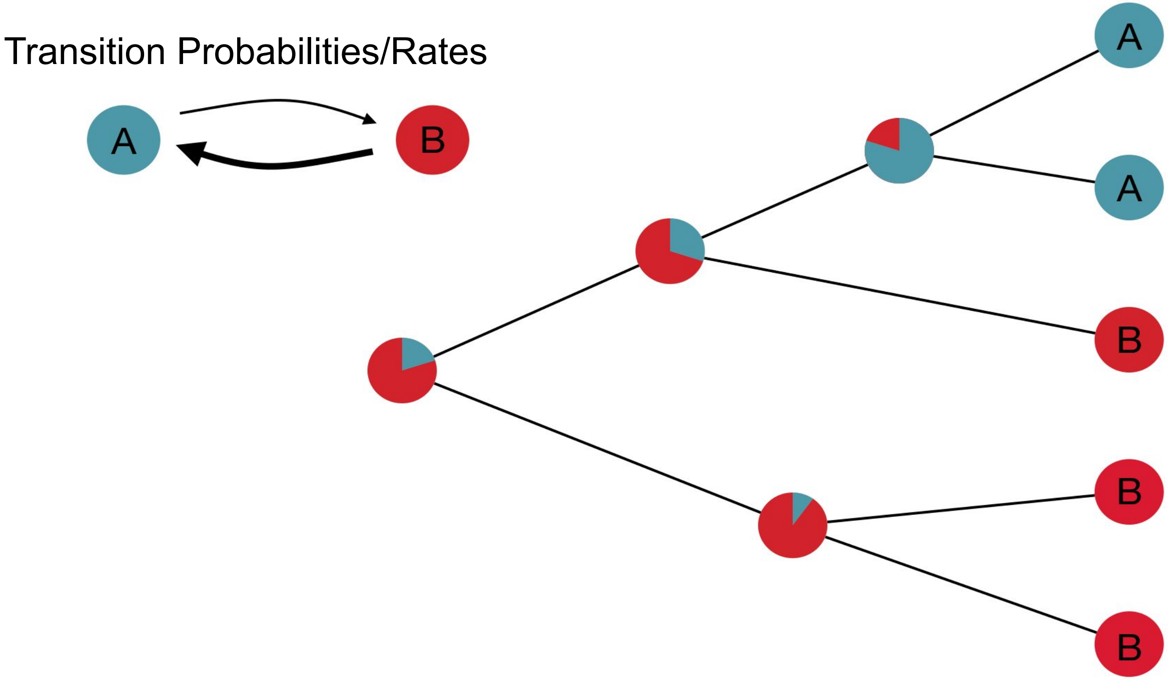
https://www.thelancet.com/journals/lanmic/article/PIIS2666-5247(21)00068-9

Inferring internal ancestral states from observed tips


Transition Probabilities/Rates

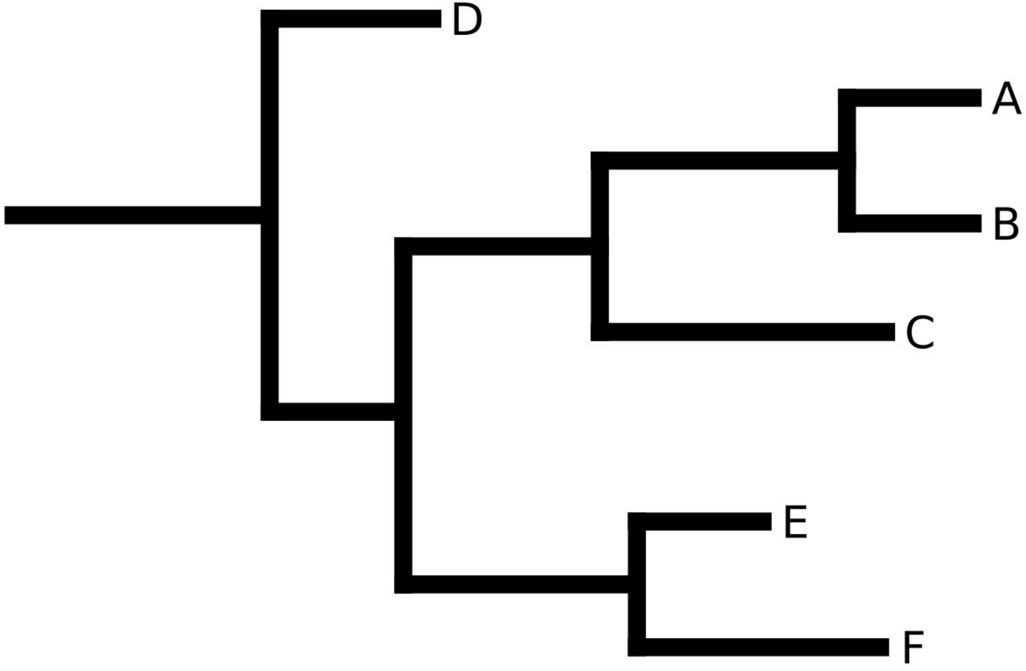
https://github.com/trvrb/gs541-phylodynamics

Inferring internal ancestral states from observed tips

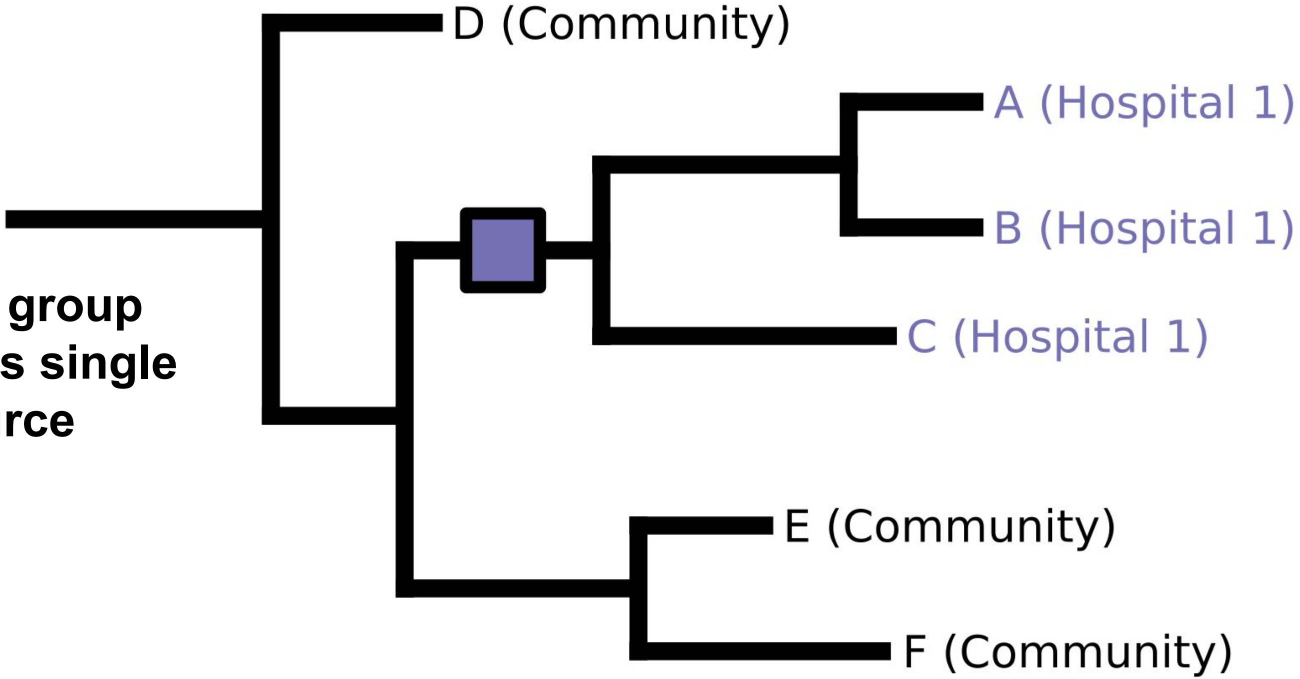


https://github.com/trvrb/gs541-phylodynamics

Trace sources of outbreaks

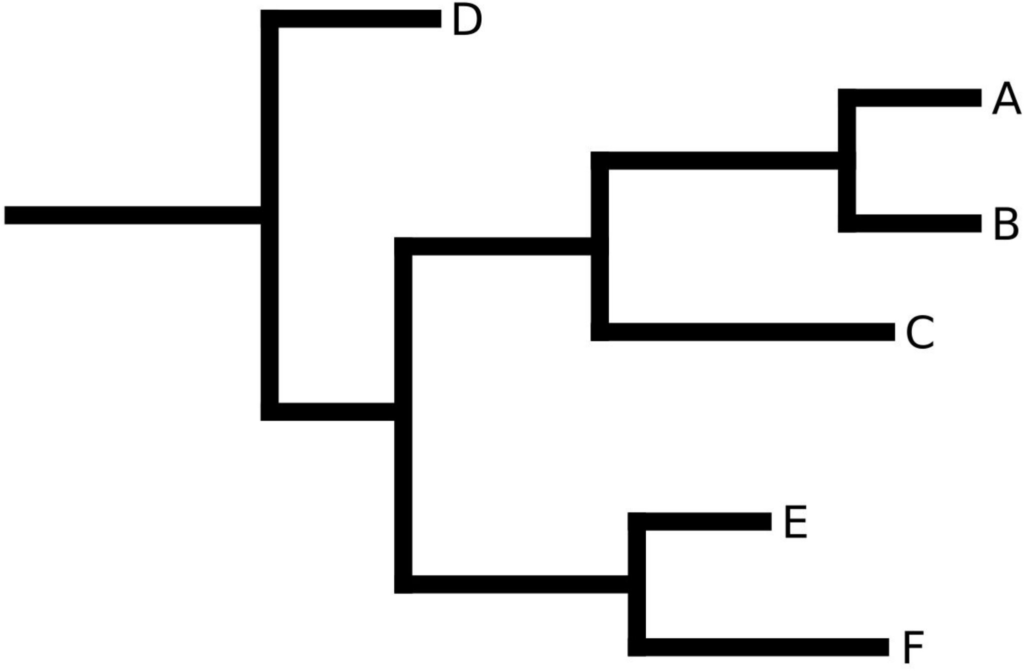


Trace sources of outbreaks

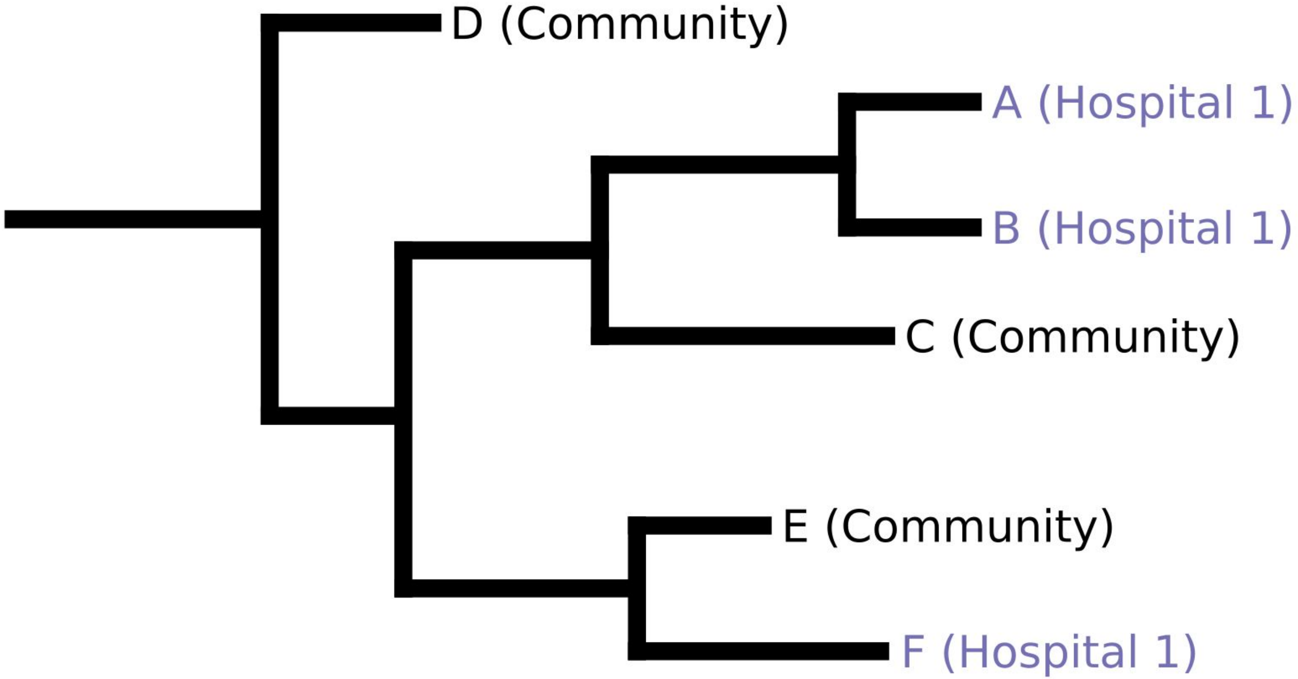


Single group suggests single source

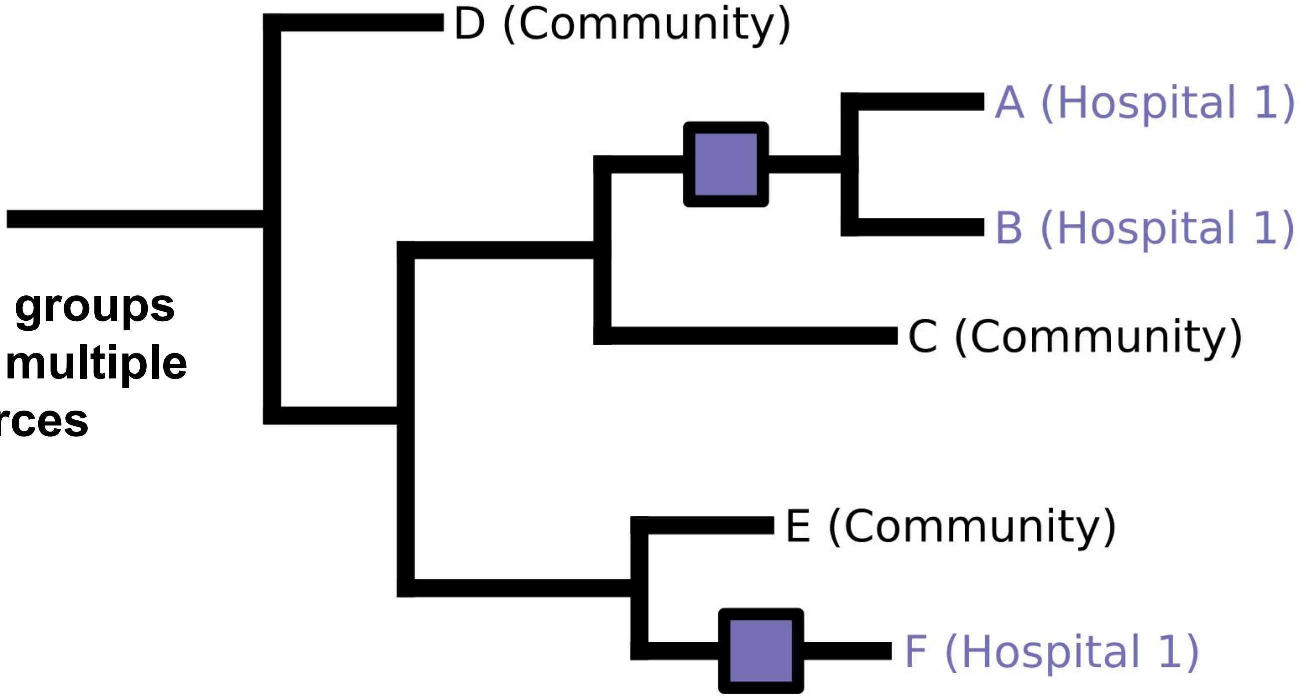
Trace sources of outbreaks



Trace sources of outbreaks

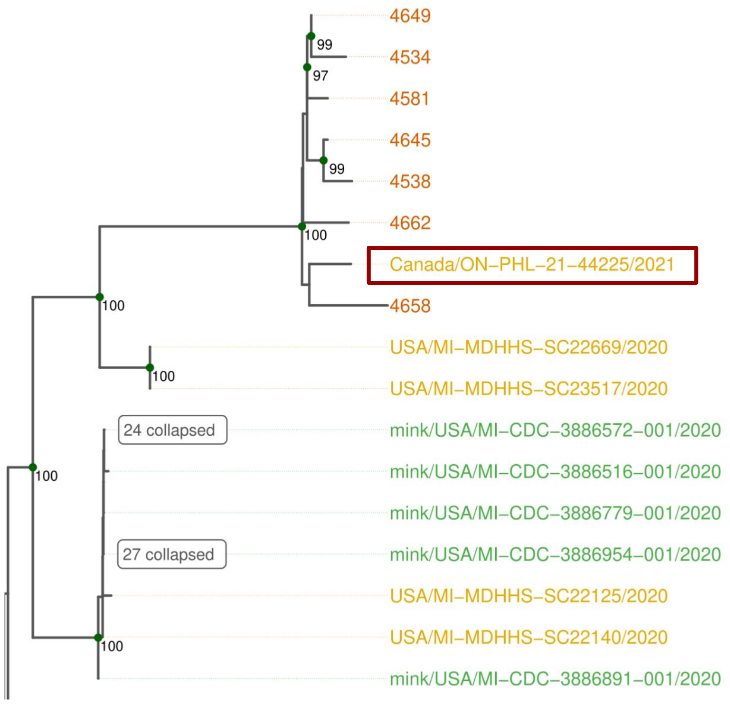


Trace sources of outbreaks



Multiple groups suggest multiple sources

Finding Deer-to-Human transmission



Host

🟨 *Homo sapien*

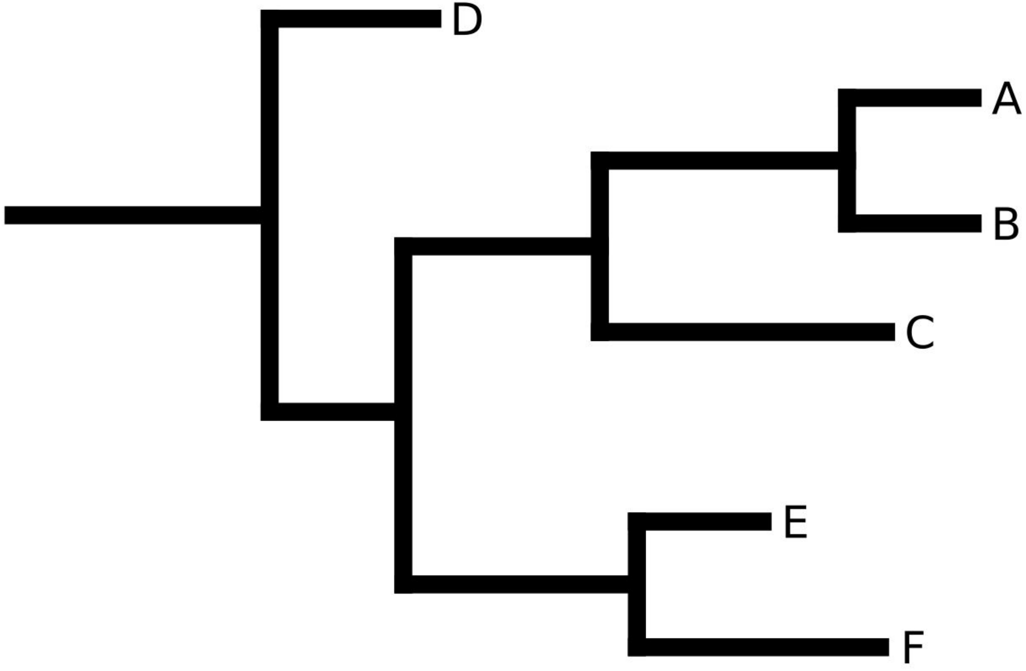
🟩 *Neovison vison*

🟧 *Odocoileus virginianus*

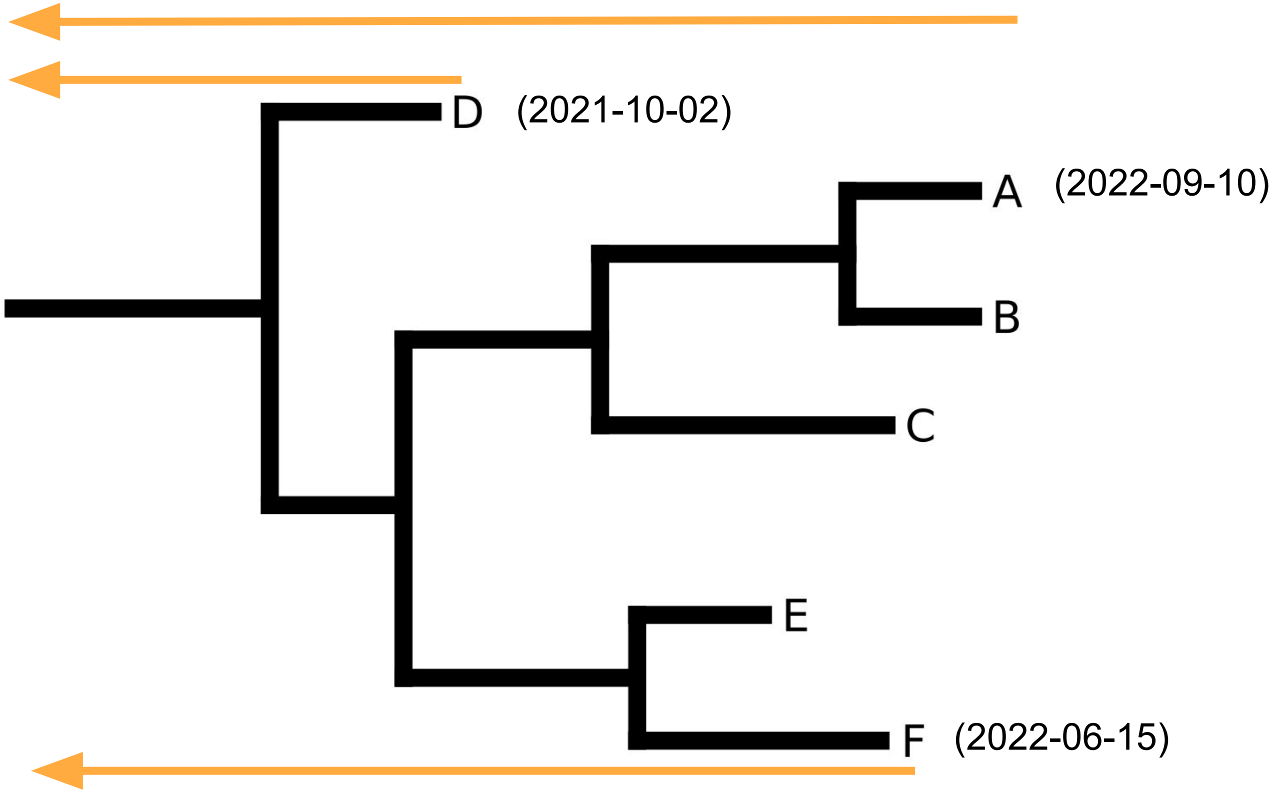
Supporting Circumstantial Evidence:

- Spatially congruent
- Temporally congruent
- Plausible Epidemiological Link

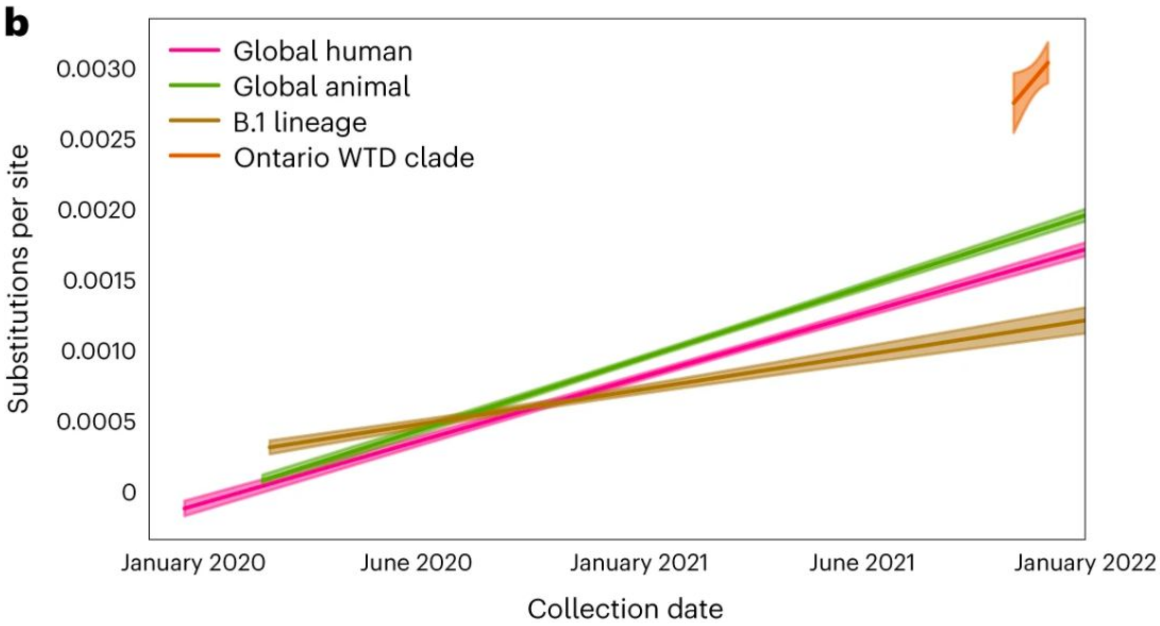
Convert genomic distance to time



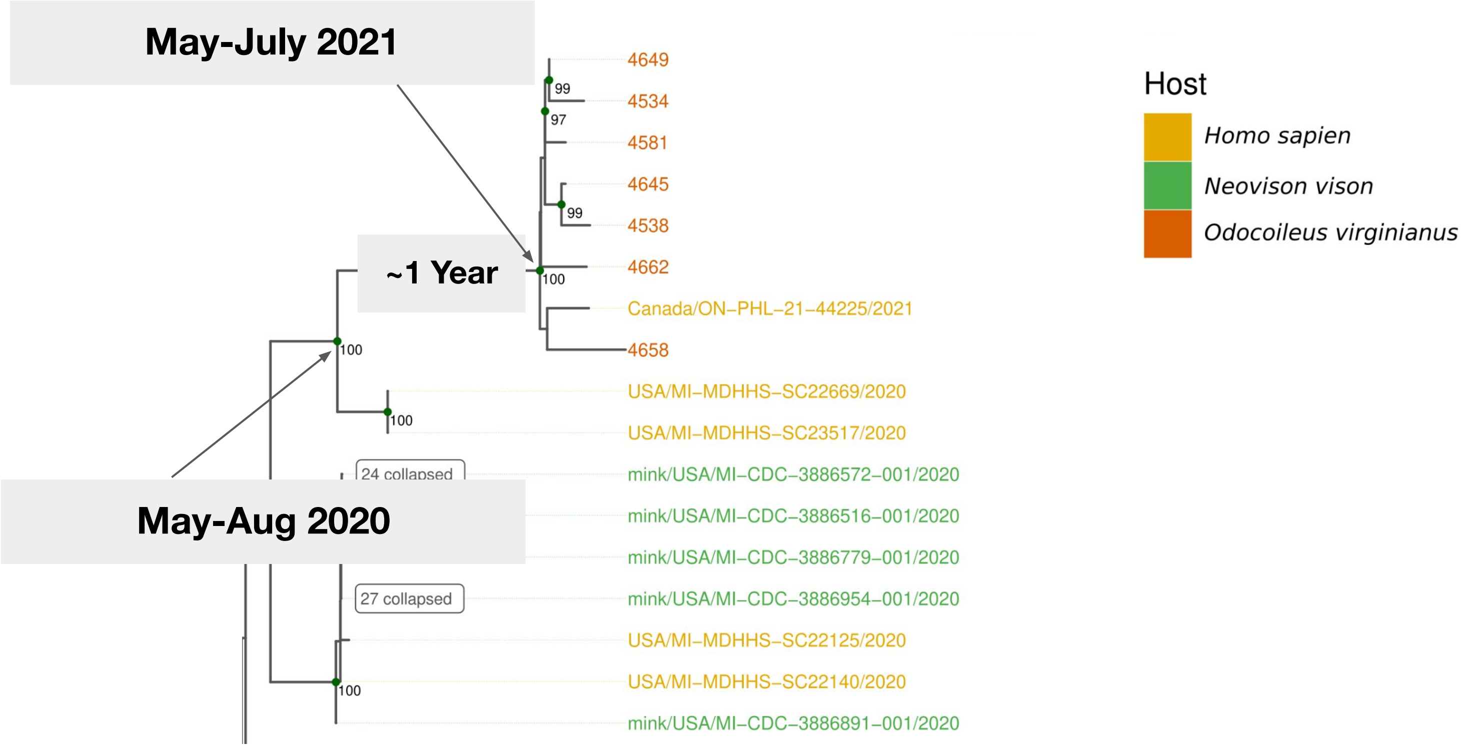
Convert genomic distance to time



Mutation rates from Root to Tip Regression



Mutations rates let us time of unobserved events

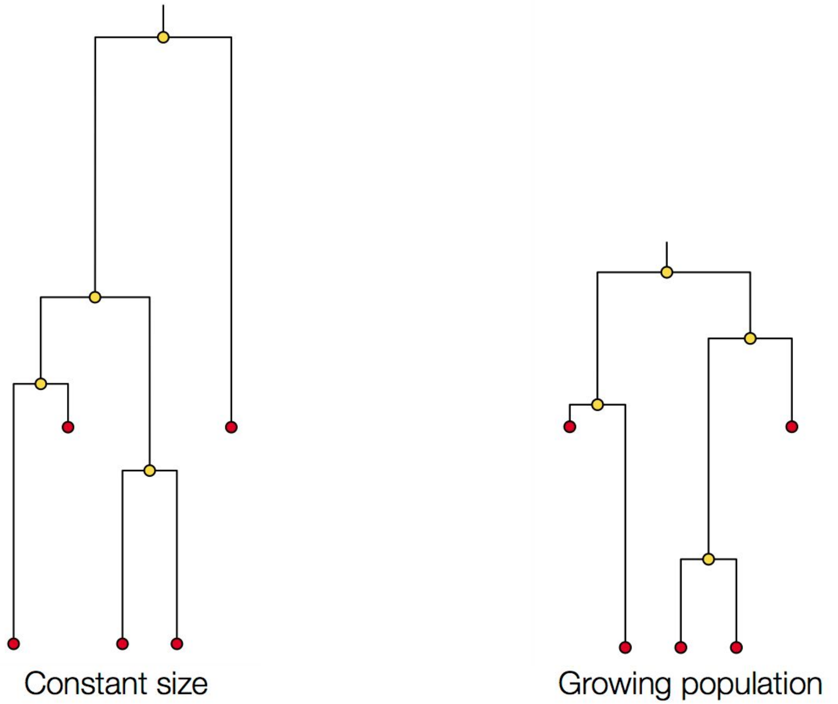


May-July 2021

~1 Year

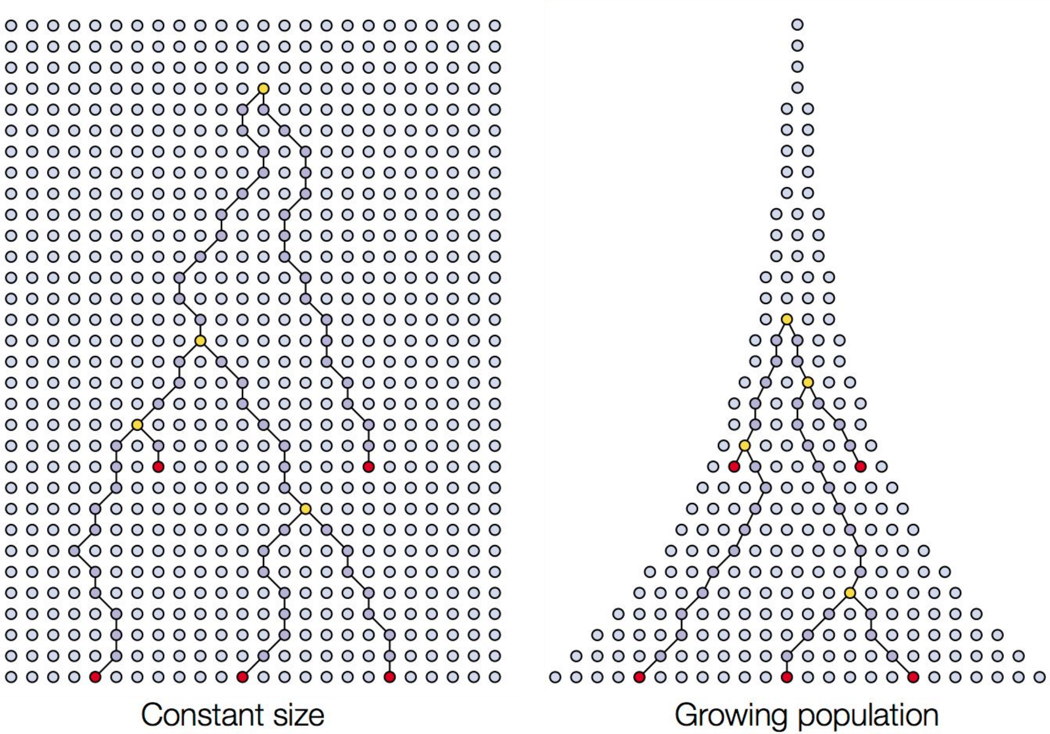
May-Aug 2020

Tree shape tells us about population size



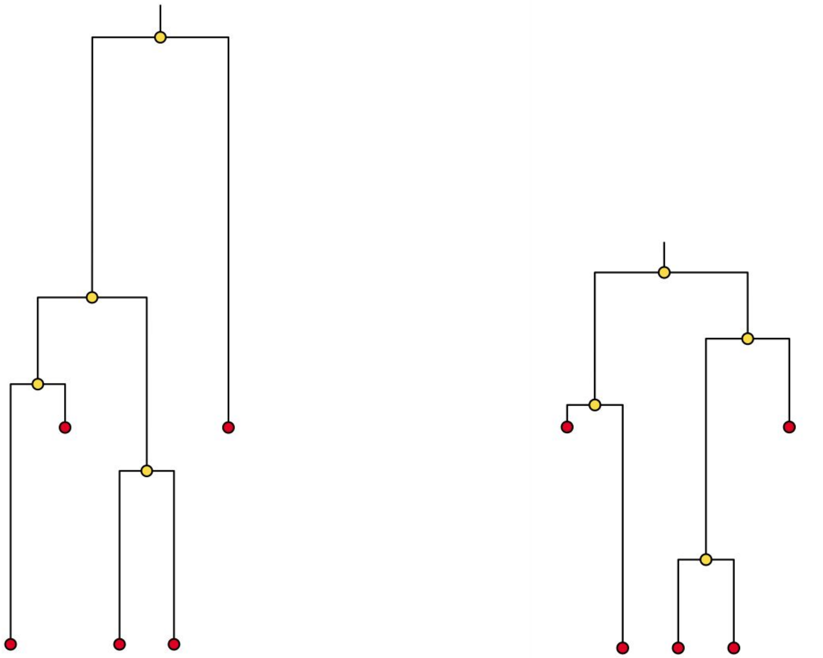
https://github.com/trvrb/gs541-phylodynamics

Tree shape tells us about population size



https://github.com/trvrb/gs541-phylodynamics

Tree shape tells us about population size

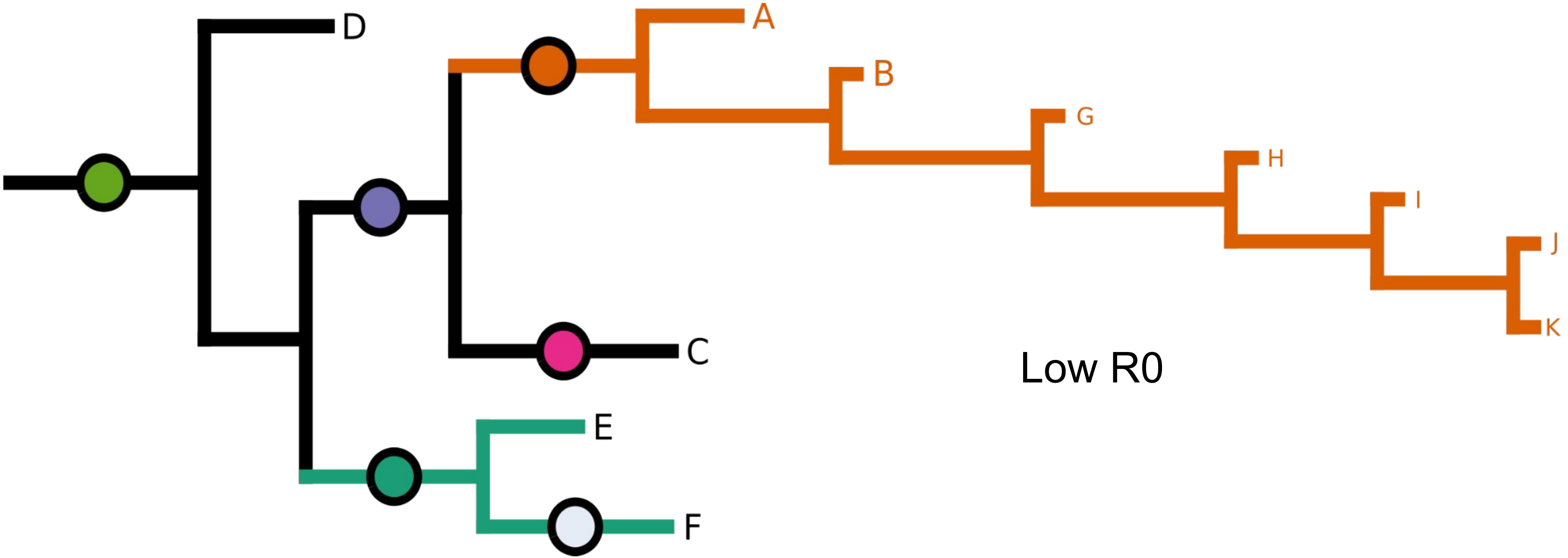


Constant size

Growing population

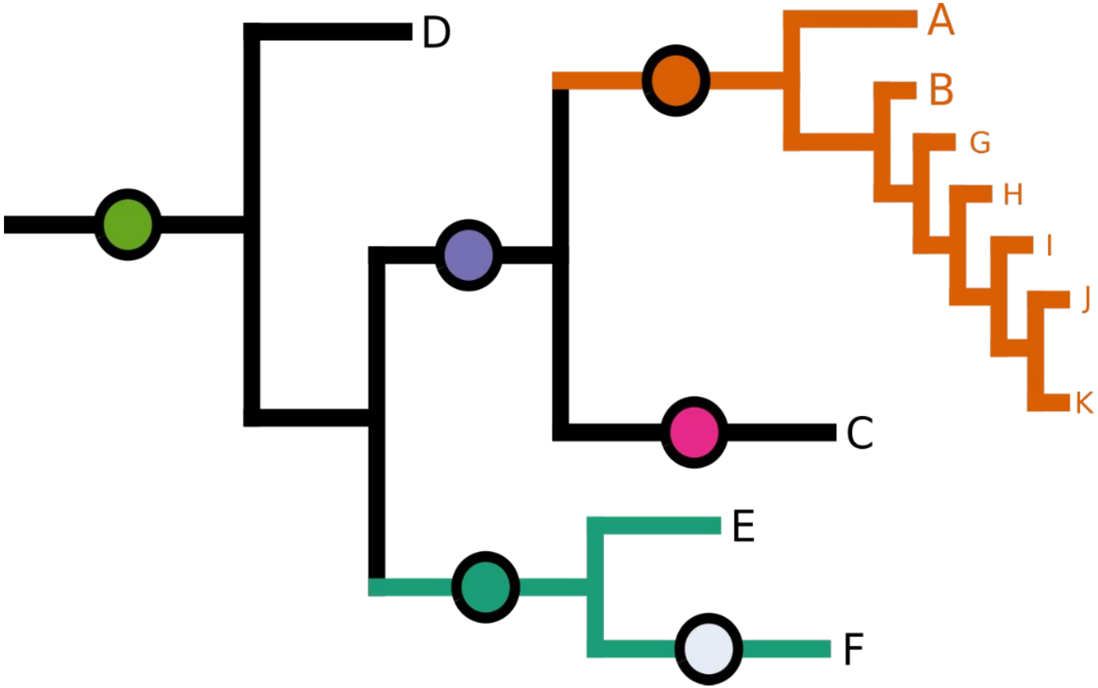
https://github.com/trvrb/gs541-phylodynamics

Inferring epidemiological parameters from shape



Low R0

Inferring epidemiological parameters from shape



High R0

What about evolutionary forces like selection?

dN/dS is one way to detect selection

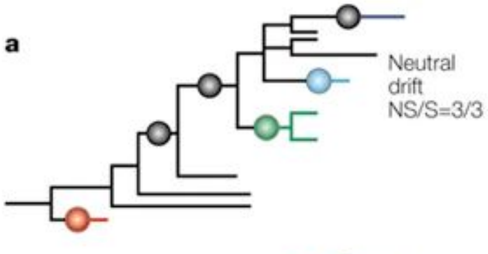
dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)

dN/dS is one way to detect selection

dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)



dN/dS ~ 1 : drift/neutral selection

Bush, R. Predicting adaptive evolution. *Nat Rev Genet* **2**, 387–392 (2001). https://doi.org/10.1038/35072023

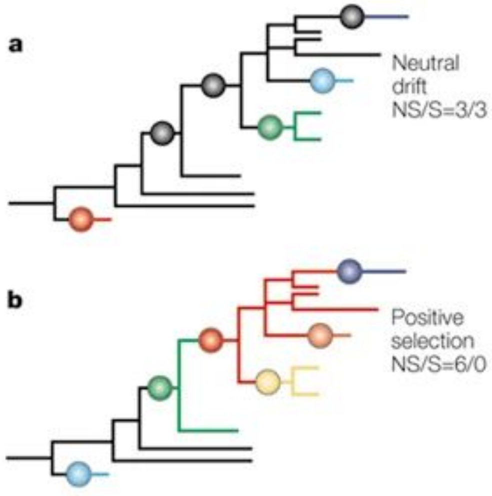
dN/dS is one way to detect selection

dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)

dN/dS > 1 : adaptive/positive selection

dN/dS ~ 1 : drift/neutral selection



Bush, R. Predicting adaptive evolution. *Nat Rev Genet* **2**, 387–392 (2001).
https://doi.org/10.1038/35072023

dN/dS is one way to detect selection

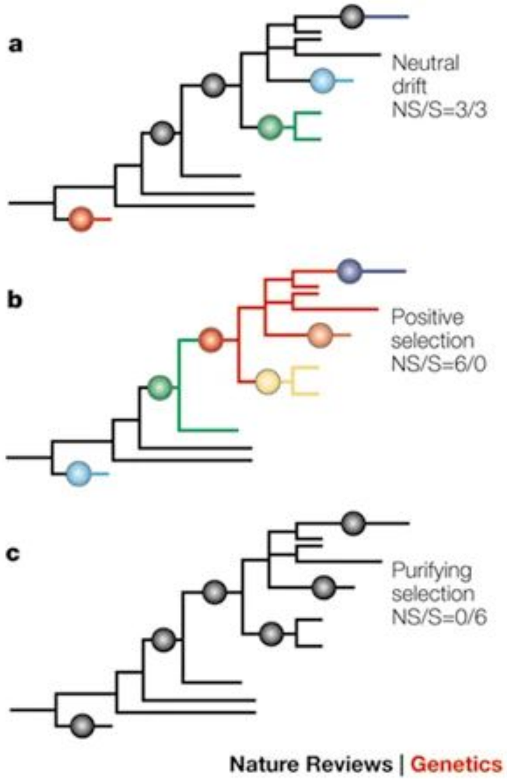
dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)

dN/dS > 1 : adaptive/positive selection

dN/dS ~ 1 : drift/neutral selection

dN/dS < 1: purifying/negative selection



Bush, R. Predicting adaptive evolution. *Nat Rev Genet* **2**, 387–392 (2001). https://doi.org/10.1038/35072023

dN/dS is one way to detect selection

dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)

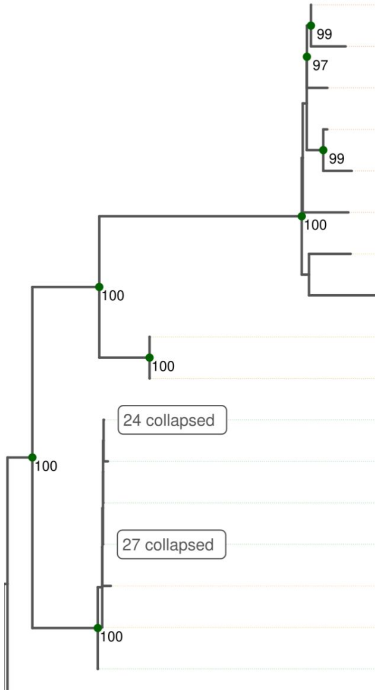
dN/dS > 1 : adaptive/positive selection

dN/dS ~ 1 : drift/neutral selection

dN/dS < 1: purifying/negative selection

Challenges:

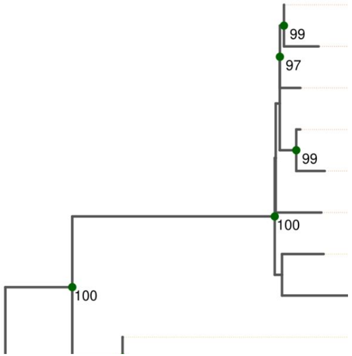
- Mutation rates vary over time/groups

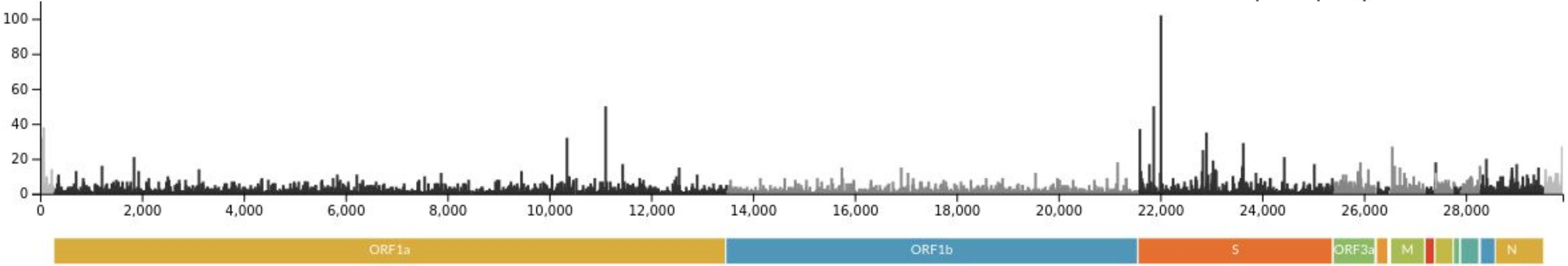


dN/dS is one way to detect selection

dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)





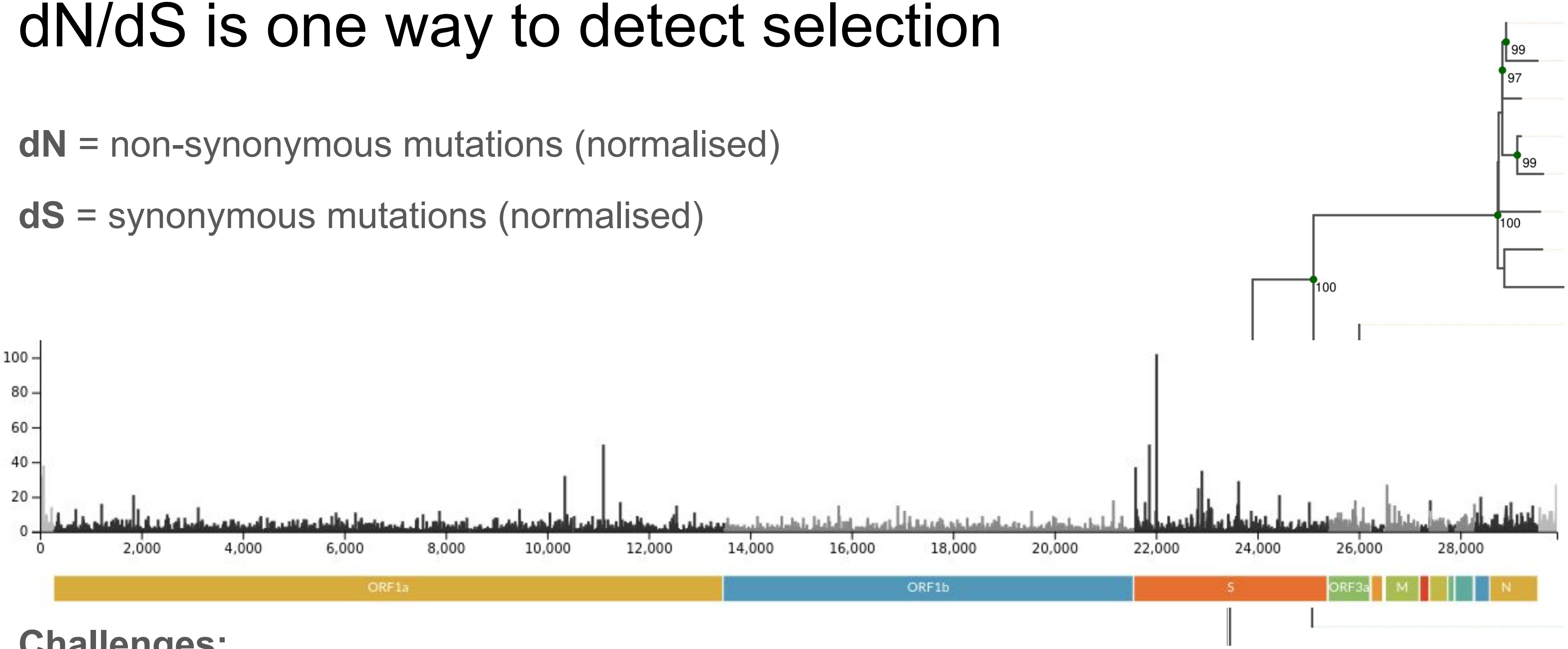
Challenges:

- Mutation rates vary over time/groups
- Mutation rates vary across genomes

dN/dS is one way to detect selection

dN = non-synonymous mutations (normalised)

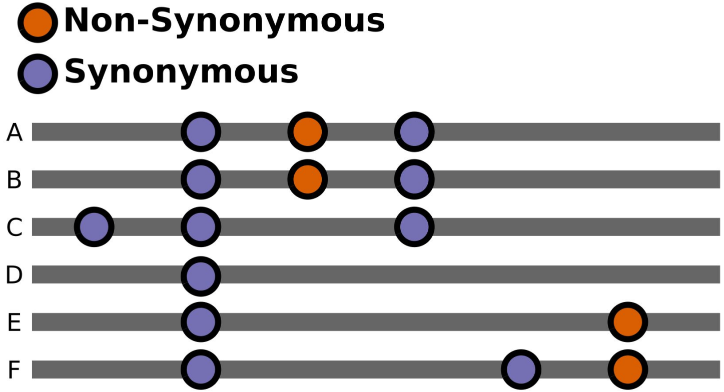
dS = synonymous mutations (normalised)



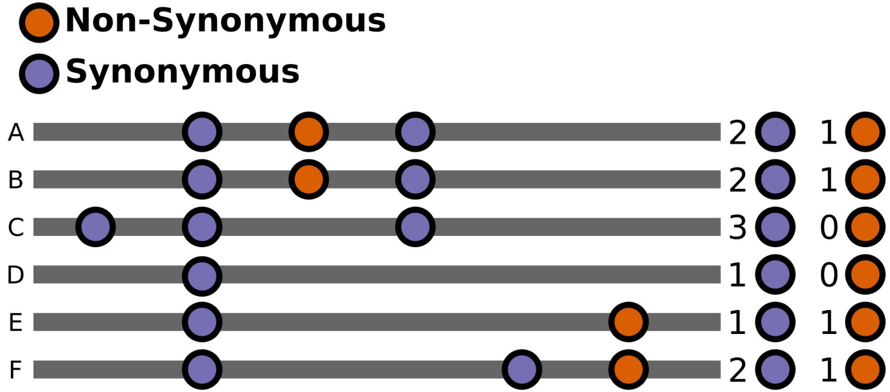
Challenges:

- Mutation rates vary over time/groups
- Mutation rates vary across genomes
- **Genomes are related** (mutations are non-independent)

Non-independence of events in related genomes



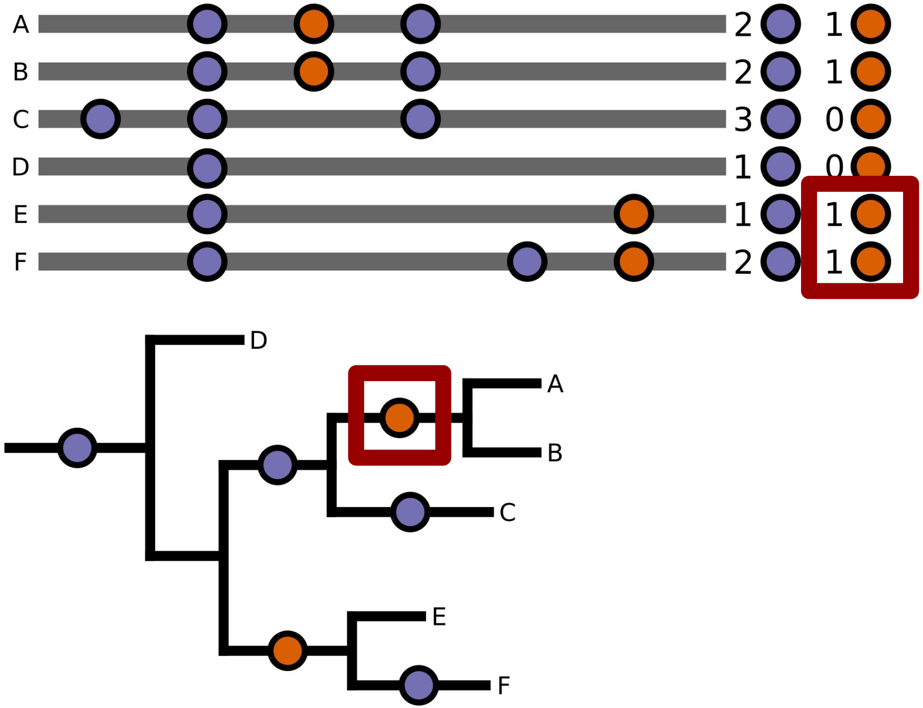
Non-independence of events in related genomes



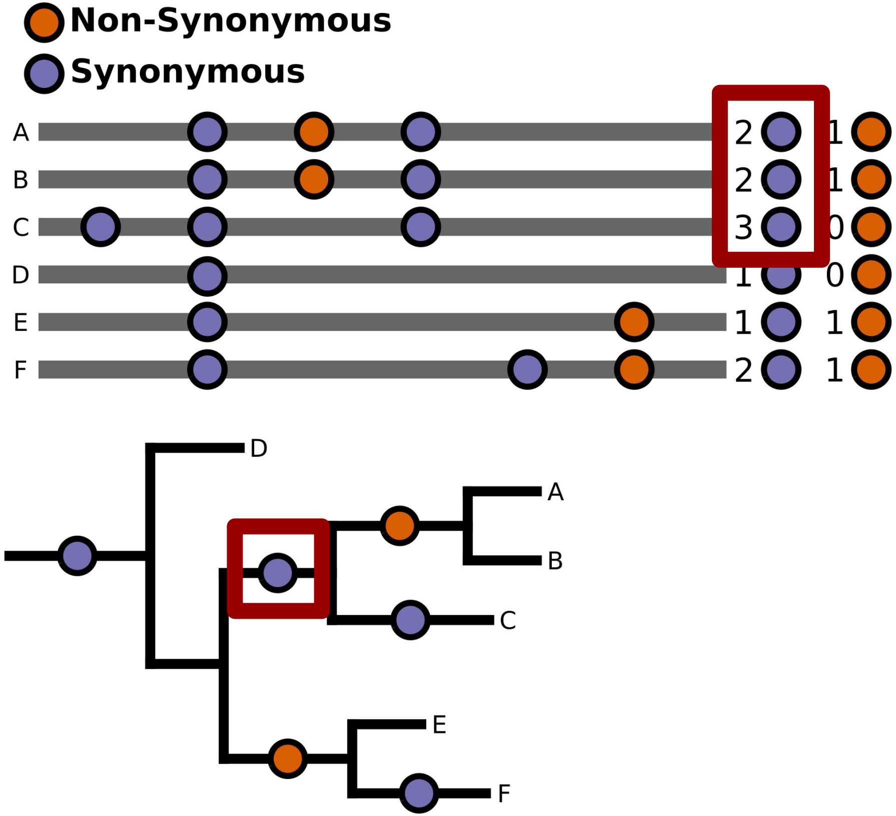
Non-independence of events in related genomes

● **Non-Synonymous**

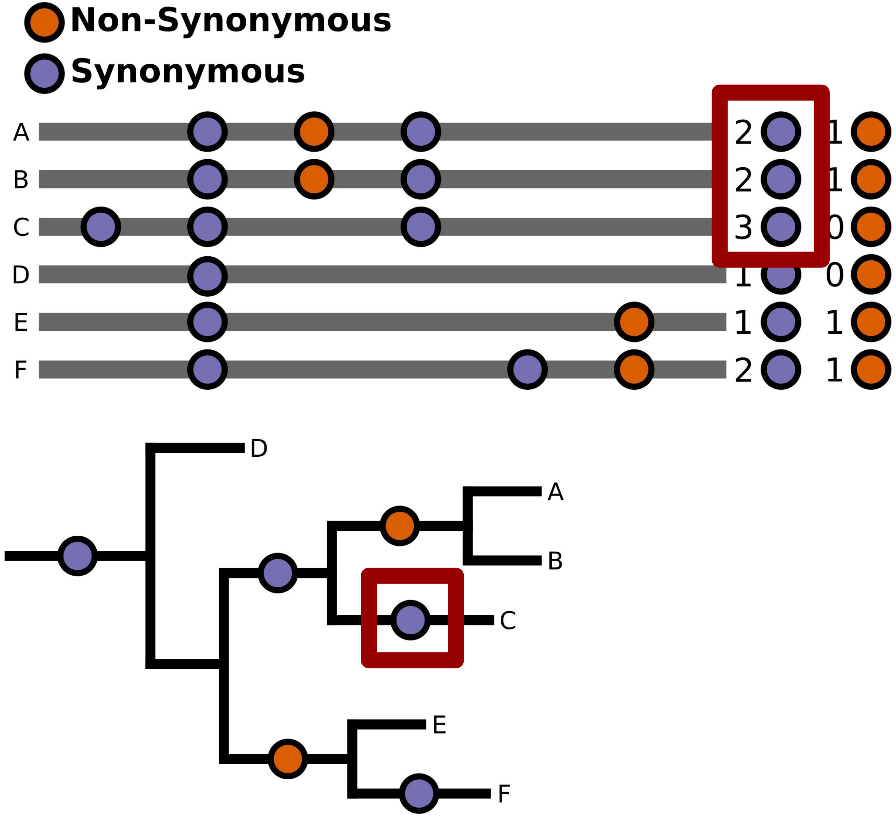
● **Synonymous**



Non-independence of events in related genomes

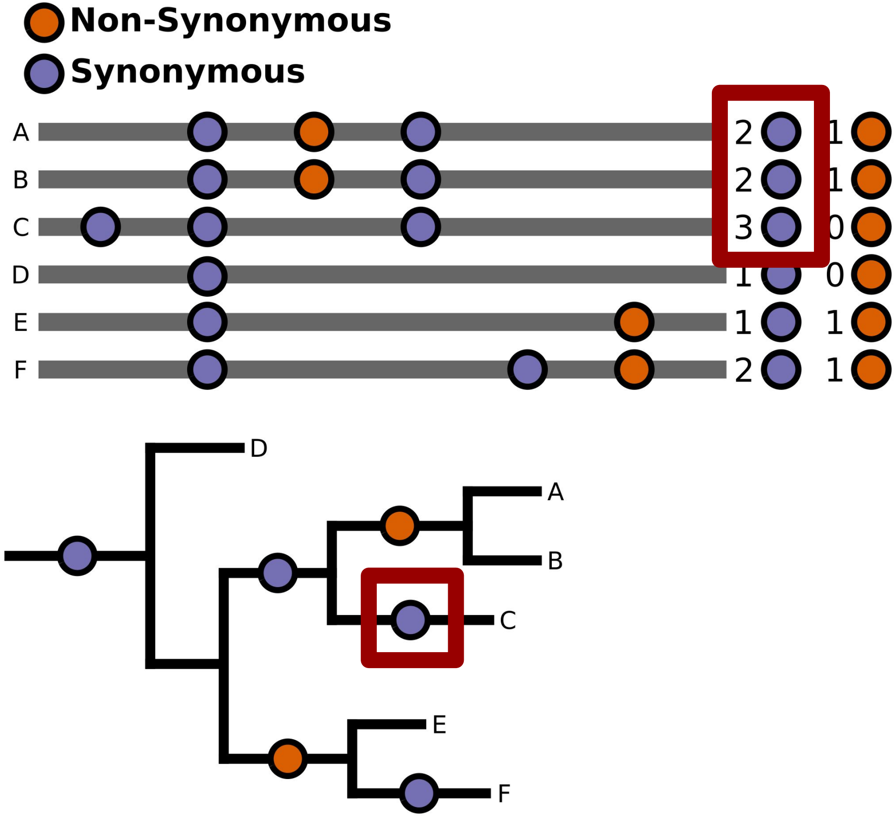


Non-independence of events in related genomes



- Phylogeny captures dependency structure of genomic data

Non-independence of events in related genomes

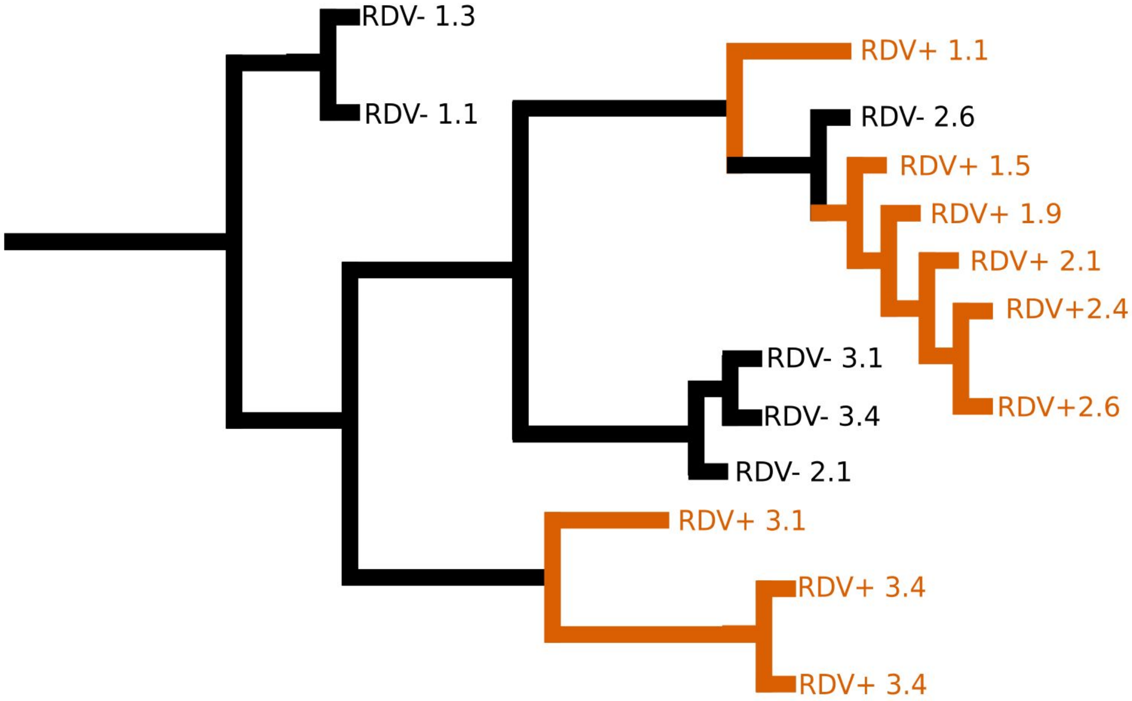


- Phylogeny captures dependency structure of genomic data
- Informs error term for models (e.g., regression)

- **a**daptive **B**ranch-**S**ite **R**andom **E**ffects **L**ikelihood: Is there a significant proportion of sites within selected branches with dN/dS > 1

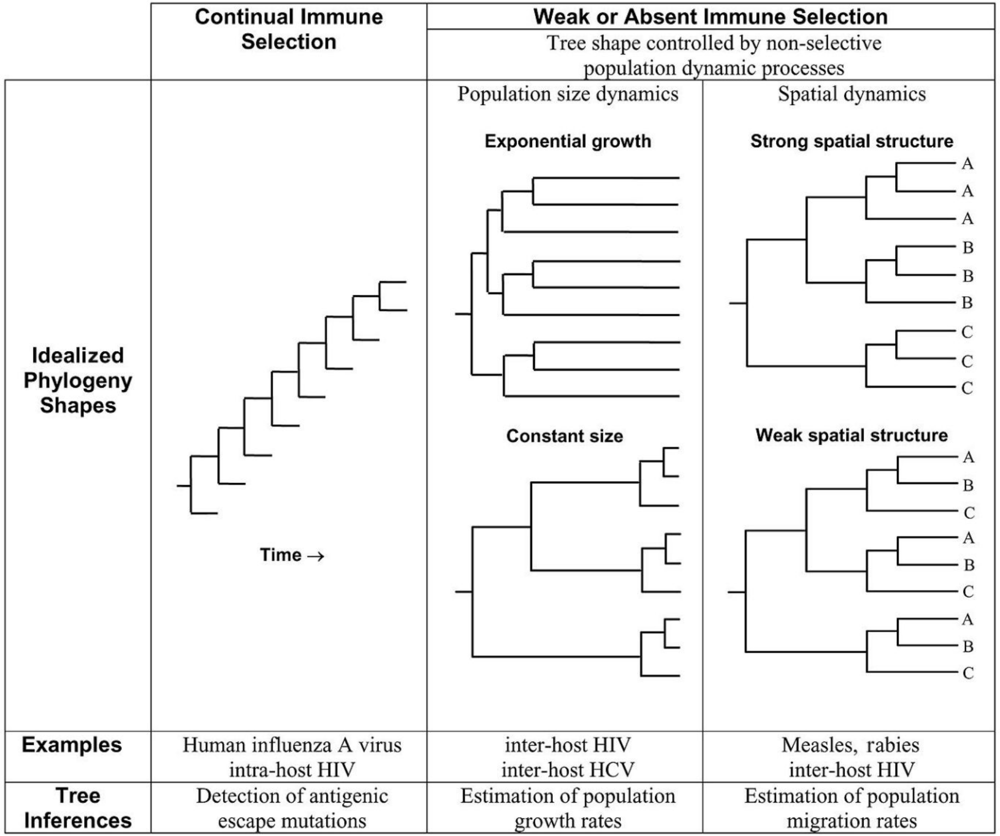
Smith, MD et al. "Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection." Mol. Biol. Evol. 32, 1342–1353 (2015).

Testing for remdesevir resistance selection



Nirmalarajah, Kuganya, et al. "Use of whole genome sequencing to identify low-frequency mutations in SARS-CoV-2 patients treated with remdesivir." *Influenza and Other Respiratory Viruses* 17.9 (2023): e13179.

Many other analyses are possible



	Continual Immune Selection	Weak or Absent Immune Selection	
		Tree shape controlled by non-selective population dynamic processes	
		Population size dynamics	Spatial dynamics
Idealized Phylogeny Shapes		Exponential growth Constant size	Strong spatial structure Weak spatial structure
Examples	Human influenza A virus intra-host HIV	inter-host HIV inter-host HCV	Measles, rabies inter-host HIV
Tree Inferences	Detection of antigenic escape mutations	Estimation of population growth rates	Estimation of population migration rates

Massive area:

- Birth-death models
- Coalescent models
- Bayesian skyline/skygrid models
- Spatiotemporal models (phylogeography)
- Recombination

Grenfell, B. T. (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science, 303(5656), 327–332. doi:10.1126/science.1090727

Summary

- Pathogen **evolution** and **epidemiology** are intrinsically linked
- Genomics provides insights into **evolution and unobserved events**
- Comparison of DNA sequences to databases can be used for **diagnostics**
- Pattern of mutations across genomes can be used to generate **phylogenies**
- Phylogenies are structured by **sampling**, **ecology, evolution, and epidemiology**
- Probabilistic **Bayesian phylogenetic inference** is a key tool
- Can use these approaches to do many things including:
    - Identify **lineages**
    - Monitor **evolution**
    - Infer **timing/location** of outbreaks/events
    - Determine **epidemiological parameters**
    - Characterise strength and direction of **selection**