MICI3119 - Genomics I

Finlay Maguire
finlay.maguire@dal.ca
maguire-lab.github.io

Overview

- Hospital-based outbreak prevention and control
- DNA Sequencing Technologies
- Reference-based assembly
- De novo assembly
- Inferring phylogenies from genomes
- Interpreting phylogenies
- Use genomic data to respond to stop an outbreak

IPAC Investigations: UCSF NICU MRSA Outbreak

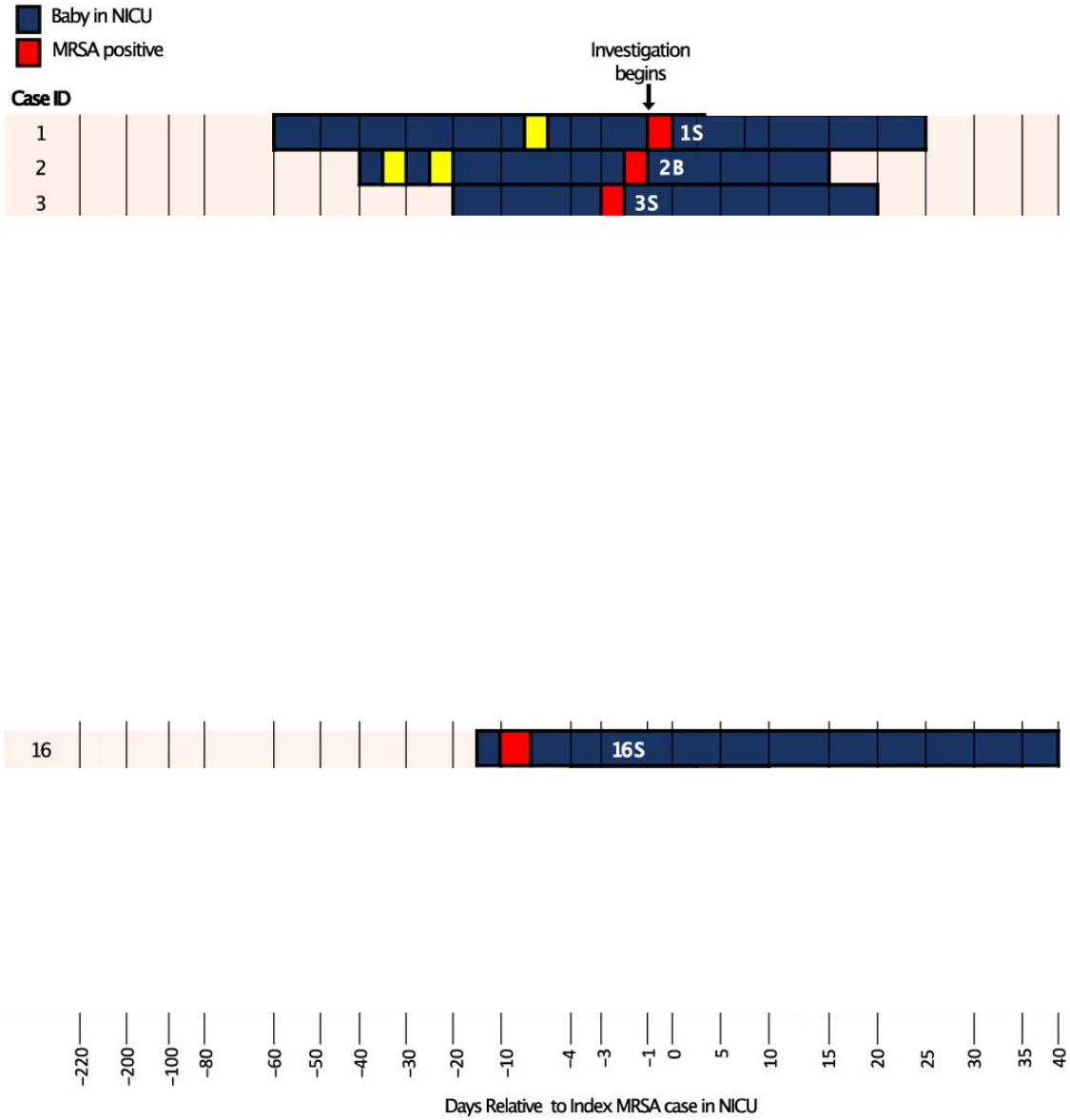
Neonatal Intensive Care Unit

https://healthtian.com/wp-content/uploads/2020/09/neonatal-intensive-care-unit1.jpg

Setting: UCSF NICU (58-bed) in UCSF Benioff Children's Hospital (183-bed)

Madera, Sharline, et al. "Prolonged silent carriage, genomic virulence potential and transmission between staff and patients characterize a neonatal intensive care unit (NICU) outbreak of methicillin-resistant Staphylococcus aureus (MRSA)." Infection Control & Hospital Epidemiology 44.1 (2023): 40-46.

IPAC Investigations: UCSF NICU MRSA Outbreak

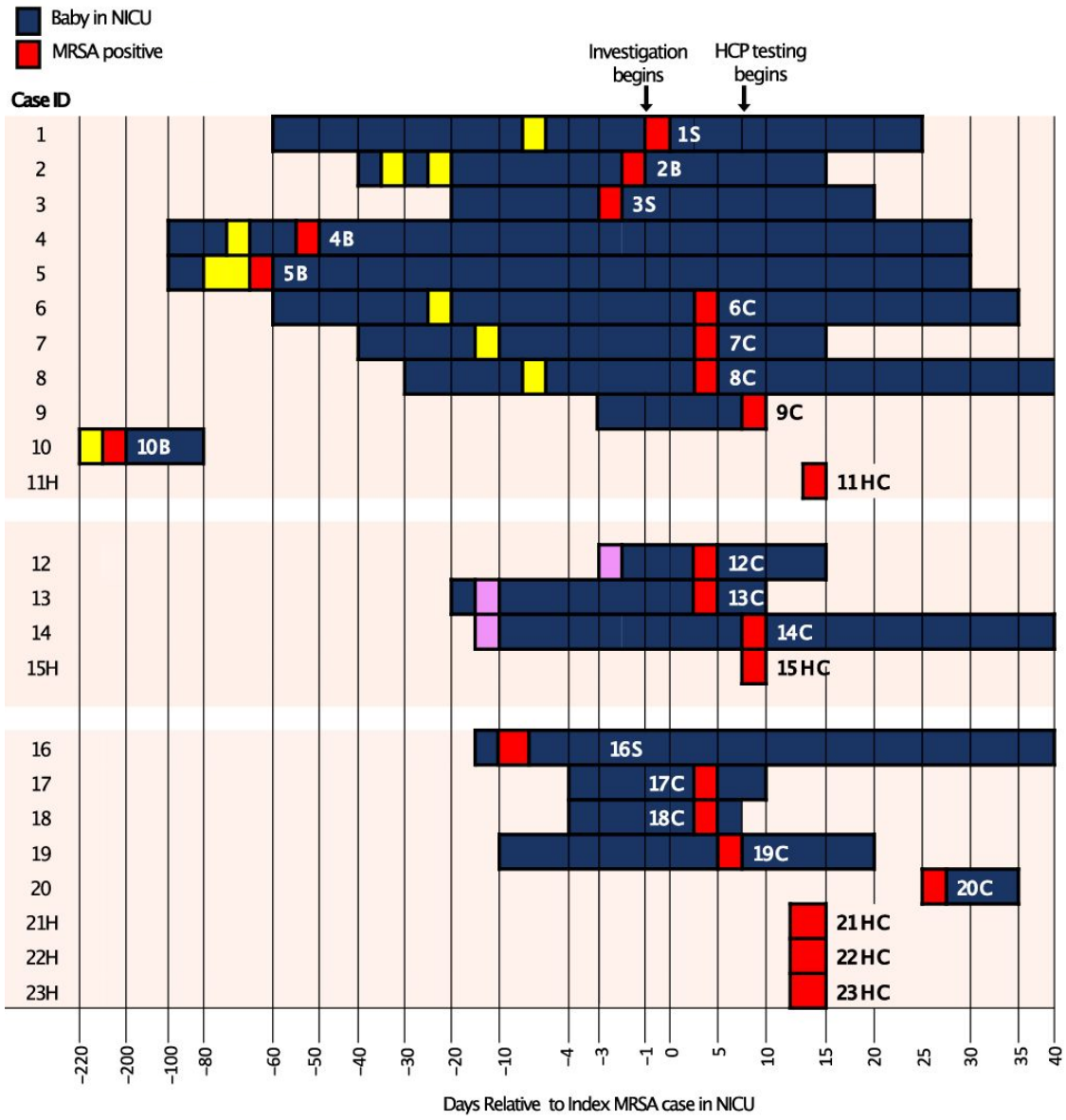


Setting: UCSF NICU (58-bed) in UCSF Benioff Children's Hospital (183-bed)

Investigation Trigger: 4 NICU infants with invasive MRSA within 8 days

Madera, Sharline, et al. "Prolonged silent carriage, genomic virulence potential and transmission between staff and patients characterize a neonatal intensive care unit (NICU) outbreak of methicillin-resistant Staphylococcus aureus (MRSA)." Infection Control & Hospital Epidemiology 44.1 (2023): 40-46.

IPAC Investigations: UCSF NICU MRSA Outbreak



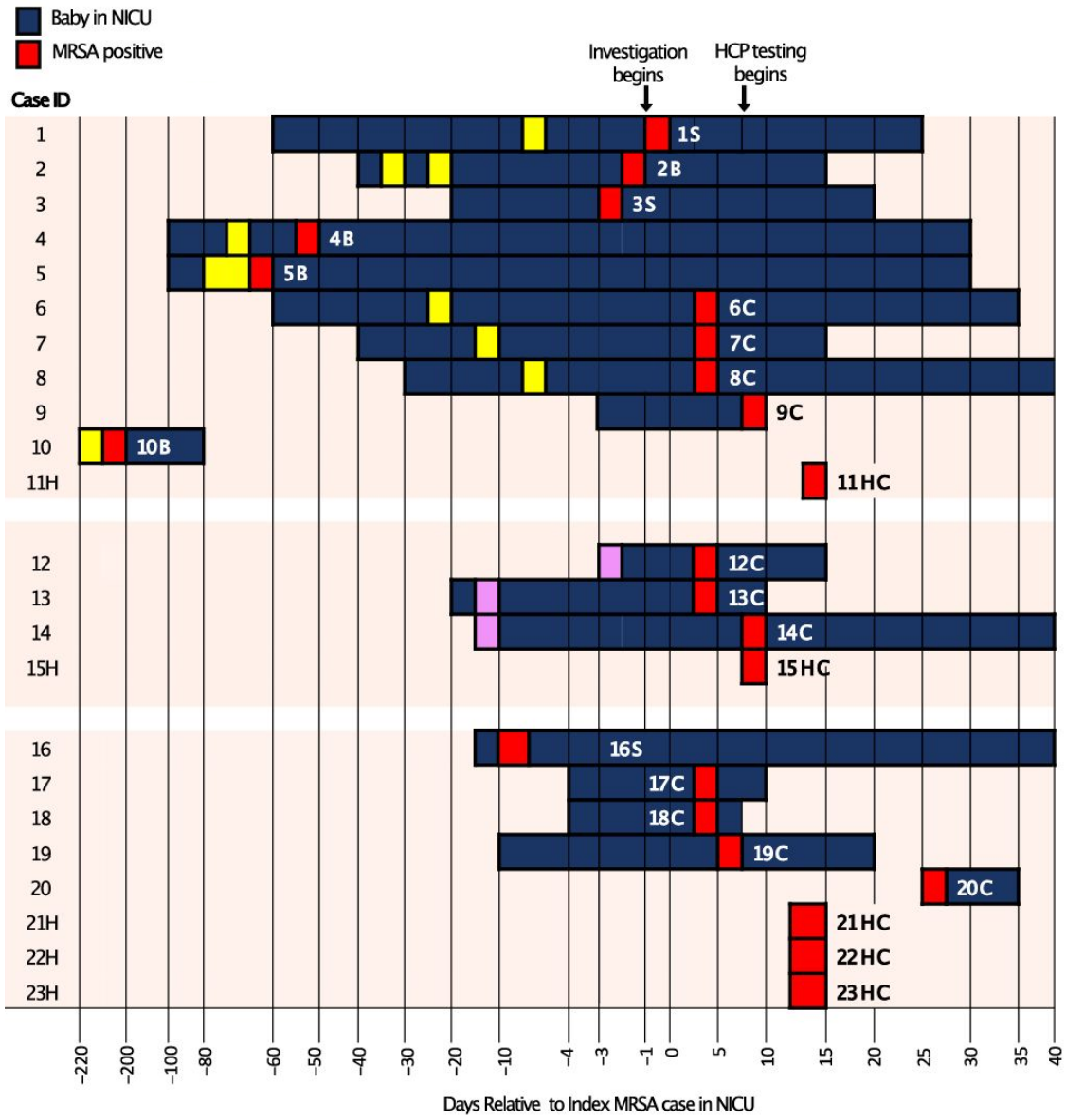
Setting: UCSF NICU (58-bed) in UCSF Benioff Children's Hospital (183-bed)

Investigation Trigger: 4 NICU infants with invasive MRSA within 8 days

Next Step: Look at recent past MRSA cases, contact trace, & expand surveillance swabbing of NICU + HCPs

Madera, Sharline, et al. "Prolonged silent carriage, genomic virulence potential and transmission between staff and patients characterize a neonatal intensive care unit (NICU) outbreak of methicillin-resistant Staphylococcus aureus (MRSA)." Infection Control & Hospital Epidemiology 44.1 (2023): 40-46.

IPAC Investigations: UCSF NICU MRSA Outbreak



Setting: UCSF NICU (58-bed) in UCSF Benioff Children's Hospital (183-bed)

Investigation Trigger: 4 NICU infants with invasive MRSA within 8 days

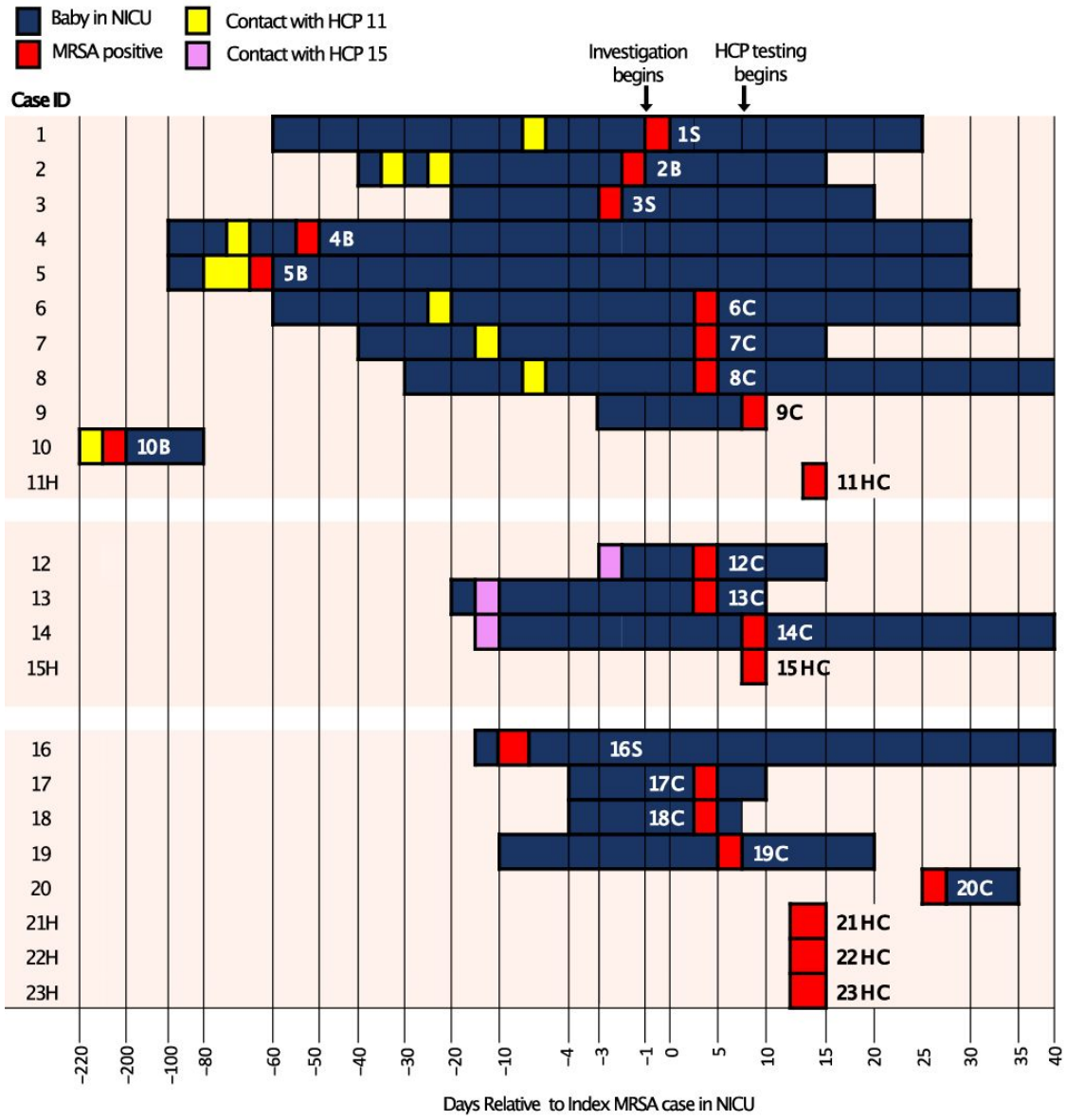
Next Step: Look at recent past MRSA cases, contact trace, & expand surveillance swabbing of NICU + HCPs

Which cases are linked?

Where did these come from?

Madera, Sharline, et al. "Prolonged silent carriage, genomic virulence potential and transmission between staff and patients characterize a neonatal intensive care unit (NICU) outbreak of methicillin-resistant Staphylococcus aureus (MRSA)." Infection Control & Hospital Epidemiology 44.1 (2023): 40-46.

IPAC Investigations: UCSF NICU MRSA Outbreak



Setting: UCSF NICU (58-bed) in UCSF Benioff Children's Hospital (183-bed)

Investigation Trigger: 4 NICU infants with invasive MRSA within 8 days

Next Step: Look at recent past MRSA cases, contact trace, & expand surveillance swabbing of NICU + HCPs

Which cases are linked?

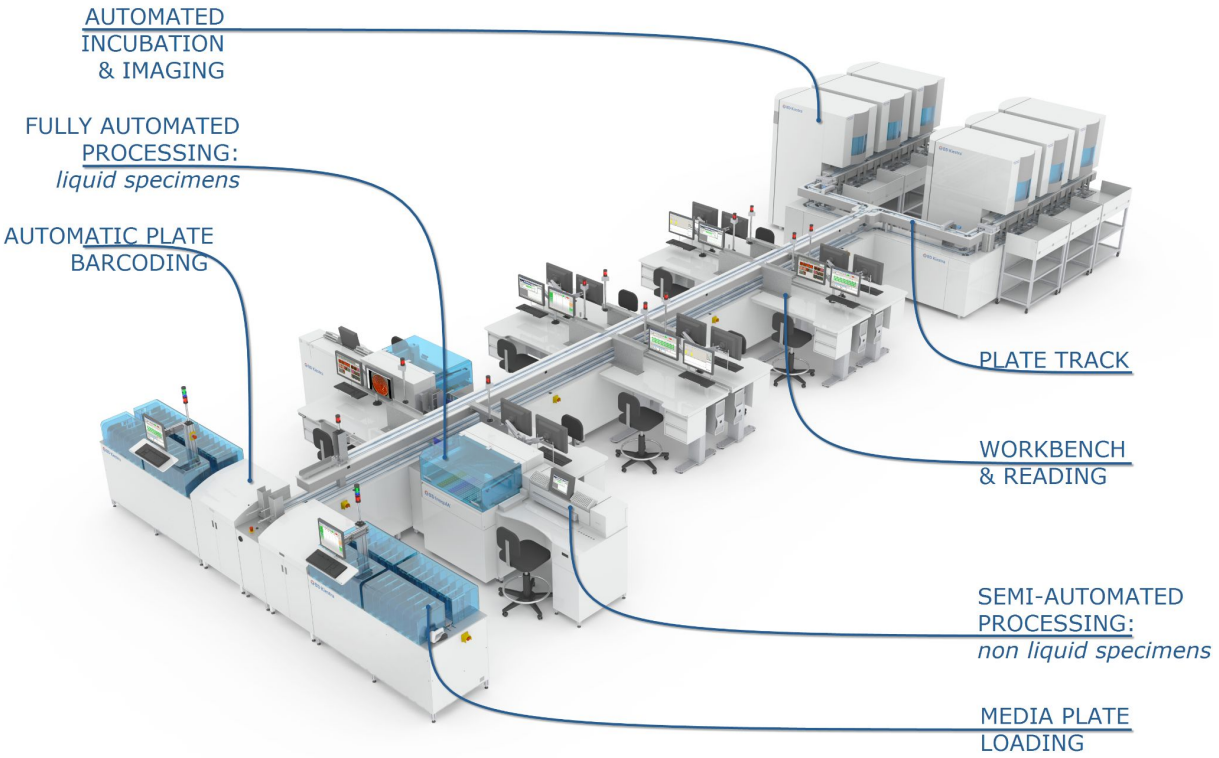
Where did these come from?

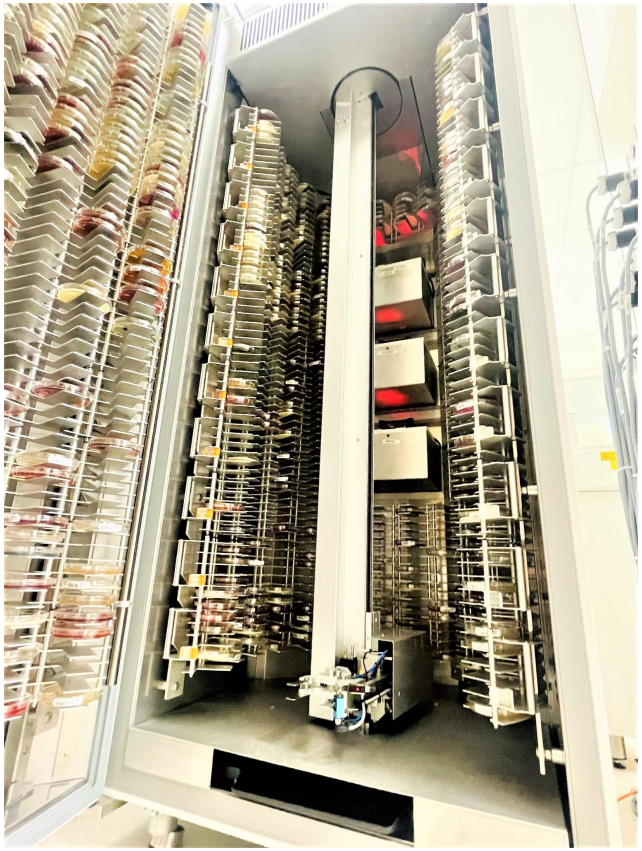
Contact tracing gives ideas but need more information => Genomes

Madera, Sharline, et al. "Prolonged silent carriage, genomic virulence potential and transmission between staff and patients characterize a neonatal intensive care unit (NICU) outbreak of methicillin-resistant Staphylococcus aureus (MRSA)." Infection Control & Hospital Epidemiology 44.1 (2023): 40-46.

So, we have lots of swabs, what now?

~~Expensively~~ Automated sample processing and culturing





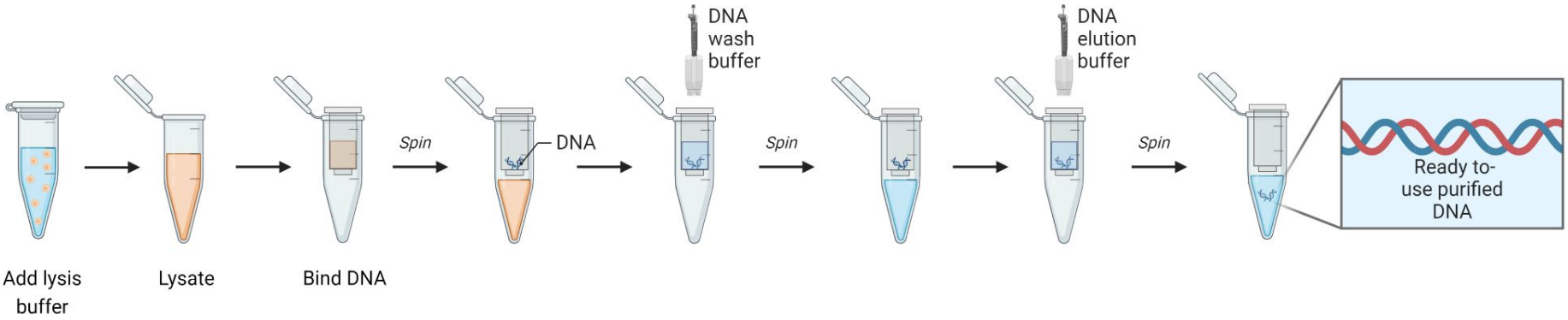
https://www.bd.com/scripts/europe/labautomation/productsdrilldown.asp?CatID=455&SubID=1836&siteID=20309&d=&s=europe%2Flabautomation&sTitle=Lab+Automation&metaTitle=Total+Lab+Automation&dc=europe&dcTitle=Europe

Christy Vermeiren

~~Expensively~~ Automated sample processing and culturing

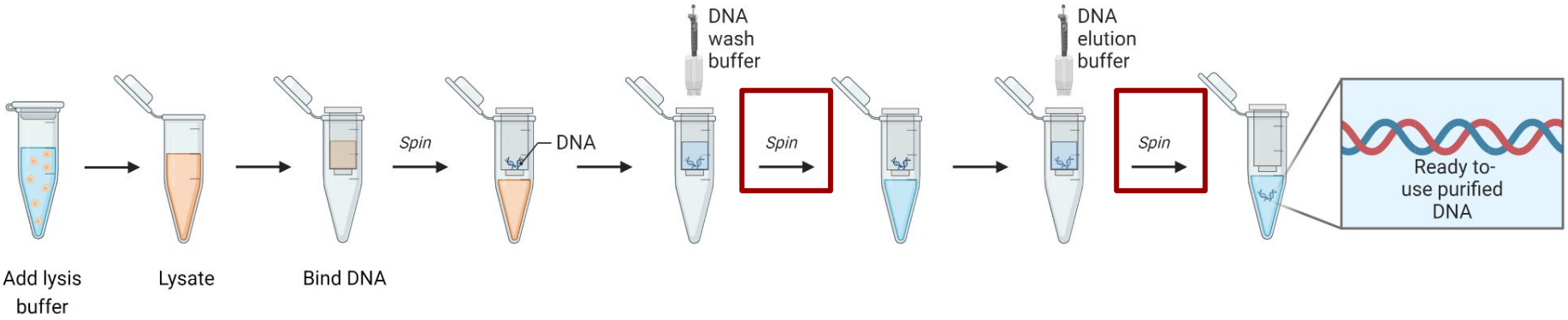
Christy Vermeiren

Automated DNA extraction with magnets + robots



https://images.aatbio.com/universal/newsletter/volume-11-1/Silica%20Gel%20DNA%20Extraction%20Workflow.png

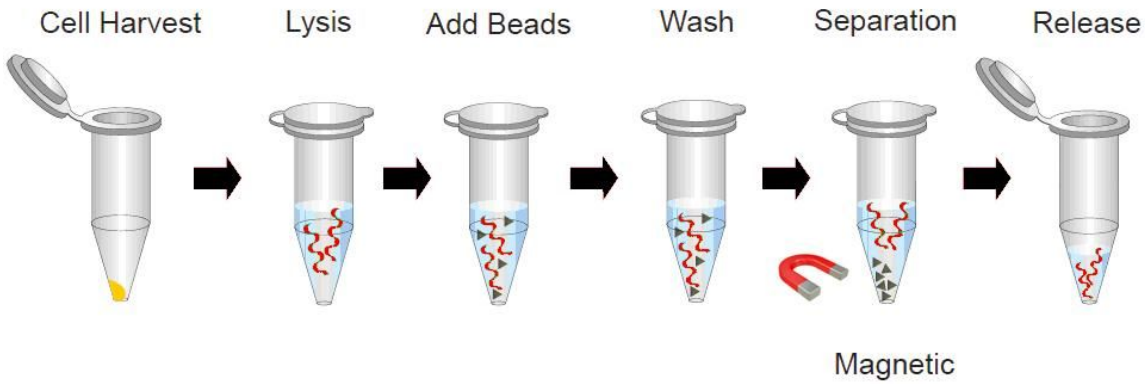
Automated DNA extraction with magnets + robots



https://craft-robotics.s3.amazonaws.com/_thumbnail/Multi-Colored-Pipetting.jpg?mtime=20180919131431
https://the-dna-universe.com/wp-content/uploads/2021/11/Hamilton-liquid-handling-robot.png

https://images.aatbio.com/universal/newsletter/volume-11-1/Silica%20Gel%20DNA%20Extraction%20Workflow.png

Automated DNA extraction with magnets + robots



https://craft-robotics.s3.amazonaws.com/_thumbnail/Multi-Colored-Pipetting.jpg?mtime=20180919131431
https://the-dna-universe.com/wp-content/uploads/2021/11/Hamilton-liquid-handling-robot.png

https://www.epruibiotech.com/wp-content/uploads/2021/03/Magnetic-beads-DNA-extraction-process.jpg

Got DNA now, how do we work out what it says?

Sequencing Technology

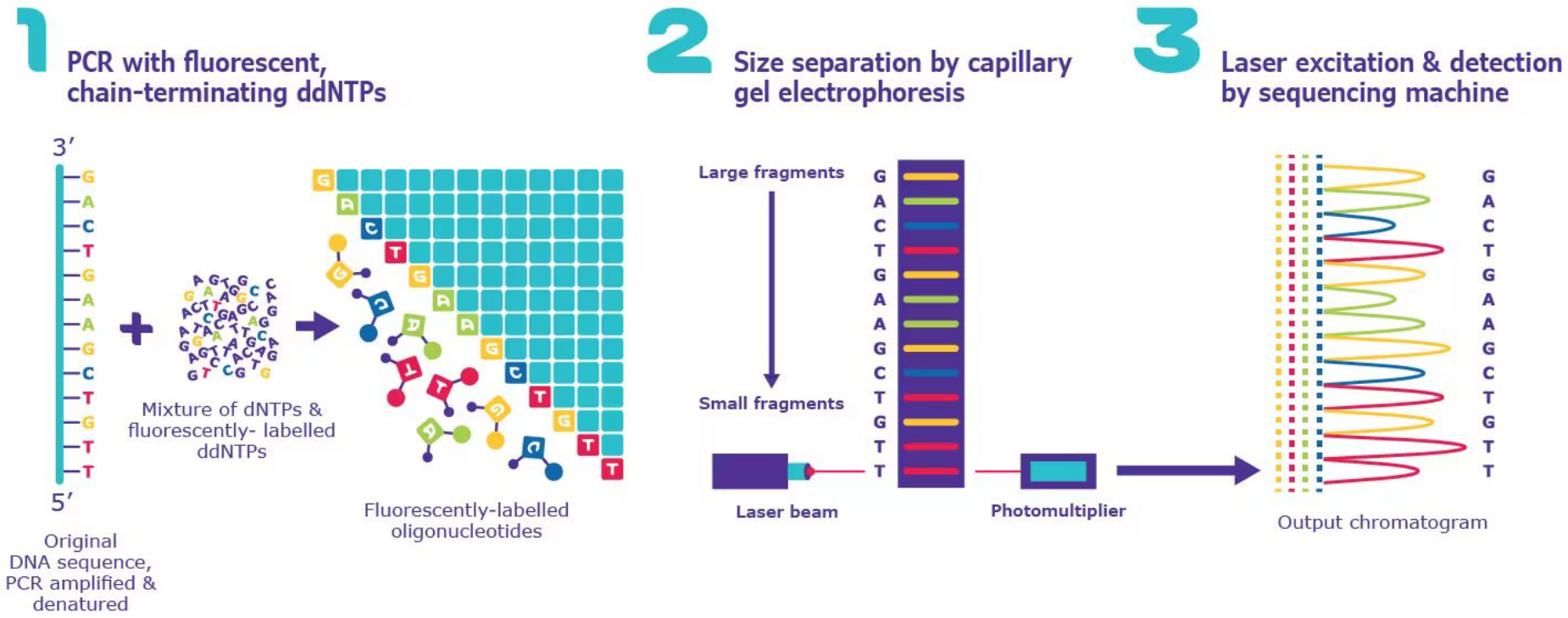
~1972-1977

First generation

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

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

~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/

Sequencing Technology

~1972-1977 ~2001-2004

First generation | **Second generation (next generation sequencing)**

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

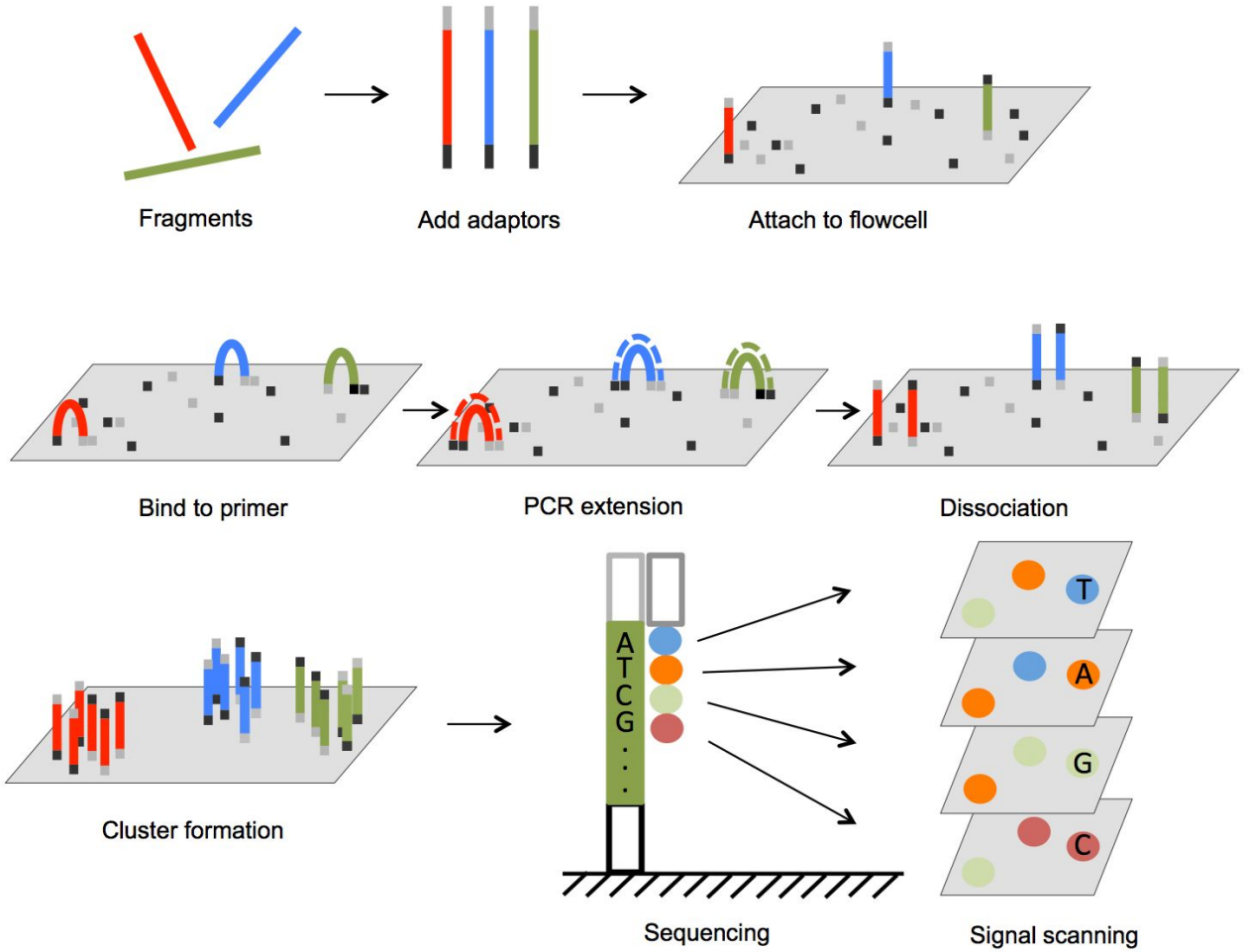
Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

454, Solexa, Ion Torrent, Illumina

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

~1972-1977 ~2001-2004

First generation

Second generation
(next generation sequencing)

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

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 Technology

~1972-1977 | ~2001-2004 | ~2011-2015

First generation | **Second generation** (next generation sequencing) | **Third generation**

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

454, Solexa,
Ion Torrent,
Illumina

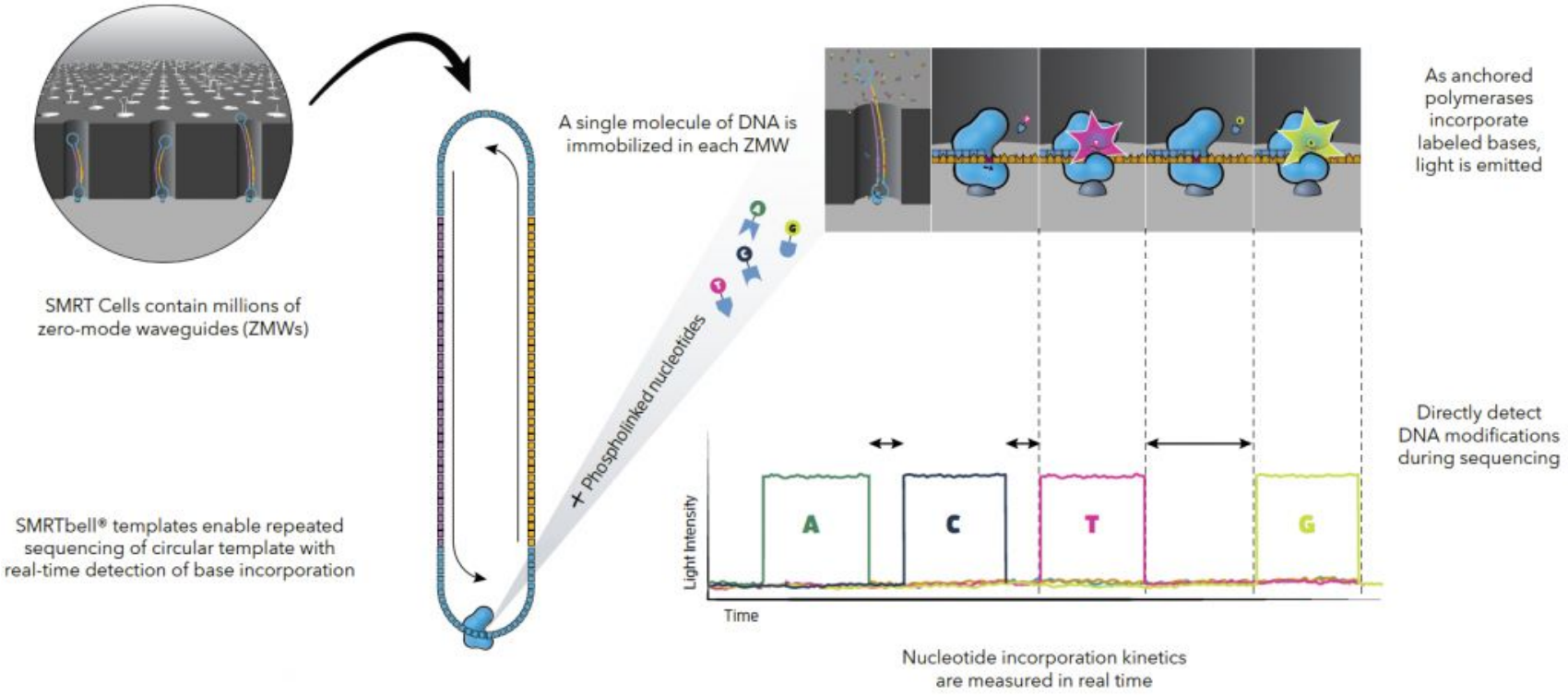
High throughput from the parallelization of sequencing reactions

~50–500 bp fragments

PacBio
Oxford Nanopore

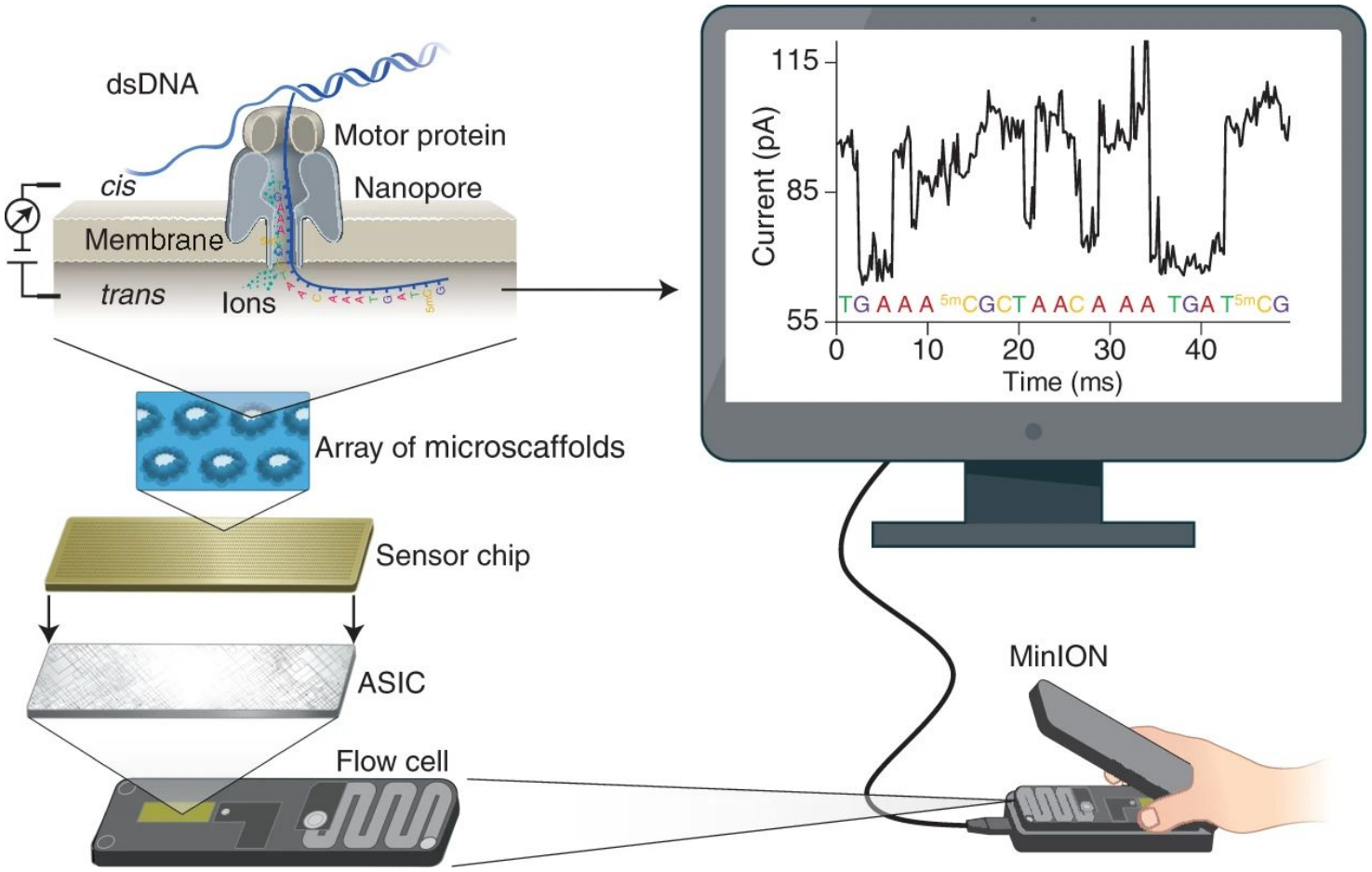
https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

PacBio Sequencing



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

~1972-1977 | ~2001-2004 | ~2011-2015

First generation | **Second generation** (next generation sequencing) | **Third generation**

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

454, Solexa, Ion Torrent, Illumina

High throughput from the parallelization of sequencing reactions

~50–500 bp fragments

PacBio
Oxford Nanopore

Sequence native DNA in real time with single-molecule resolution

Tens of kb fragments, on average

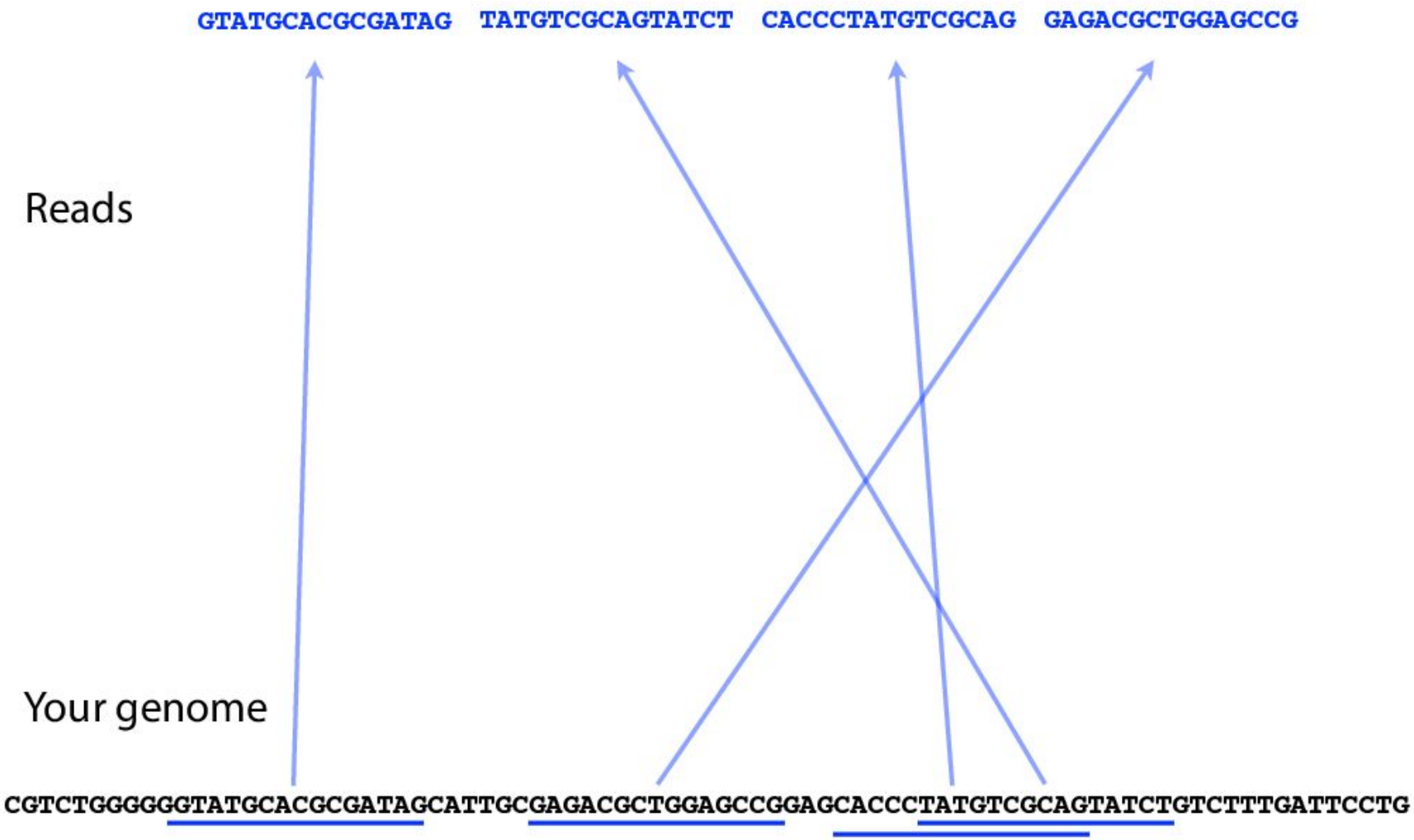
https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

Sequencing Technology

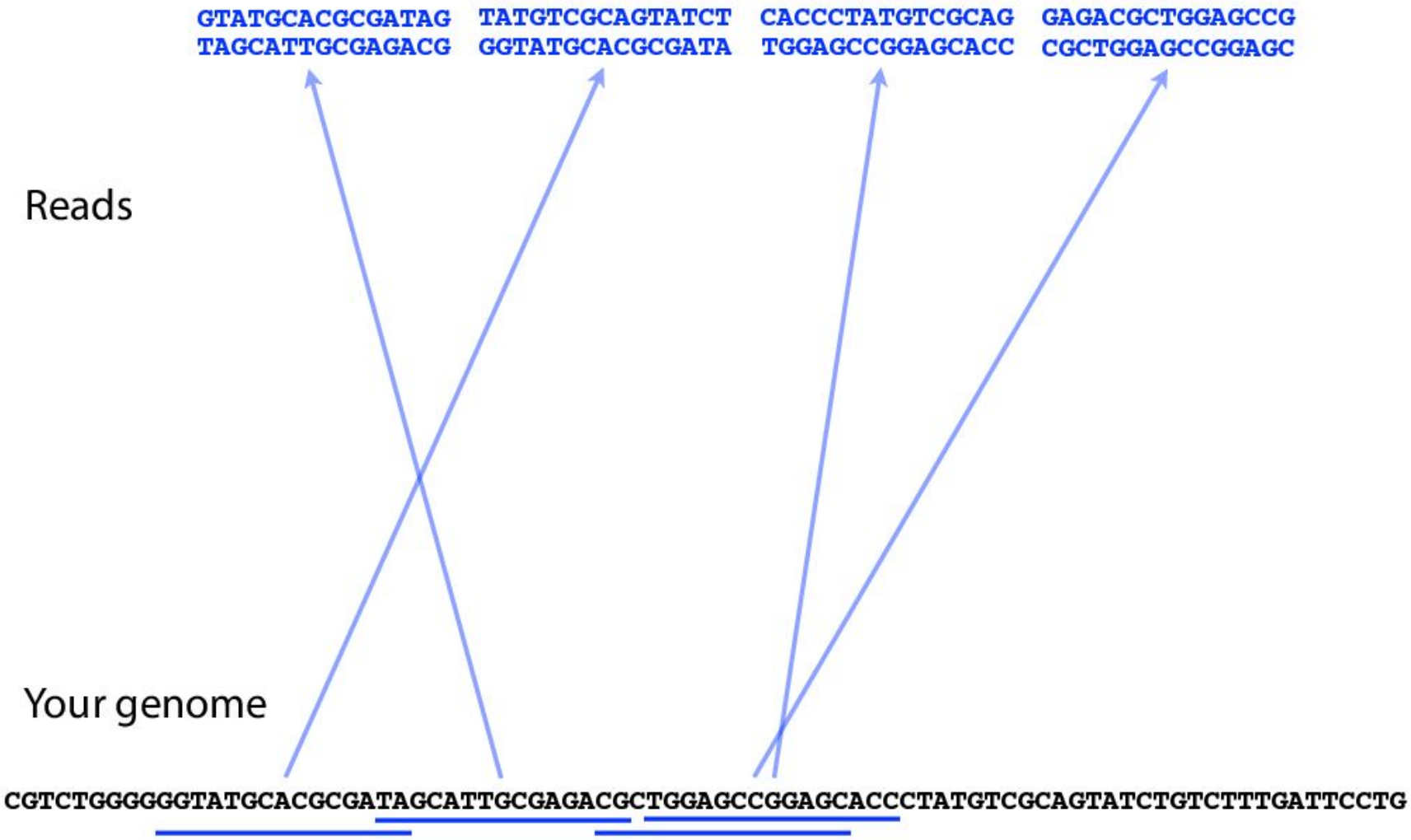
~1972-1977	~2001-2004	~2011-2015
First generation	Second generation (next generation sequencing)	Third generation
  	  	 
Sanger sequencing Maxam and Gilbert Sanger chain termination	454, Solexa, Ion Torrent, Illumina	PacBio Oxford Nanopore
Infer nucleotide identity using dNTPs, then visualize with electrophoresis	High throughput from the parallelization of sequencing reactions	Sequence native DNA in real time with single-molecule resolution
500–1,000 bp fragments	~50–500 bp fragments	Tens of kb fragments, on average
Short-read sequencing		Long-read sequencing

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

Reads are randomly(-ish) sampled from the DNA



Reads are randomly(-ish) sampled from the DNA



Reads are randomly(-ish) sampled from the DNA

Reads are randomly(-ish) sampled from the DNA

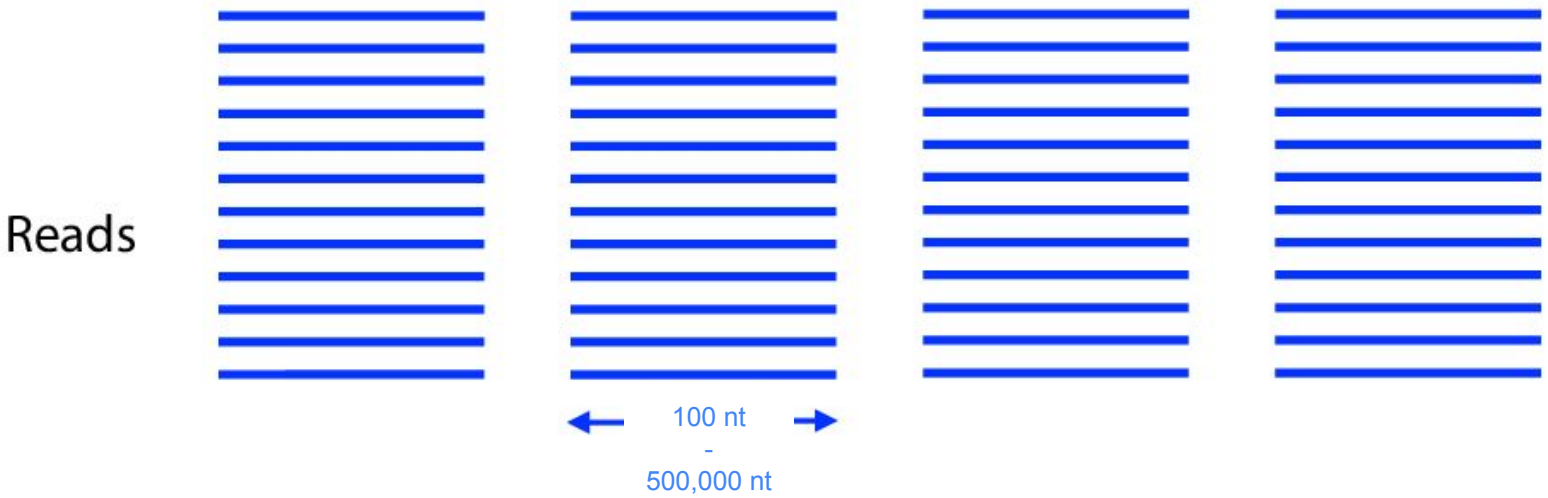
Reads

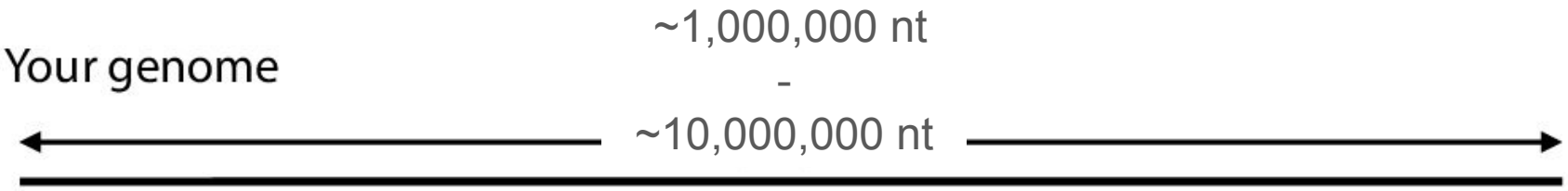
GTATGCACGCGATAG	TATGTCGCAGTATCT	CACCCTATGTCGCAG	GAGACGCTGGAGCCG
TAGCATTGCGAGACG	GGTATGCACGCGATA	TGGAGCCGGAGCACC	CGCTGGAGCCGGAGC
TGTCTTTGATTCCTG	CGCGATAGCATTGCG	GCATTGCGAGACGCT	CCTATGTCGCAGTAT
GACGCTGGAGCCGGA	GCACCCTATGTCGCA	GTATCTGTCTTTGAT	CCTCATCCTATTATT
TATCGCACCTACGTT	CAATATTCGATCATG	GATCACAGGTCTATC	ACCCTATTAACCACT
CACGGGAGCTCTCCA	TGCATTTGGTATTTT	CGTCTGGGGGGTATG	CACGCGATAGCATTG
GTATGCACGCGATAG	ACCTACGTTCAATAT	TATTTATCGCACCTA	CCACTCACGGGAGCT
GCGAGACGCTGGAGC	CTATCACCCTATTAA	CTGTCTTTGATTCCT	ACTCACGGGAGCTCT
CCTACGTTCAATATT	GCACCTACGTTCAAT	GTCTGGGGGGTATGC	AGCCGGAGCACCCTA
GACGCTGGAGCCGGA	GCACCCTATGTCGCA	GTATCTGTCTTTGAT	CCTCATCCTATTATT
TATCGCACCTACGTT	CAATATTCGATCATG	GATCACAGGTCTATC	ACCCTATTAACCACT
CACGGGAGCTCTCCA	TGCATTTGGTATTTT	CGTCTGGGGGGTATG	CACGCGATAGCATTG

Your genome

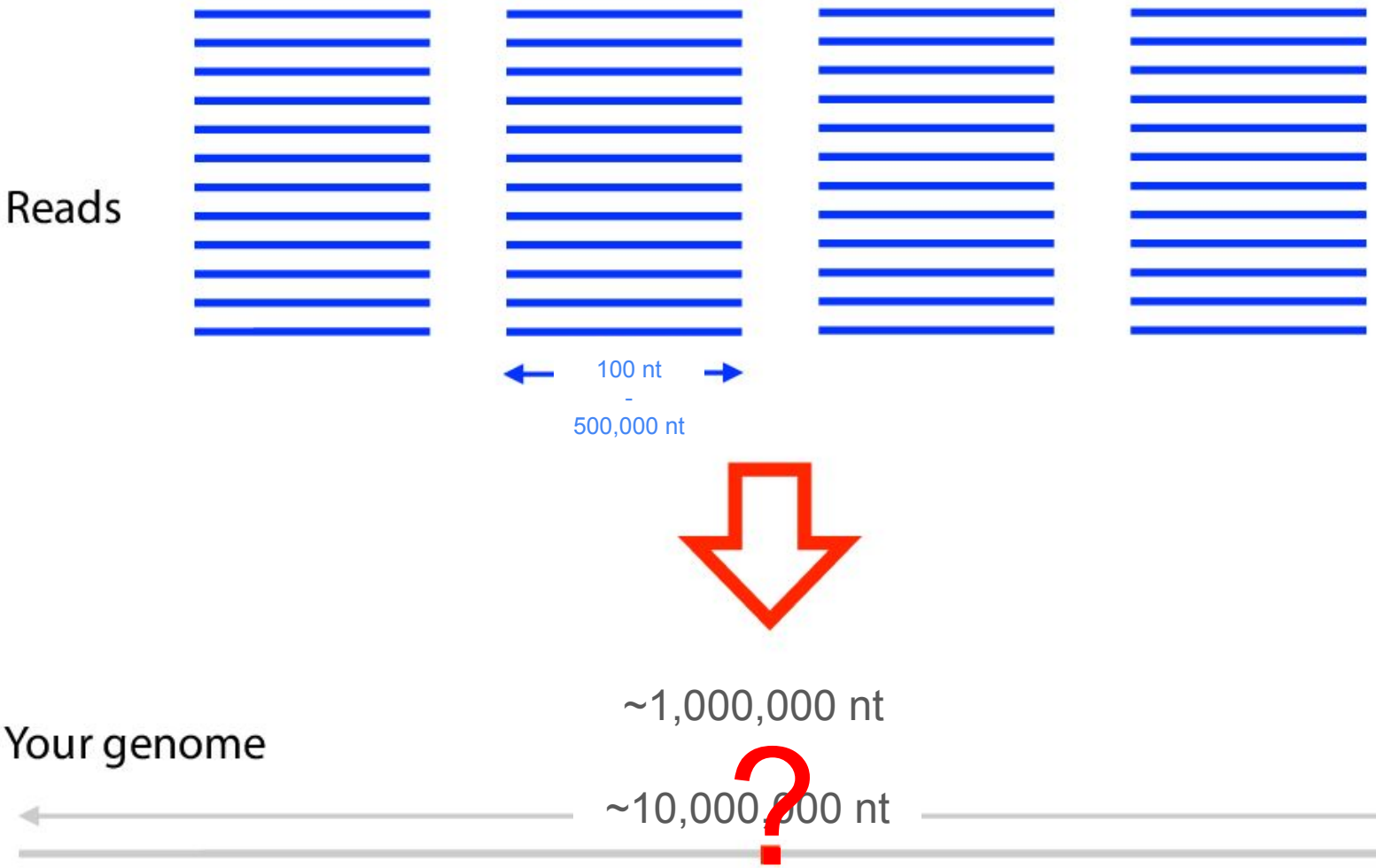
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG

Reads are randomly(-ish) sampled from the DNA

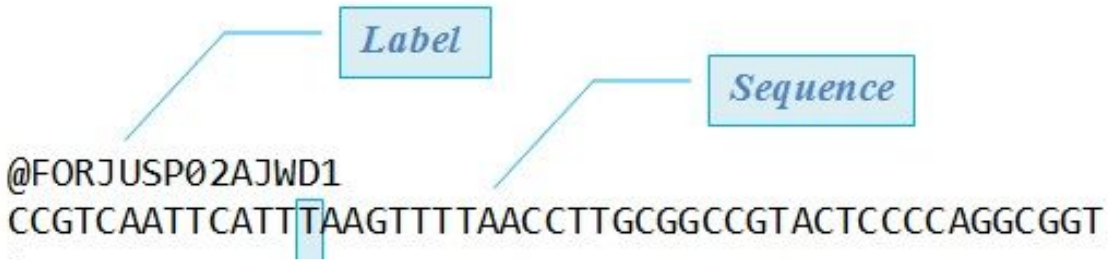




Reads are randomly(-ish) sampled from the DNA

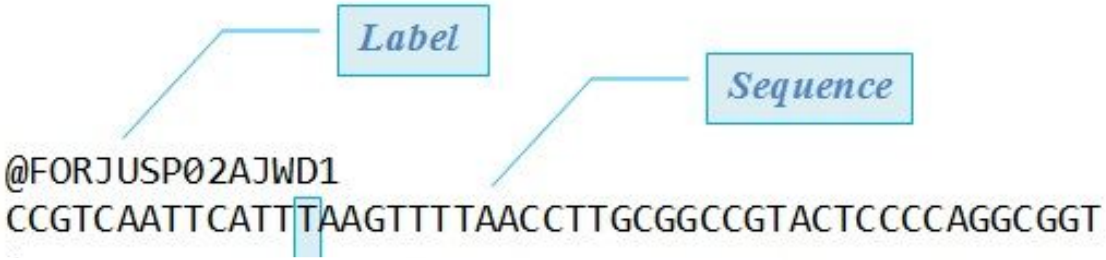


Reads also have measurement error: FASTQ



https://www.drive5.com/usearch/manual/fastq_files.html
https://learn.gencore.bio.nyu.edu/ngs-file-formats/quality-scores/

Reads also have measurement error: FASTQ

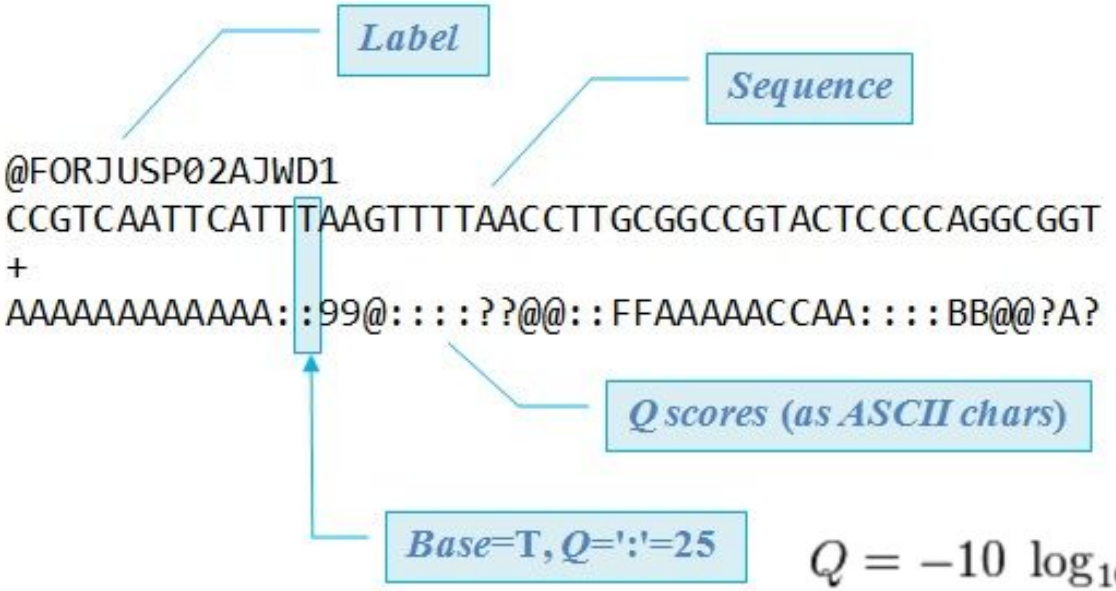


$$Q = -10 \log_{10} P \quad \Rightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

https://www.drive5.com/usearch/manual/fastq_files.html
https://learn.gencore.bio.nyu.edu/ngs-file-formats/quality-scores/

Reads also have measurement error: FASTQ



$$Q = -10 \log_{10} P \quad \Rightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

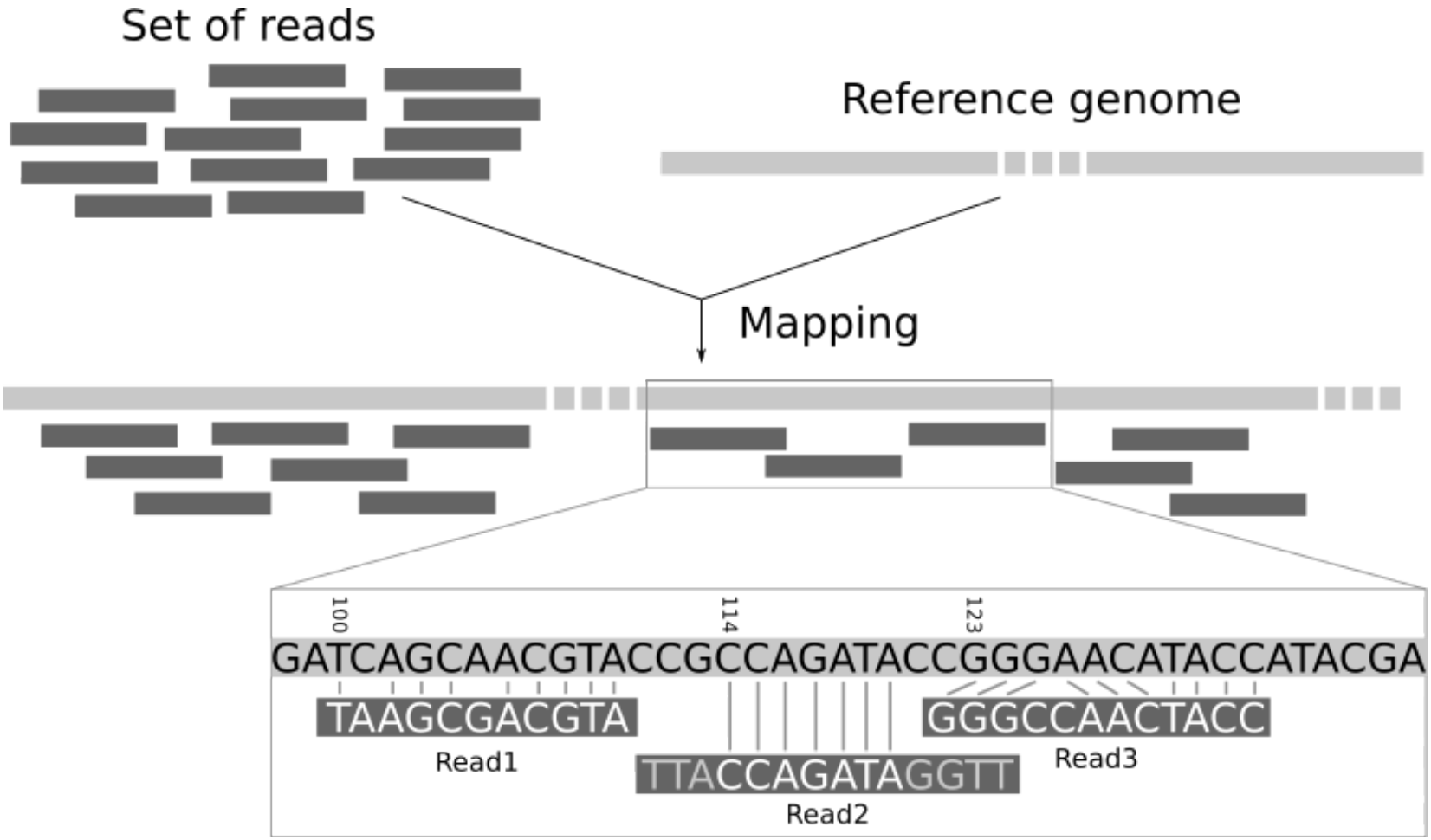
https://www.drive5.com/usearch/manual/fastq_files.html
https://learn.gencore.bio.nyu.edu/ngs-file-formats/quality-scores/

What do we do with these reads?

Reference-based assembly

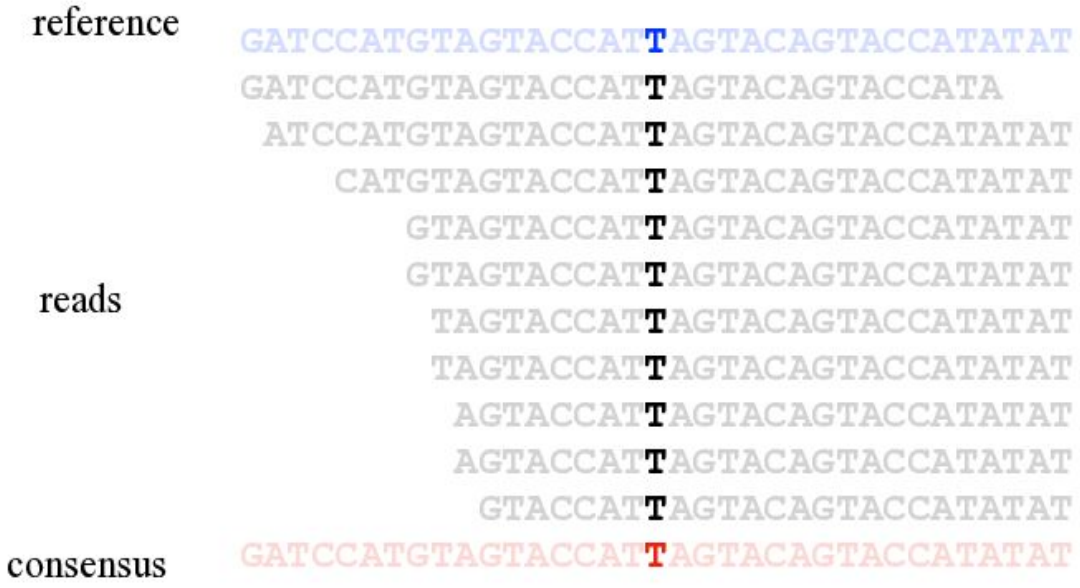
https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

Reference-based assembly



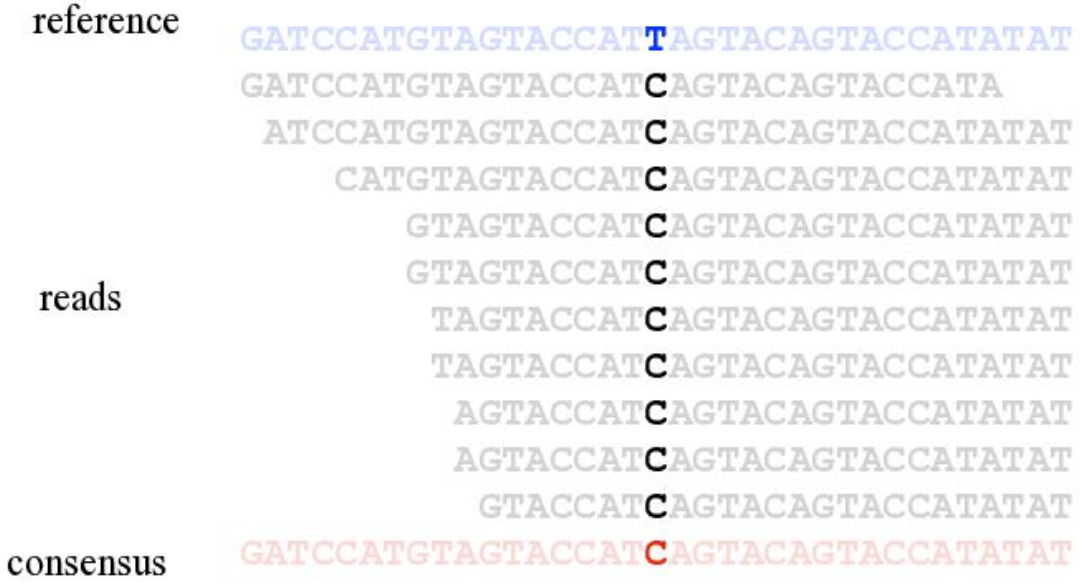
https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

Variant Calling / Consensus generation



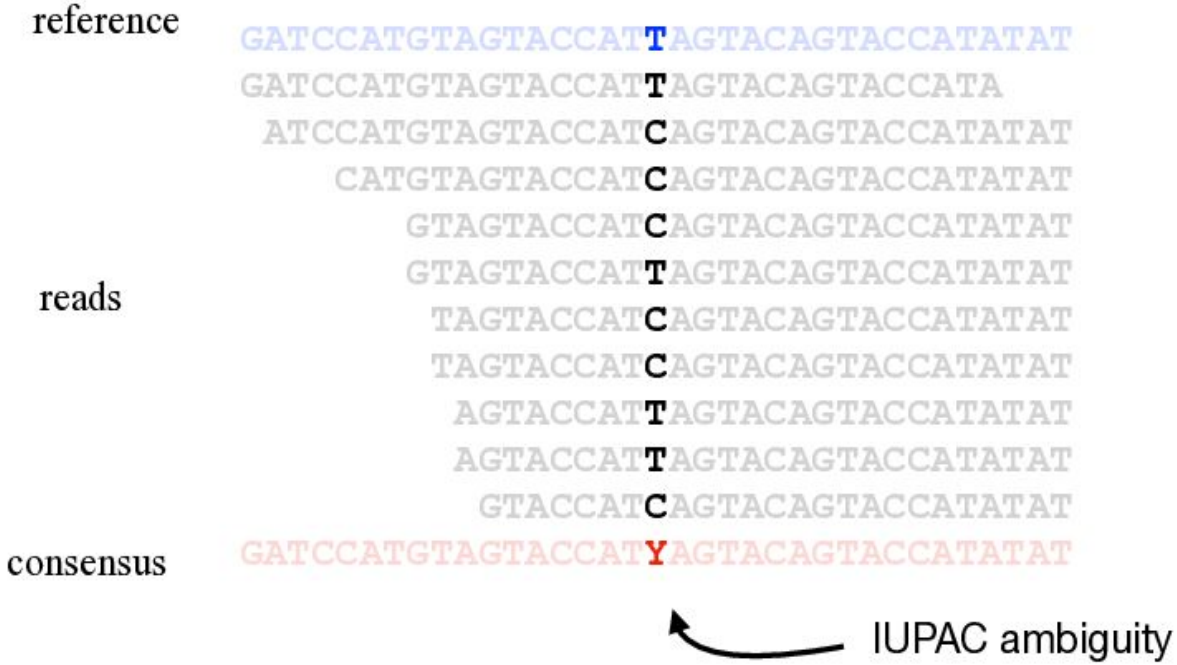
Variant Calling / Consensus generation

- *variants* or *mutations* are detected when the aligned reads differ from the reference

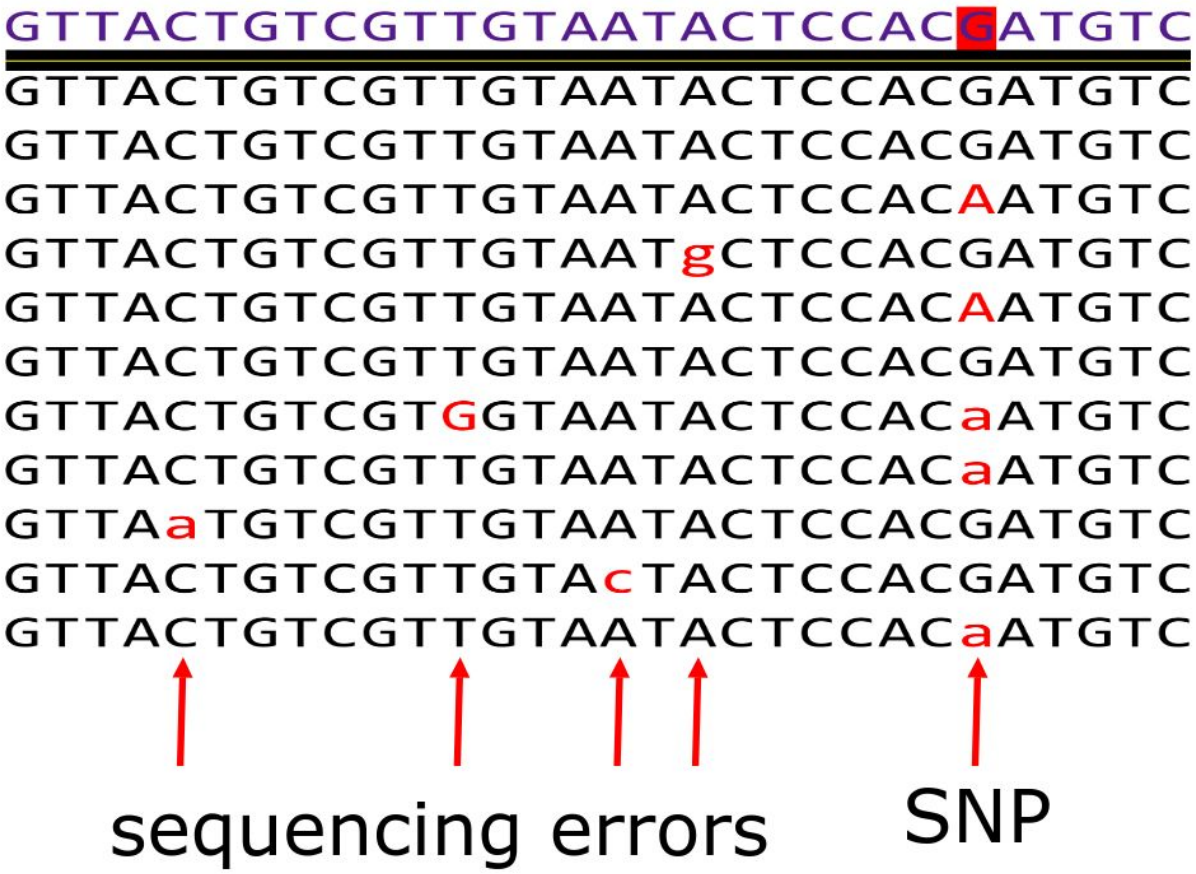


Variant Calling / Consensus generation

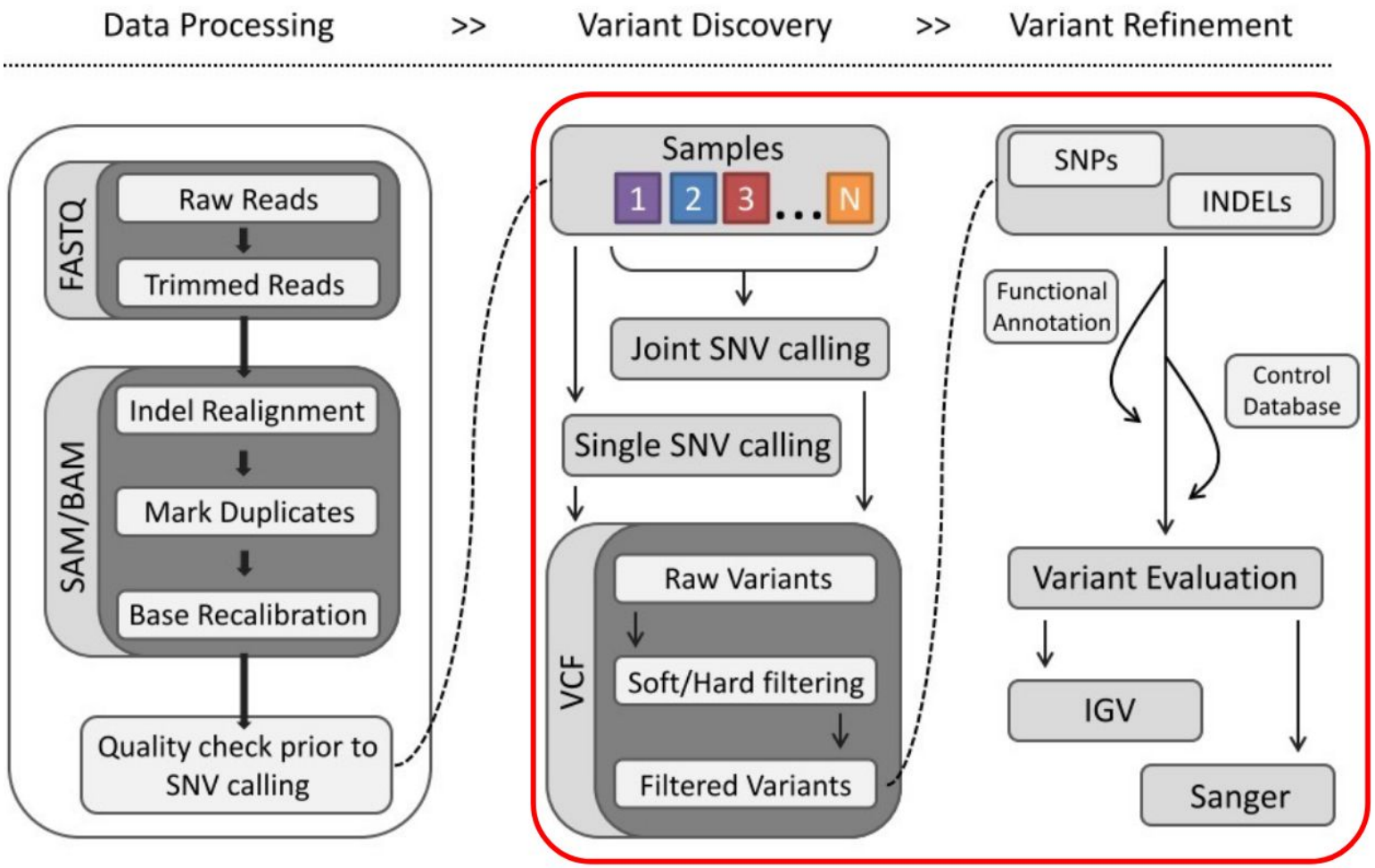
- *ambiguous* or *mixed* bases are detected when the aligned reads have evidence for more than one type of base



Distinguishing real variants from error is non-trivial

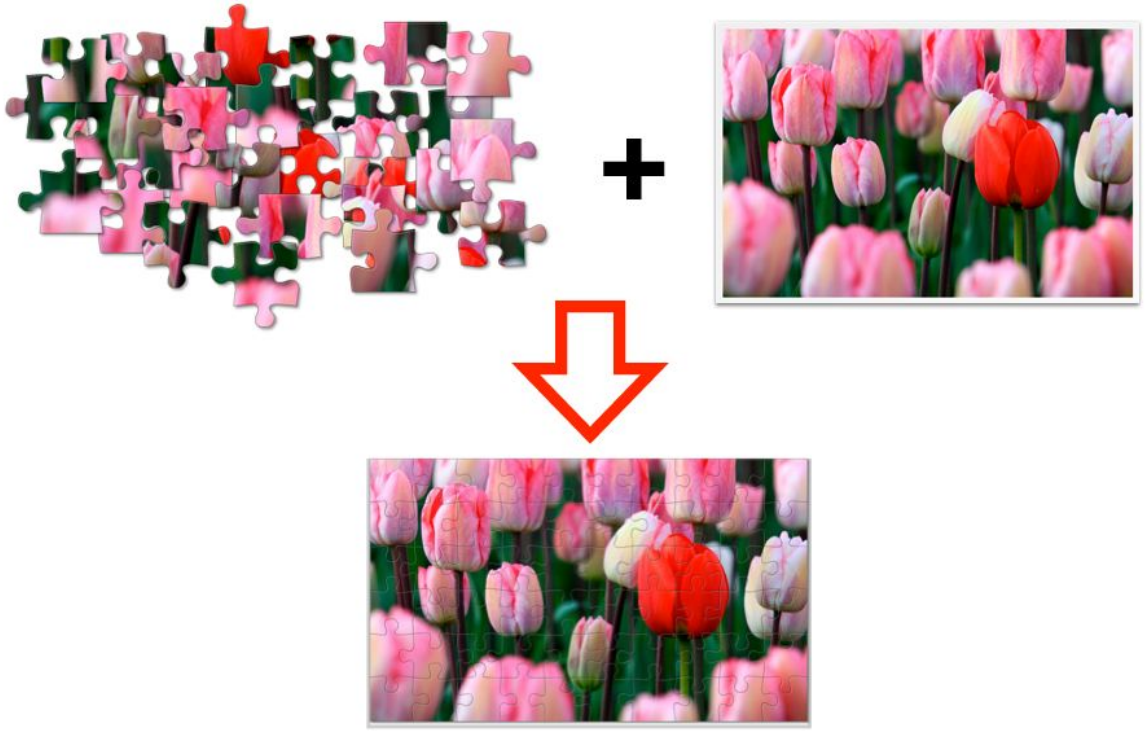


Lots of steps to get accurate reference-based data

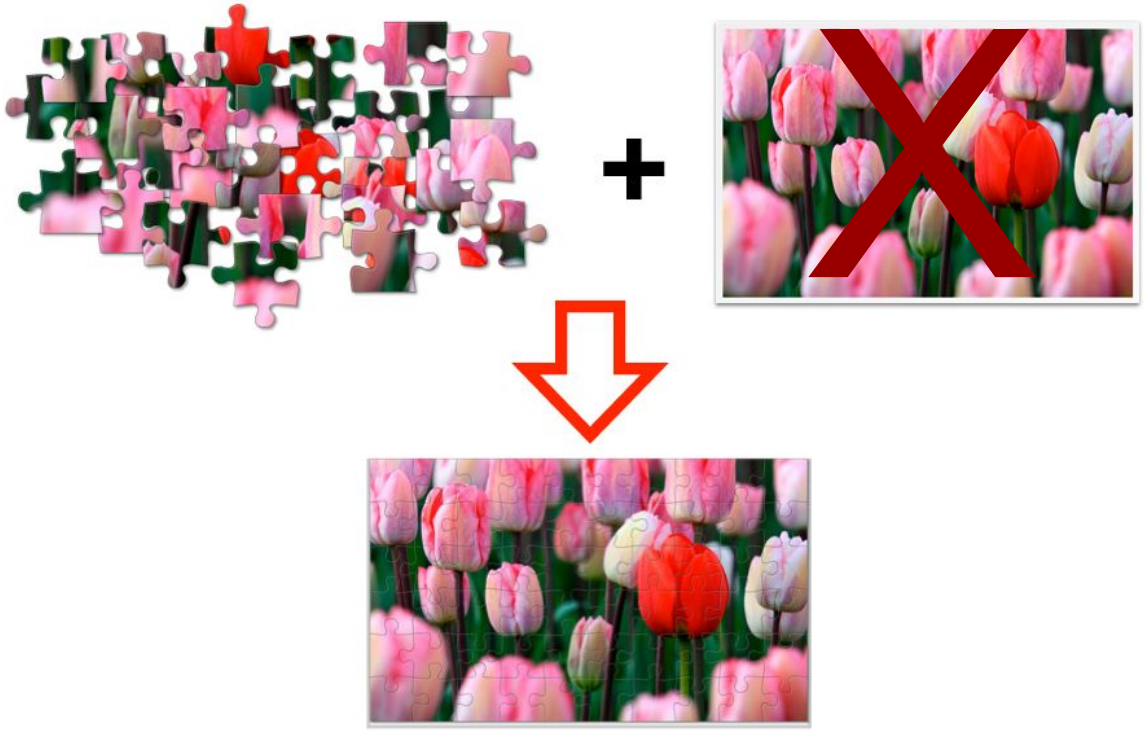


What if we don’t have a reference?

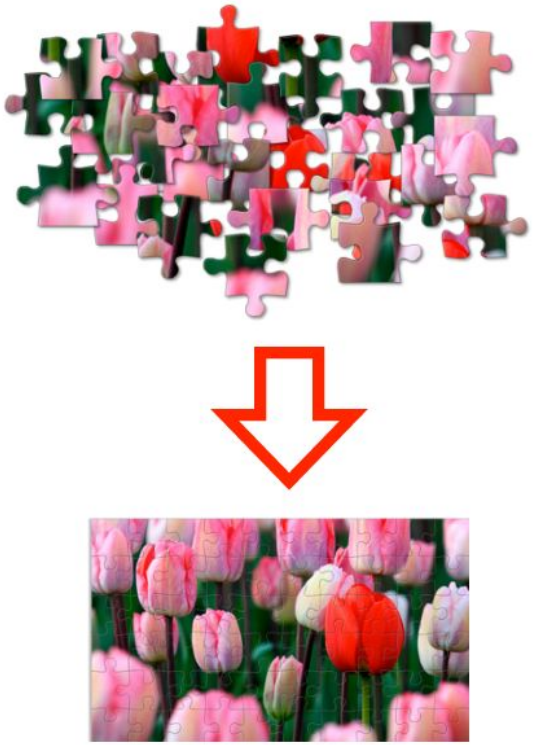
de novo Assembly



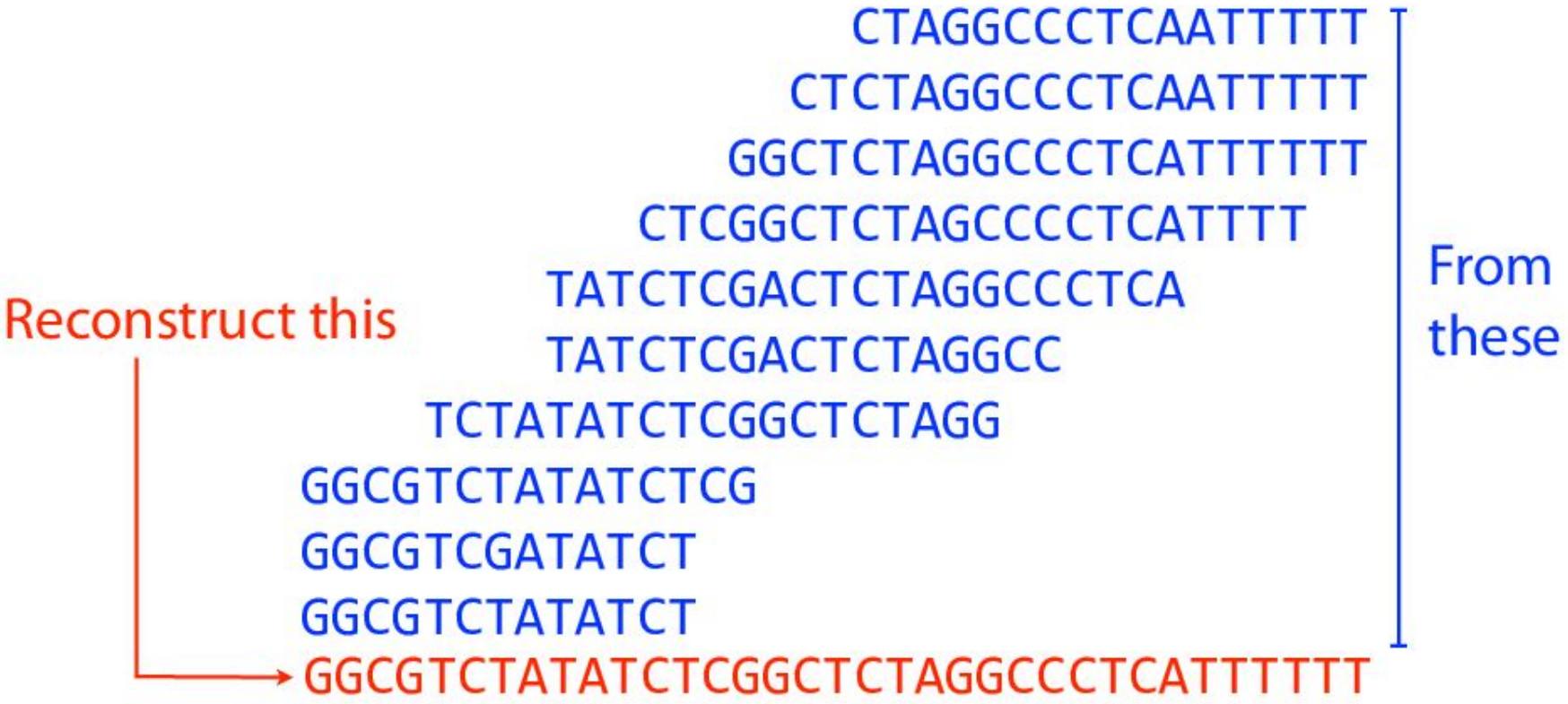
de novo Assembly



de novo Assembly



de novo Assembly



de novo Assembly

Enough sampling (depth) and reads will overlap

If the end of 1 read matches the start of another:

```
TCTATATCTCGGCTCTAGG
    ||||||| |||||||
    TATCTCGACTCTAGGCC
```

Enough sampling (depth) and reads will overlap

If the end of 1 read matches the start of another:

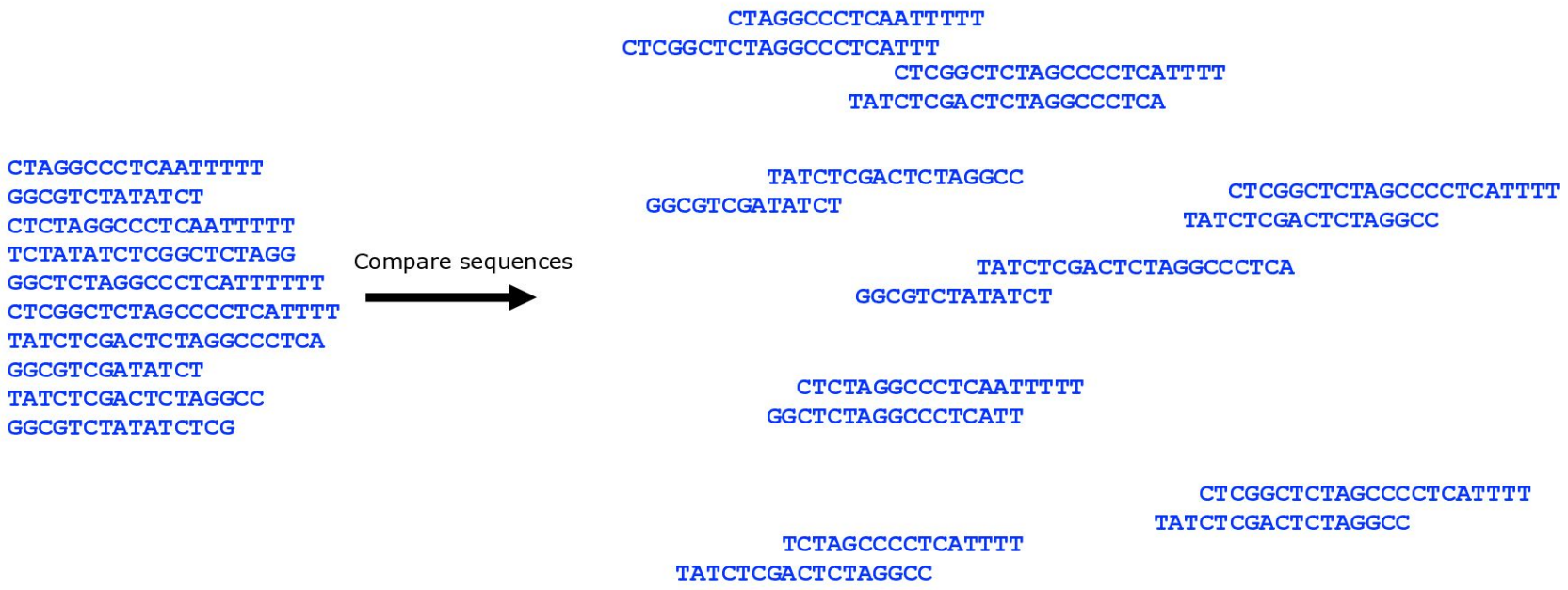
```
TCTATATCTCGGCTCTAGG
      ||||||| |||||||
      TATCTCGACTCTAGGCC
```

Then they **MIGHT** be from overlapping bits of the genome:

```
            TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
            TATCTCGACTCTAGGCC
```

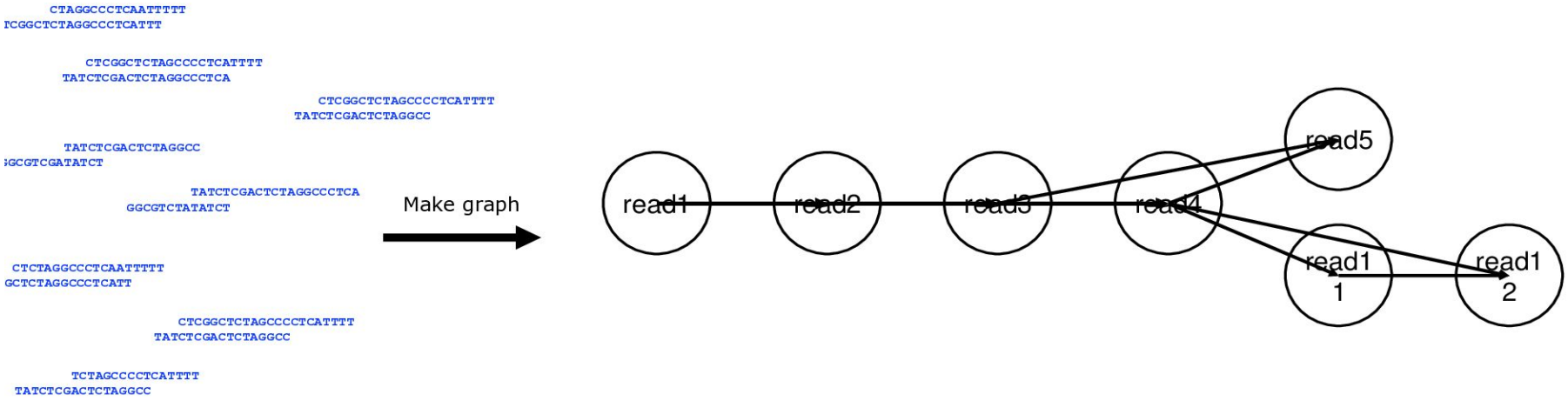
Use shared sequences to get longer fragments (contigs)

- Step 1: Find all overlapping pairs of reads



Use shared sequences to get longer fragments (contigs)

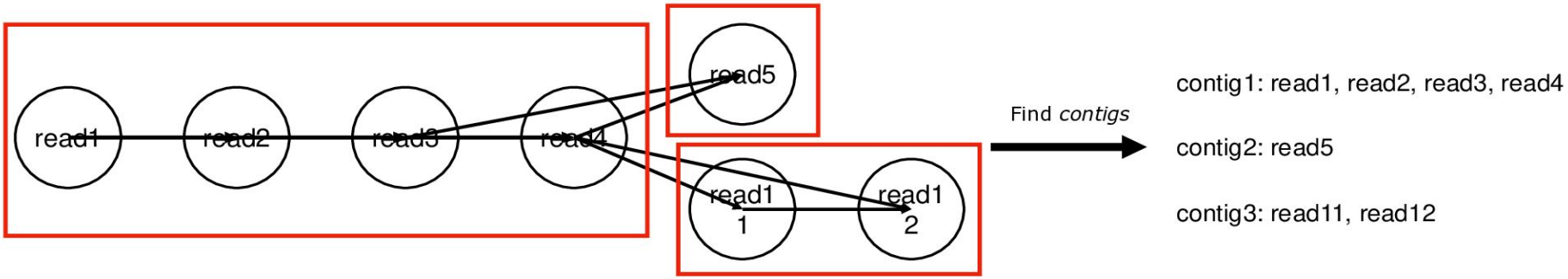
- Step 2: Construct a graph representing read-read overlaps



Every read is a vertex
Every overlapping pair of reads is connected with an edge

Use shared sequences to get longer fragments (contigs)

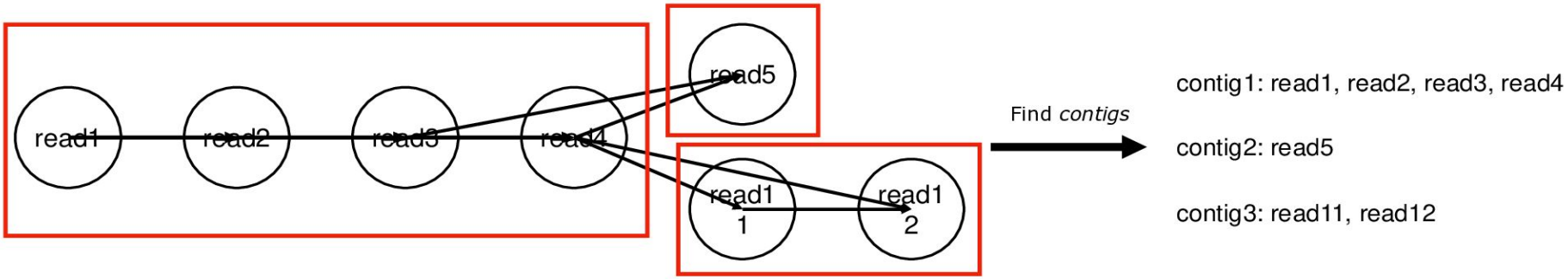
- Step 3: Analyze graph structure to find possible repeat boundaries



Key term: *contig*. Short for “contiguous sequence” it is the result of assembling reads together

Use shared sequences to get longer fragments (contigs)

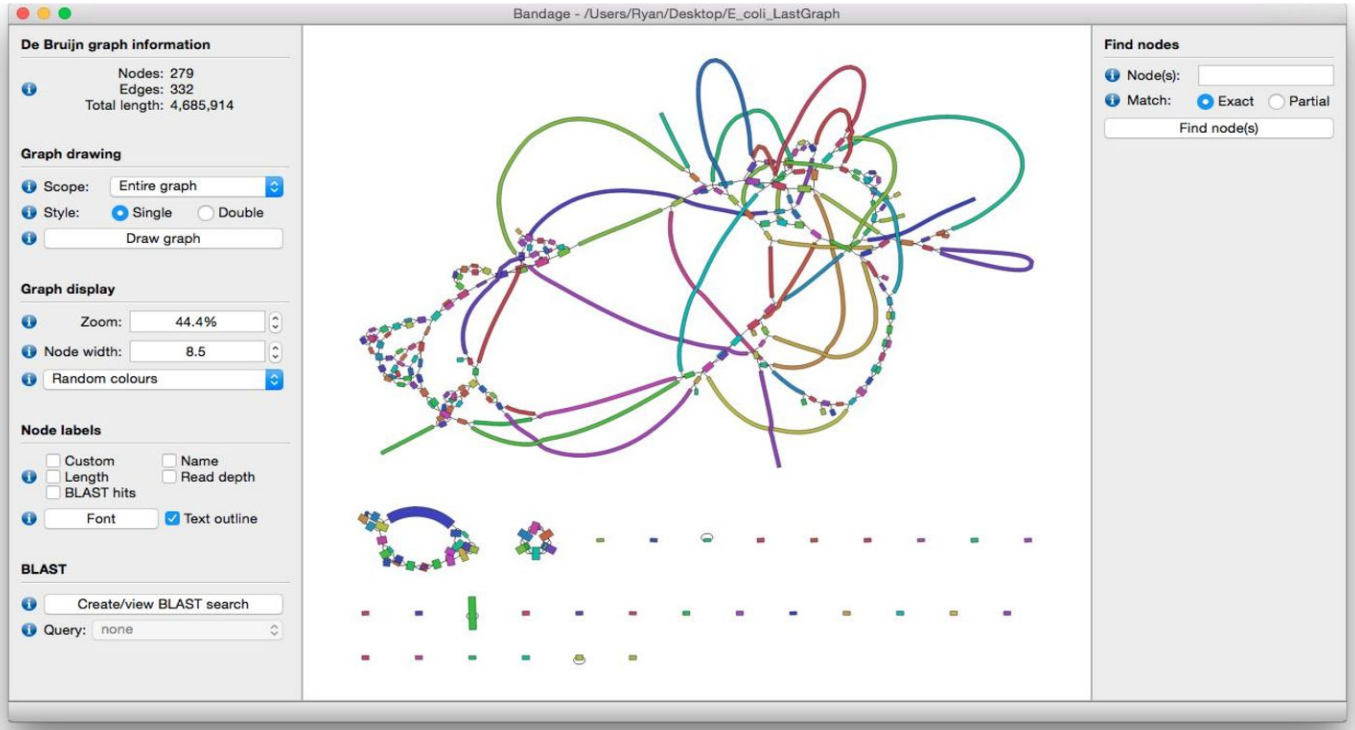
- Step 3: Analyze graph structure to find possible repeat boundaries



Key term: *contig*. Short for "contiguous sequence" it is the result of assembling reads together

Note: assembly is often done with a variant of this which uses exact matches of even shorter bits of reads (called k-mers)

Real assembly graphs are messy and complicated!



Simplified graph of *E. Coli* (4.6Mbp) genome, viewed with Bandage: https://rrwick.github.io/B[illegible]

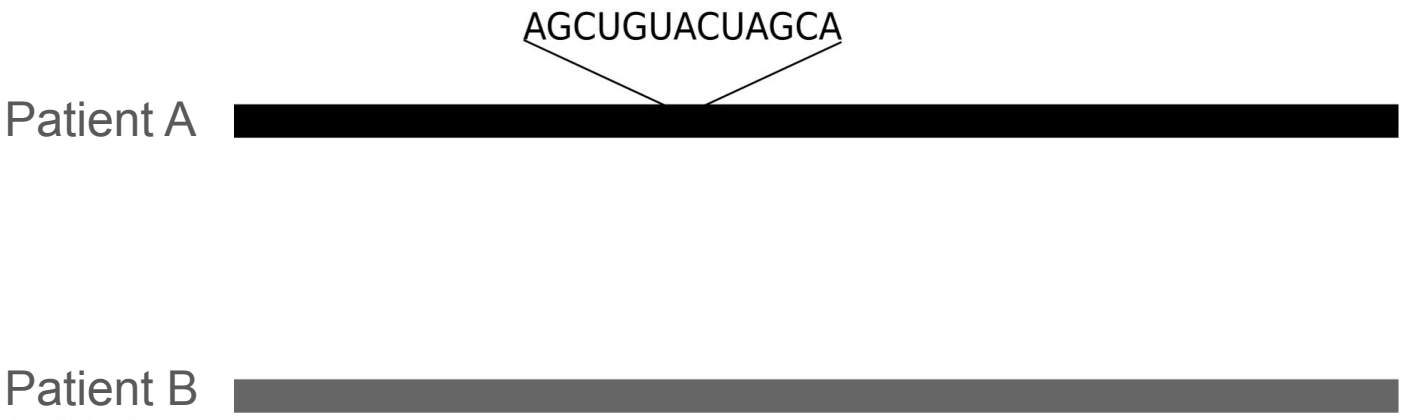
Got genomes, but how do we use them to stop our outbreak?

Compare genomes from each patient

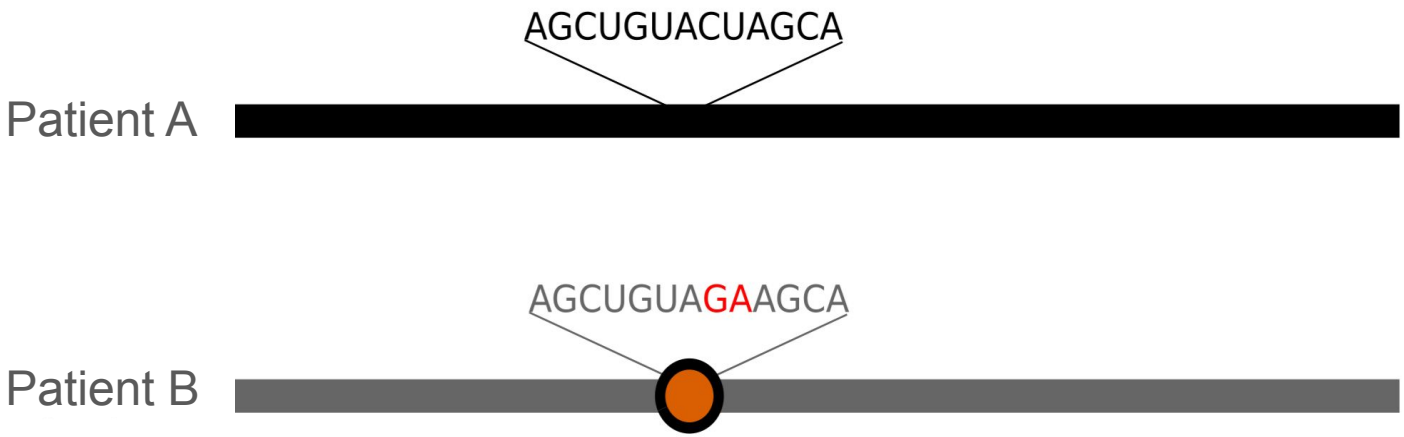
Patient A

Patient B

Compare genomes from each patient



Compare genomes from each patient

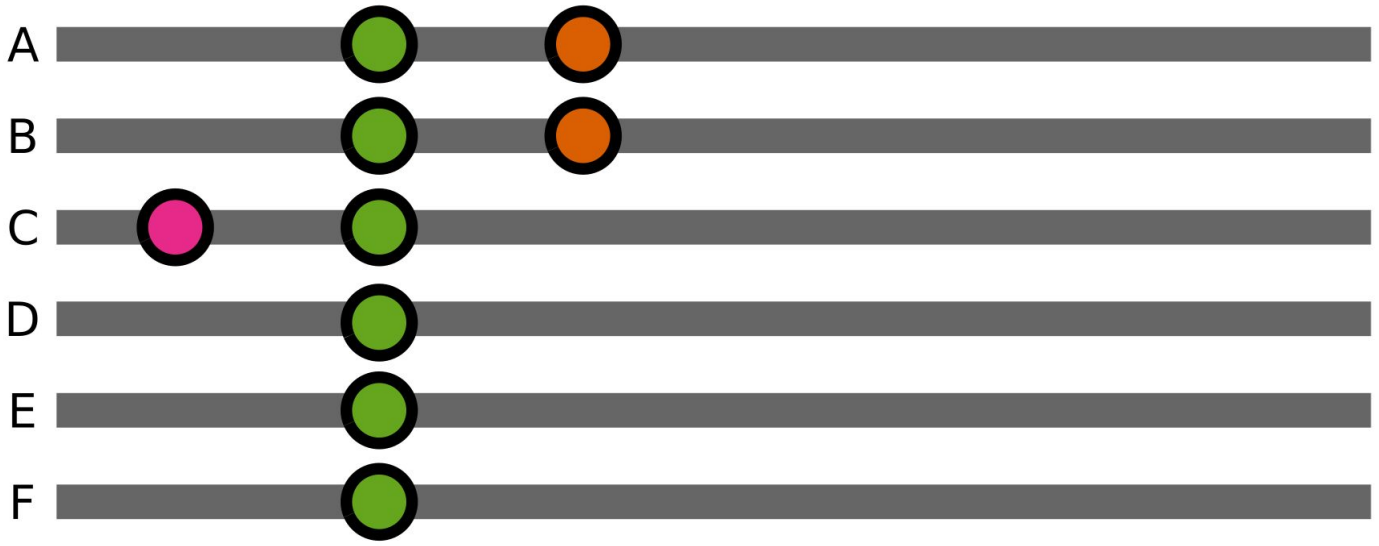


Compare (align) genomes from each patient

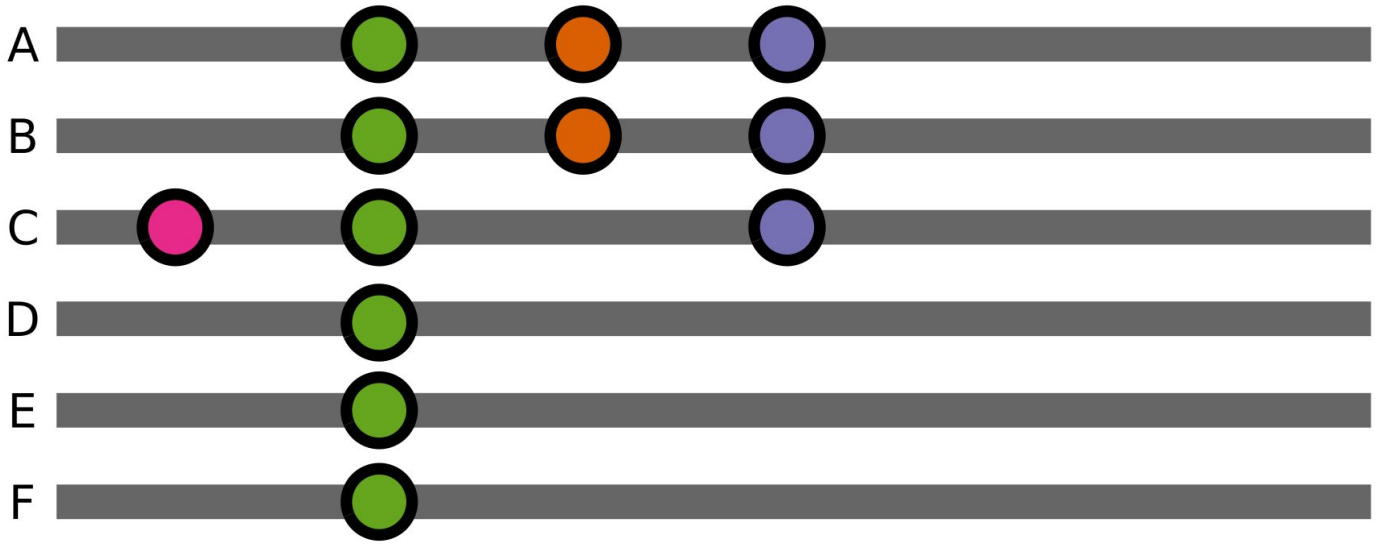
Compare (align) genomes from each patient

Compare (align) genomes from each patient

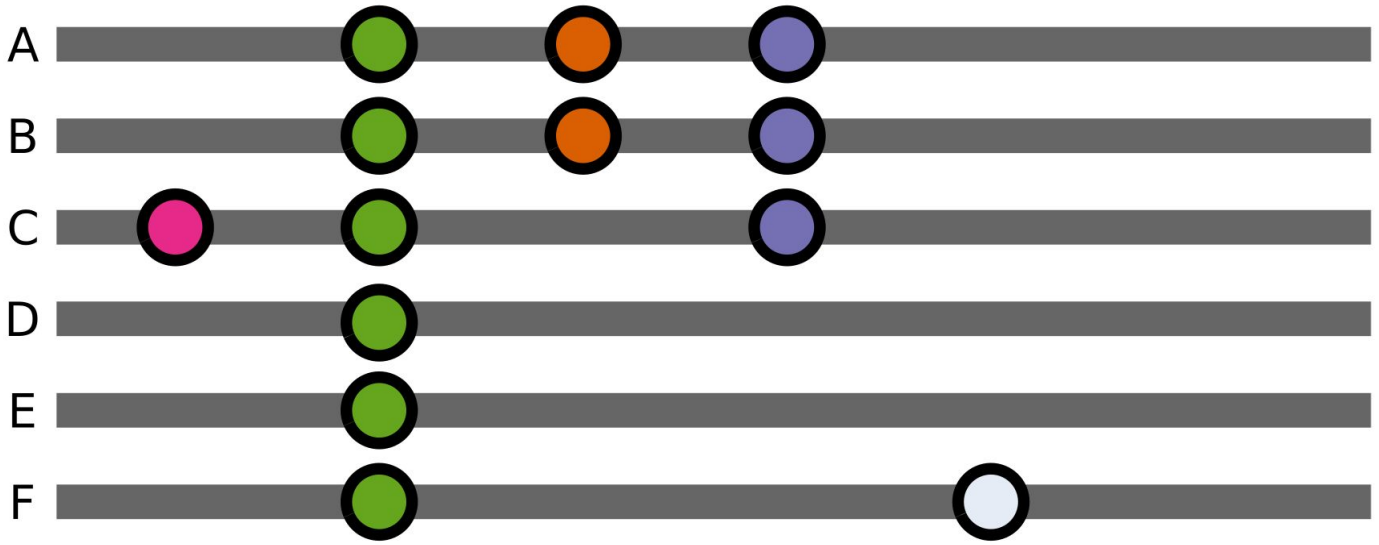
Compare (align) genomes from each patient



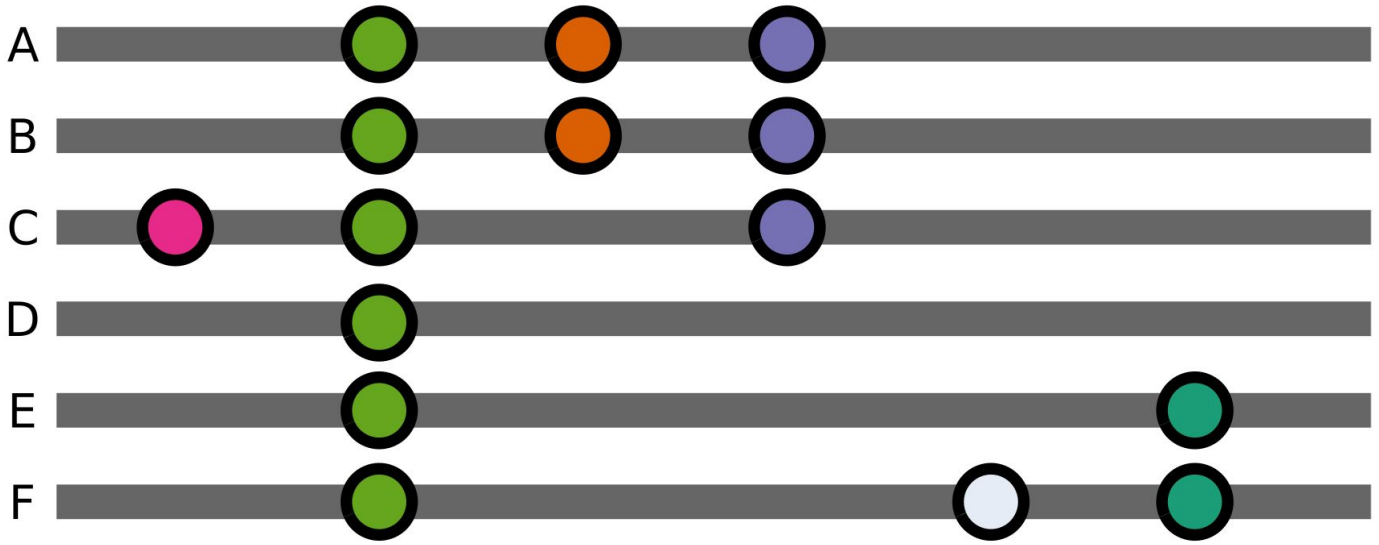
Compare (align) genomes from each patient



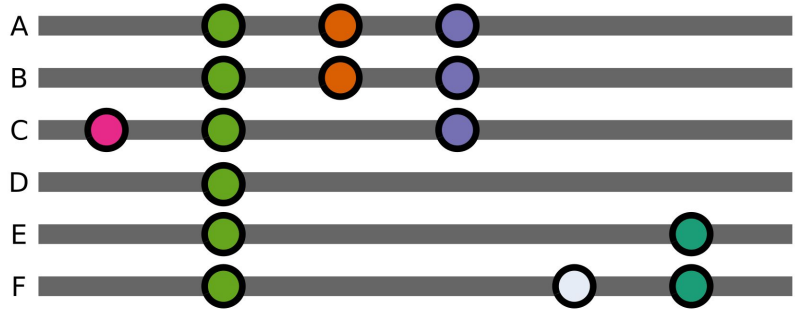
Compare (align) genomes from each patient



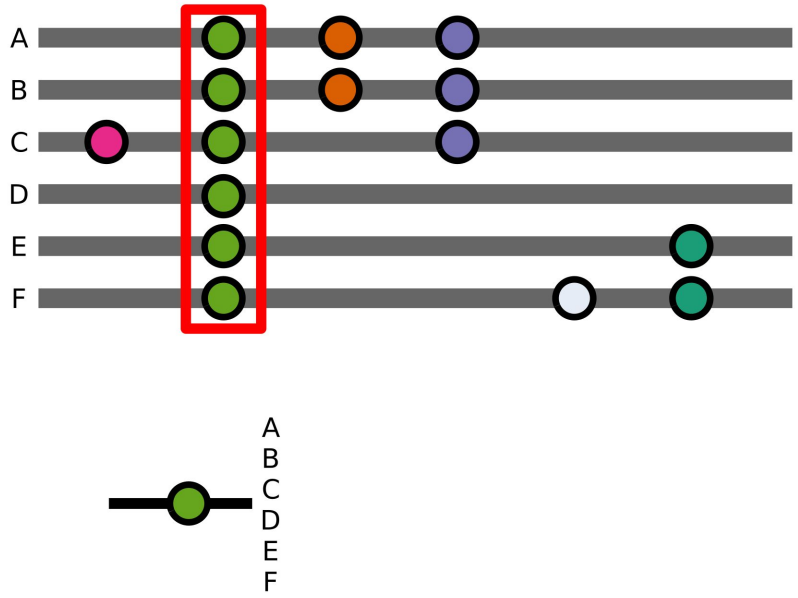
Compare (align) genomes from each patient



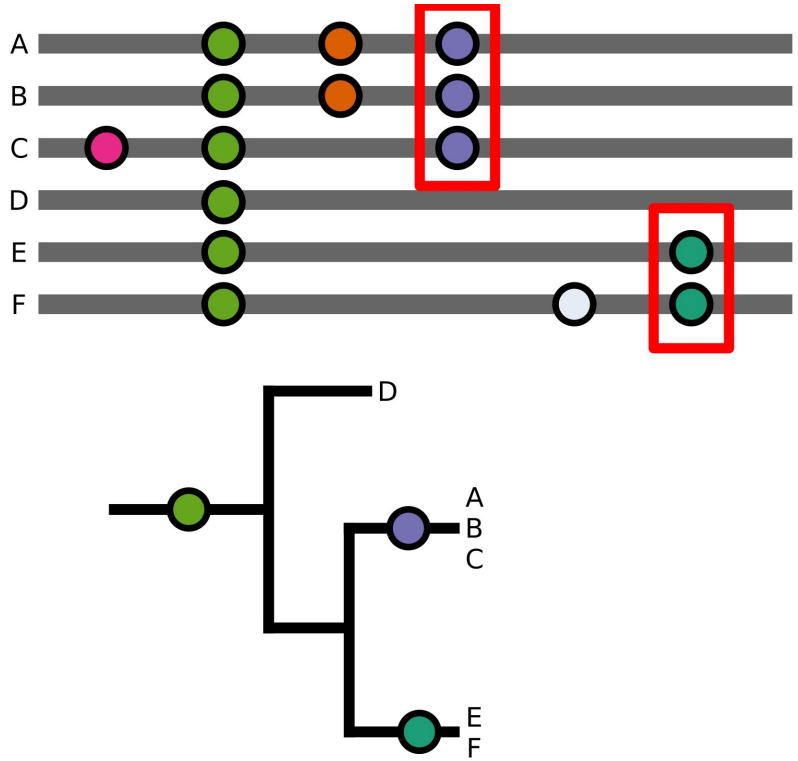
Use pattern of similarities/differences to infer relationships



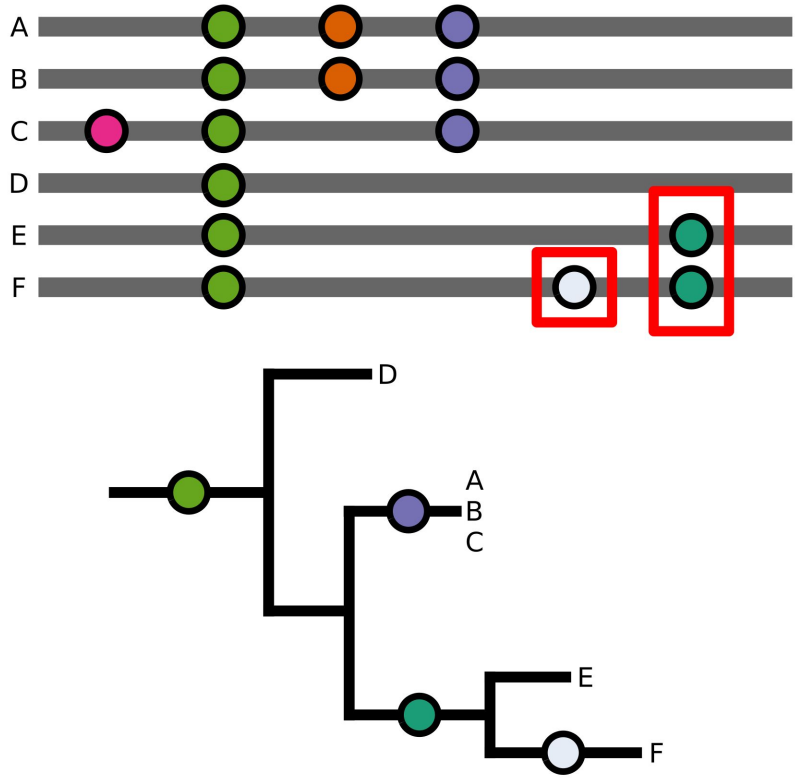
Use pattern of similarities/differences to infer relationships



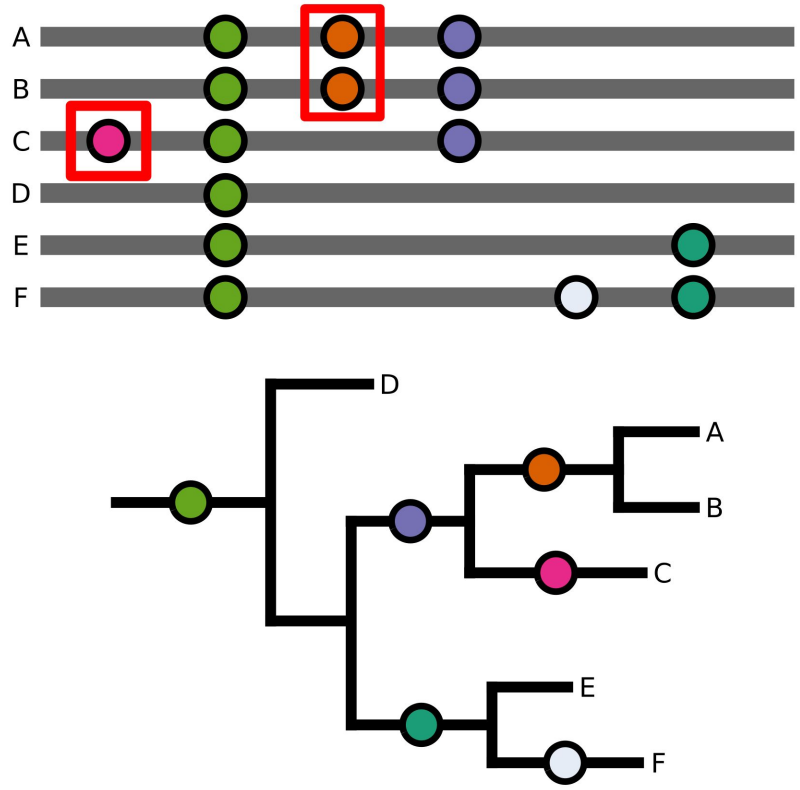
Use pattern of similarities/differences to infer relationships



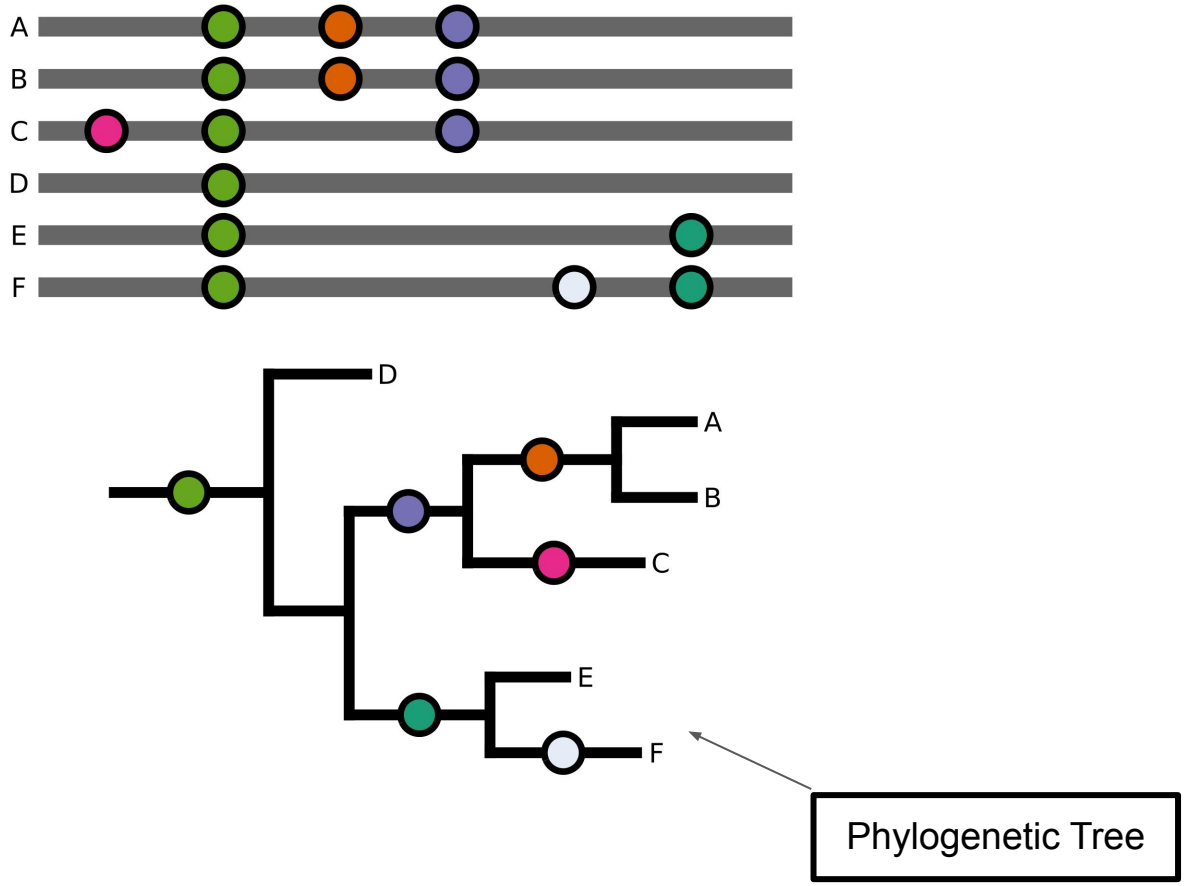
Use pattern of similarities/differences to infer relationships



Use pattern of similarities/differences to infer relationships



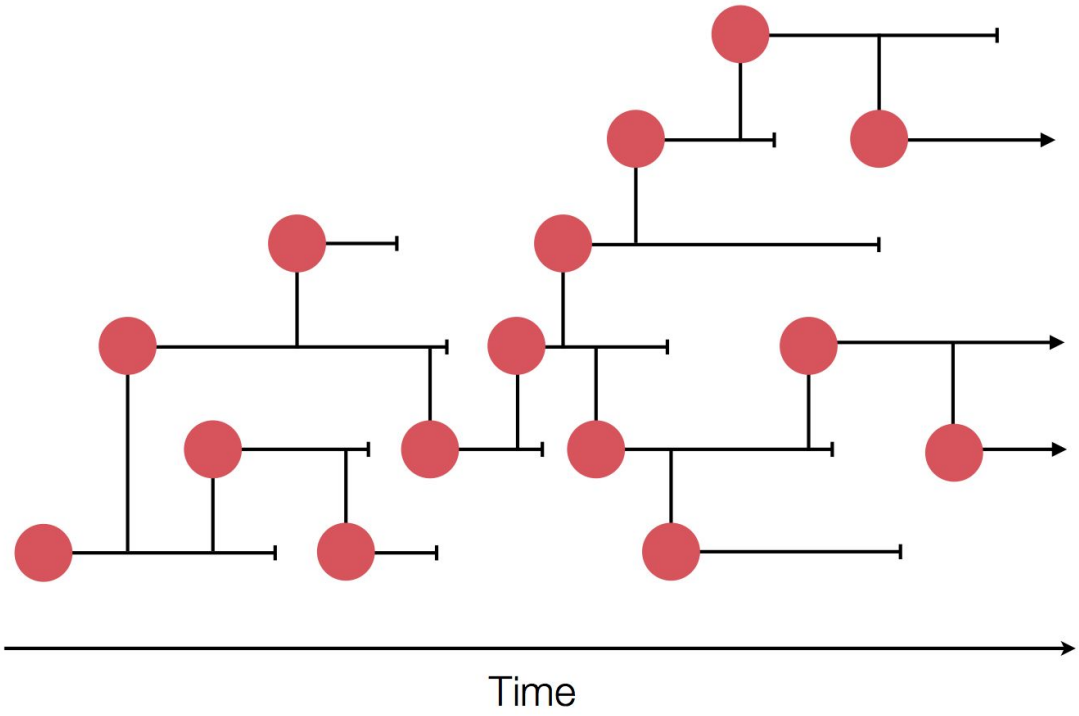
Use pattern of similarities/differences to infer relationships



Note: this generally uses statistical models which incorporate differences in substitution rates and mutation rates

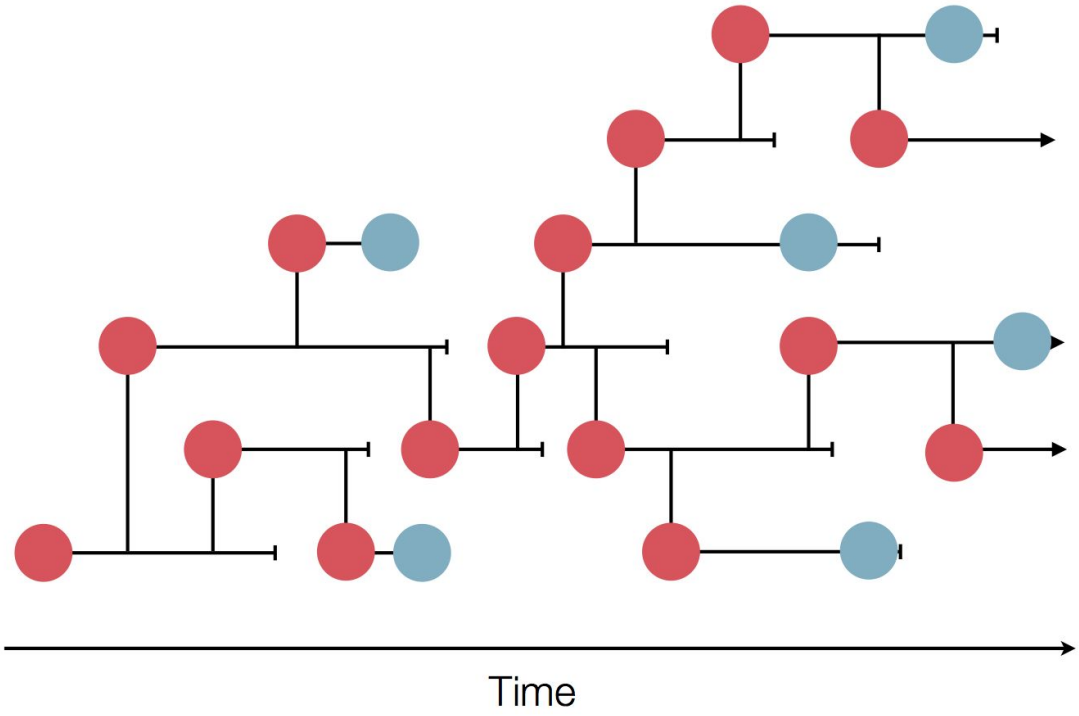
How does this tree help us?

Trees sample the underlying transmission network



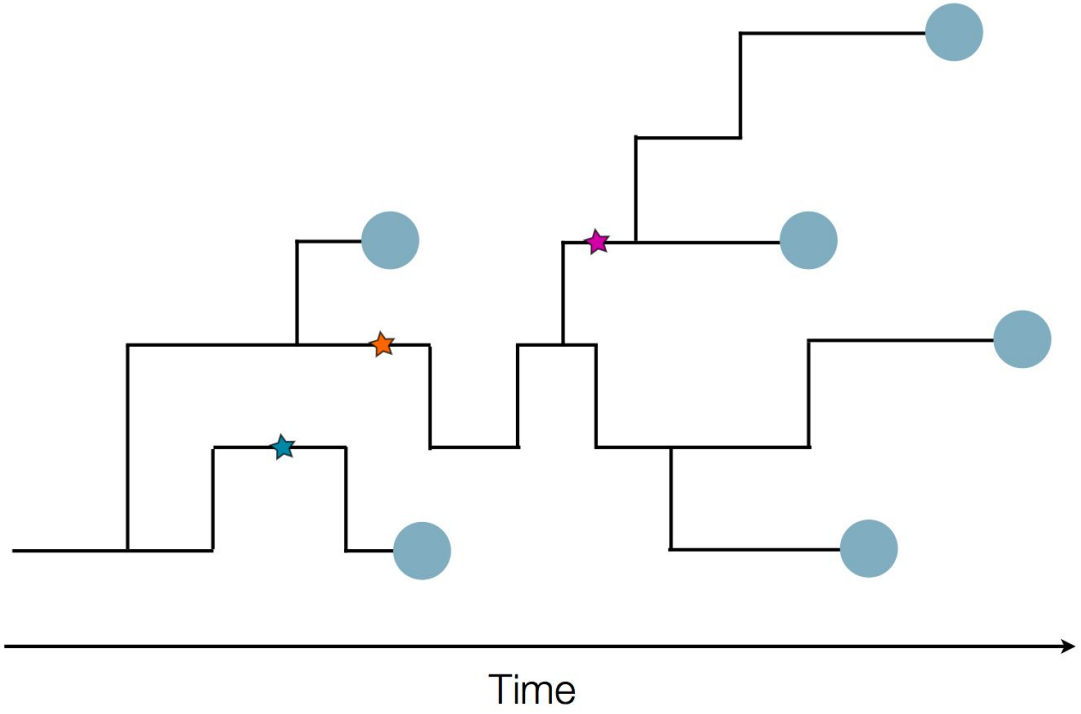
https://github.com/trvrb/phylodynamics-lecture

Trees sample the underlying transmission network



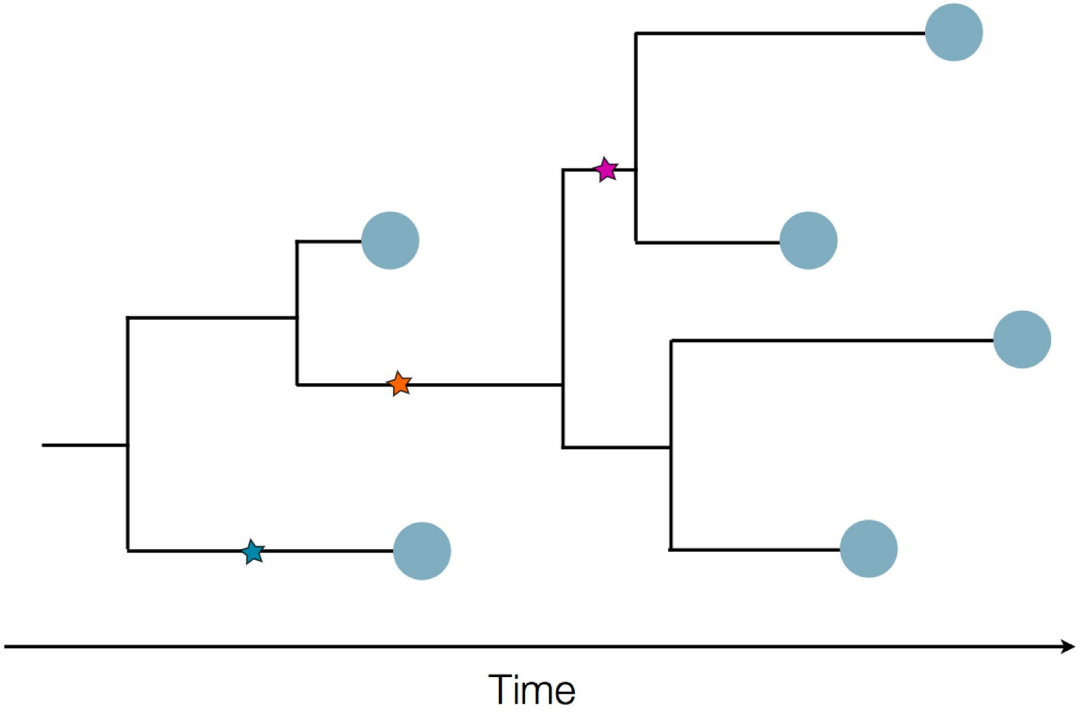
https://github.com/trvrb/phylodynamics-lecture

Trees sample the underlying transmission network



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

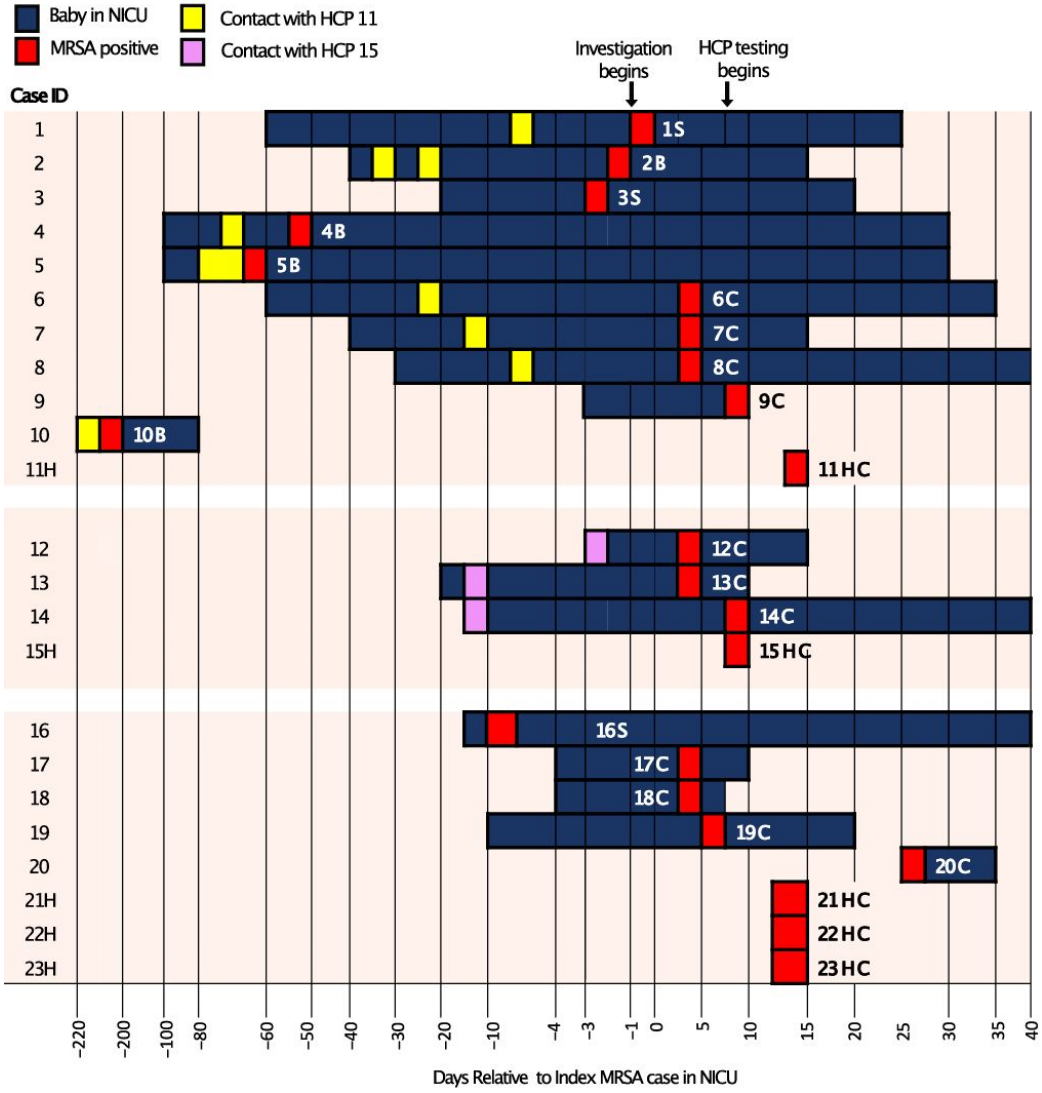
Trees sample the underlying transmission network



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

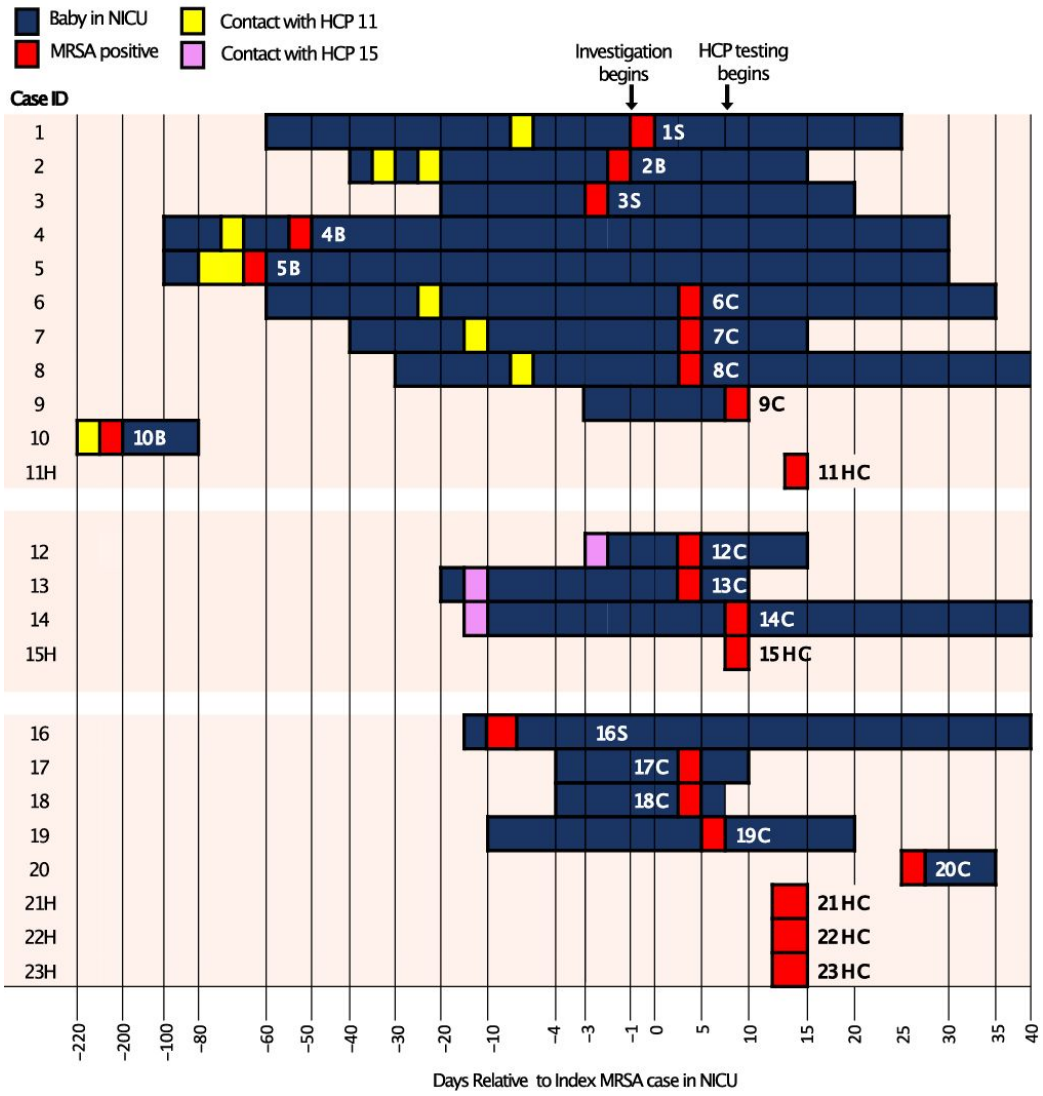
What do our genomes & trees tell us about our NICU-MRSA outbreak?

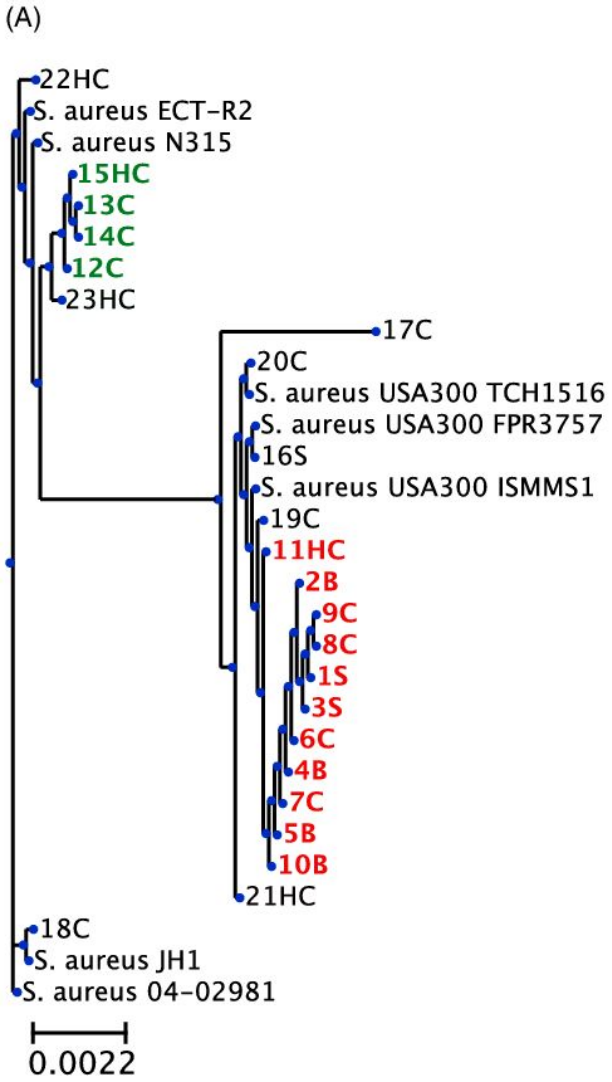
Genomics tells us there are 2 outbreaks!



Madera, Sharline, et al. 2023

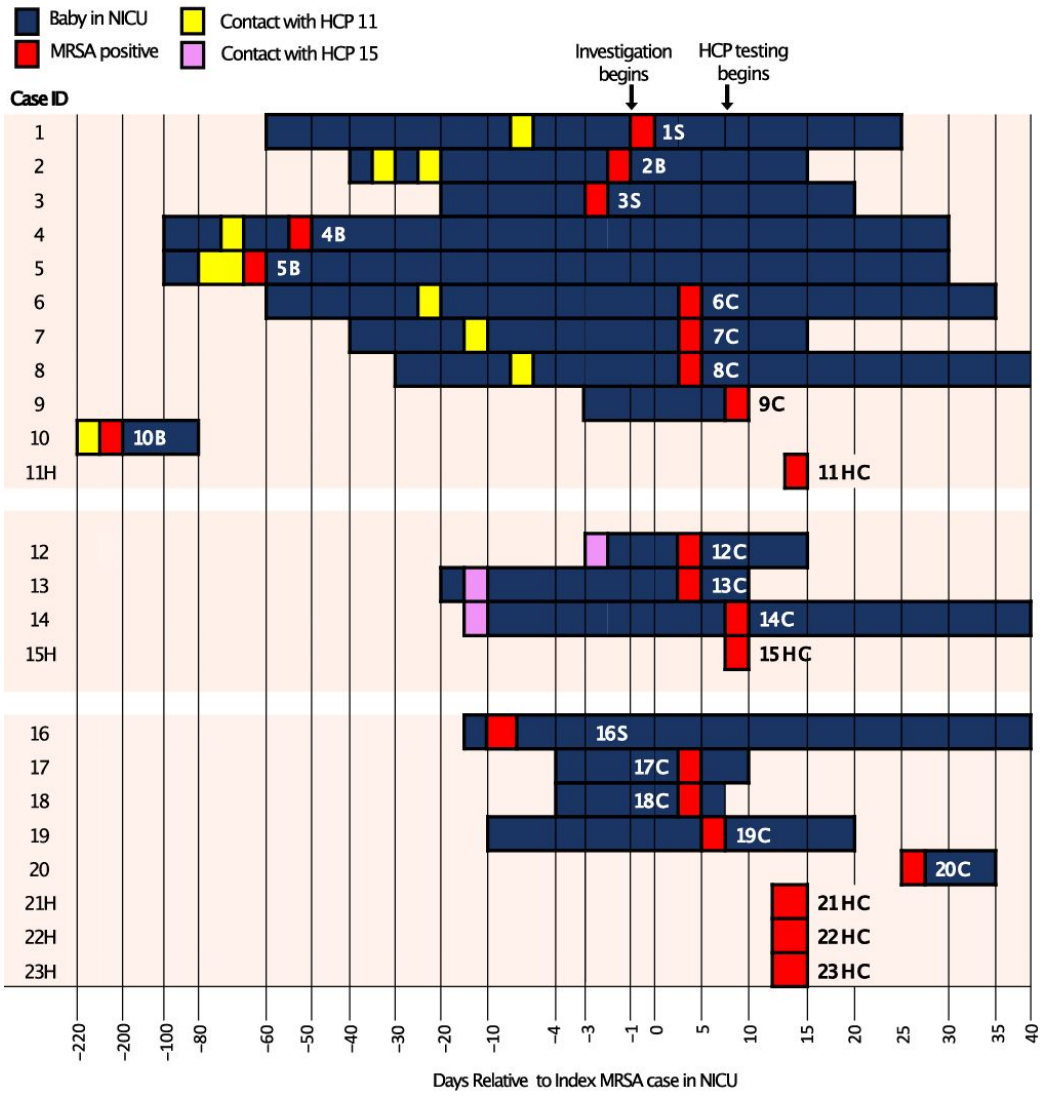
Genomics tells us there are 2 outbreaks!

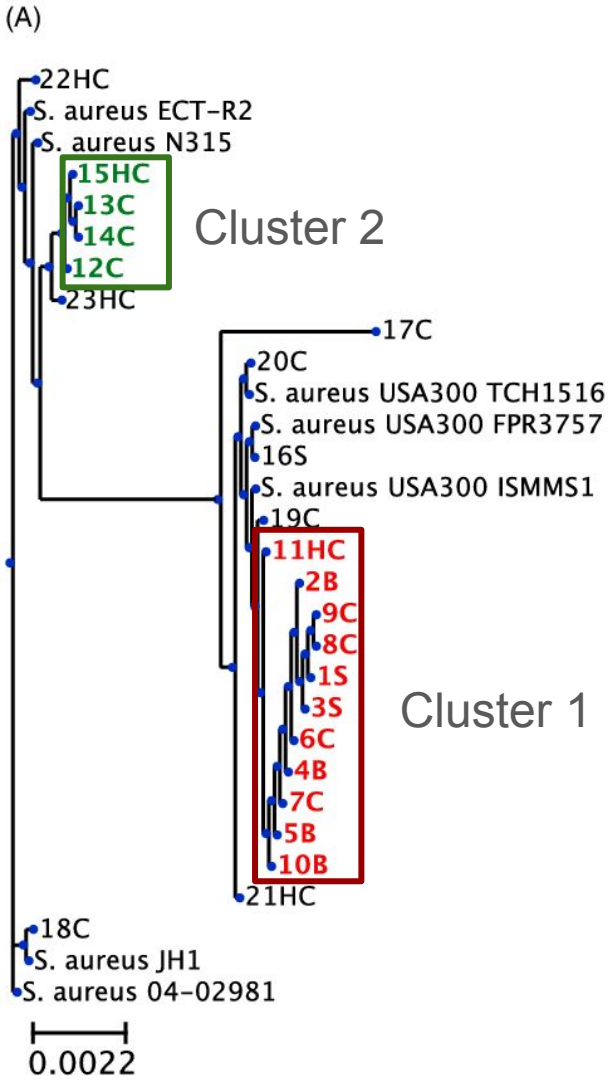




Madera, Sharline, et al. 2023

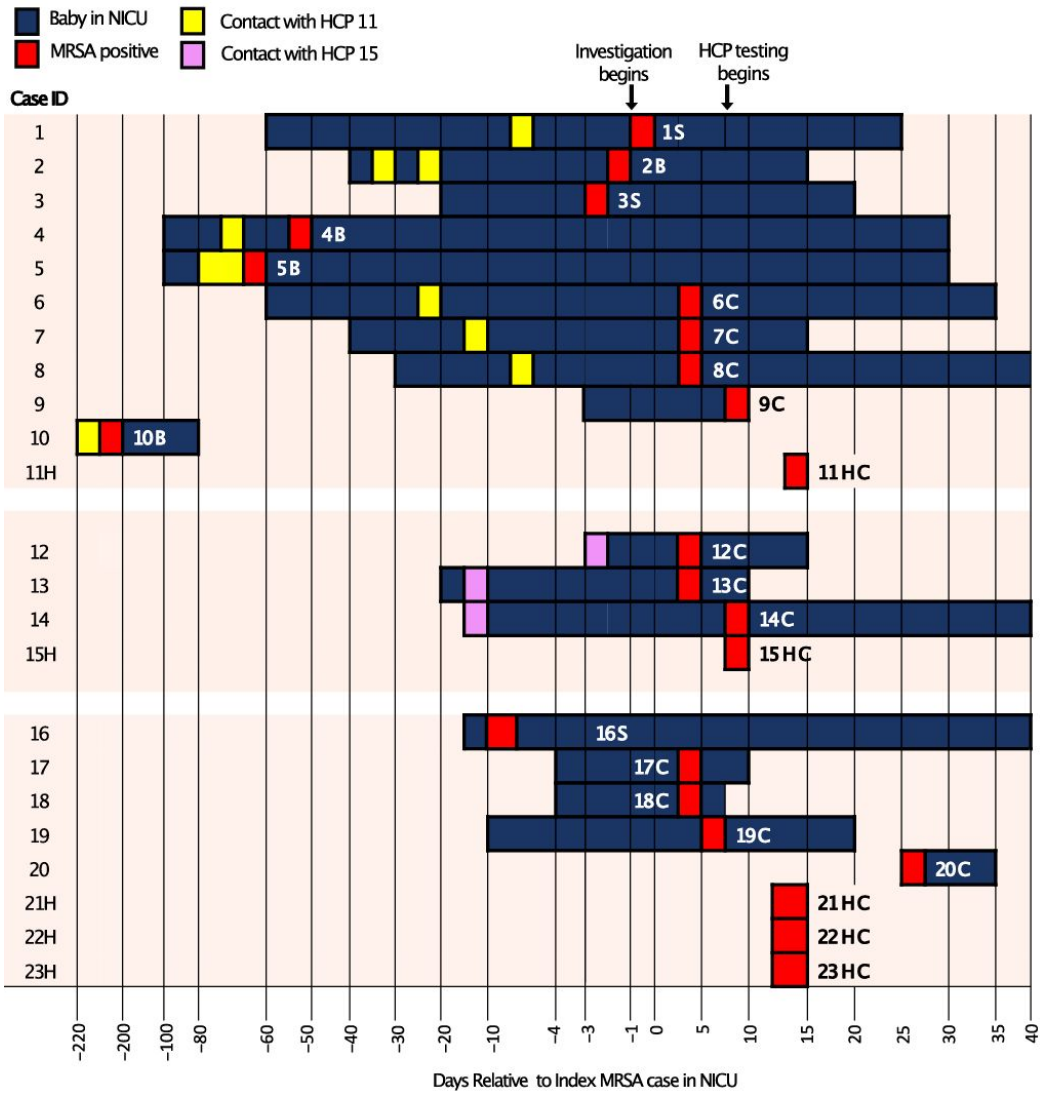
Genomics tells us there are 2 outbreaks!

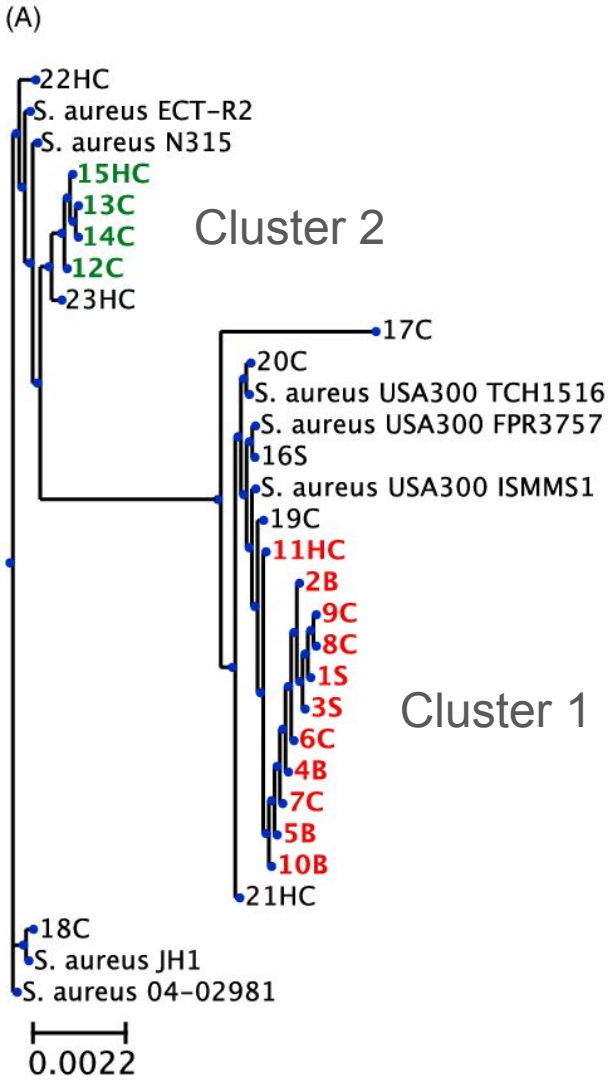




Madera, Sharline, et al. 2023

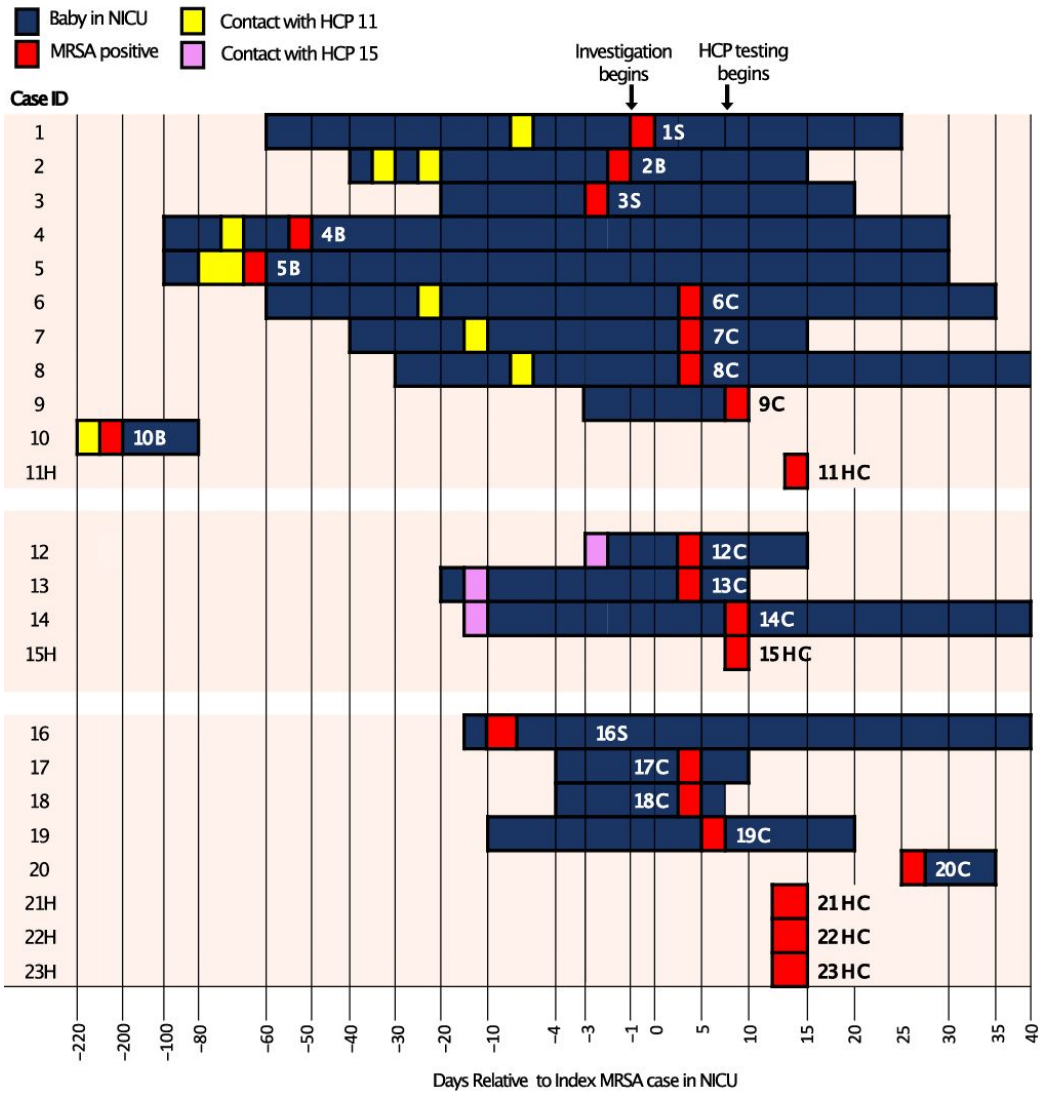
Genomics tells us there are 2 outbreaks!

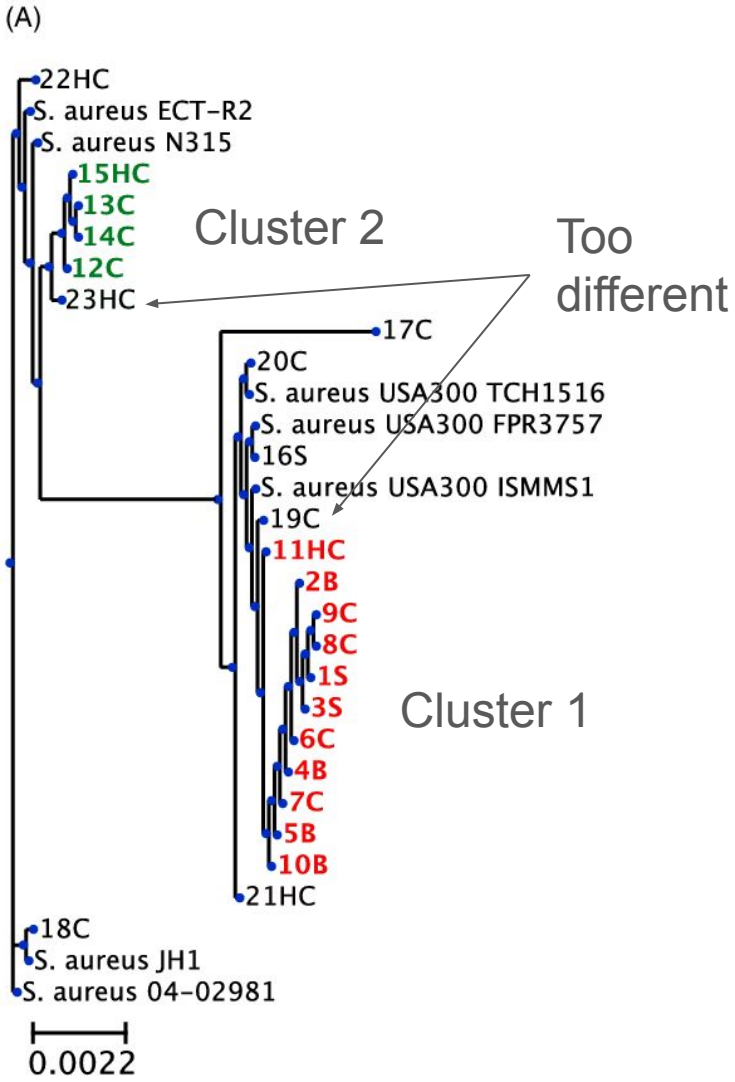




Madera, Sharline, et al. 2023

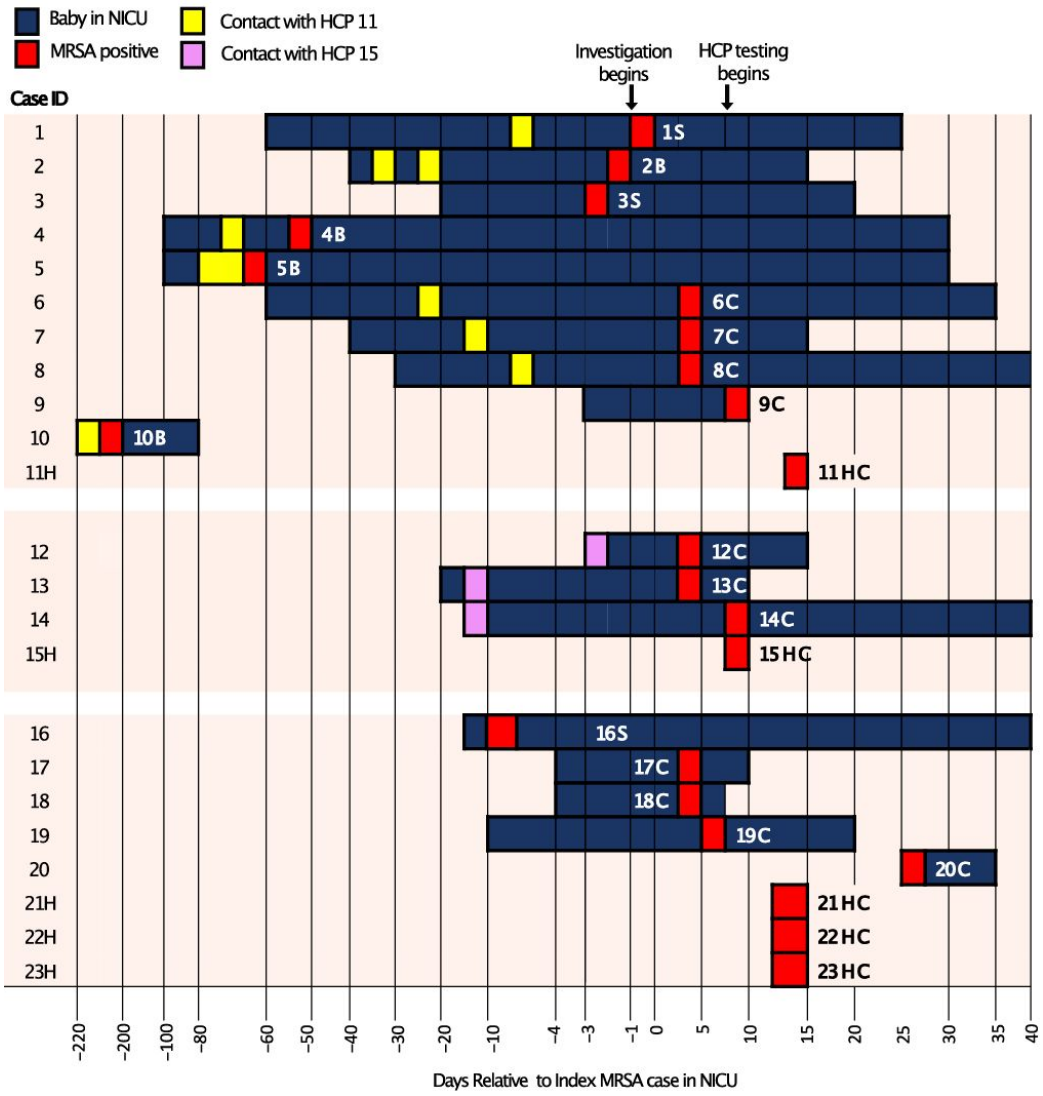
Genomics tells us there are 2 outbreaks!

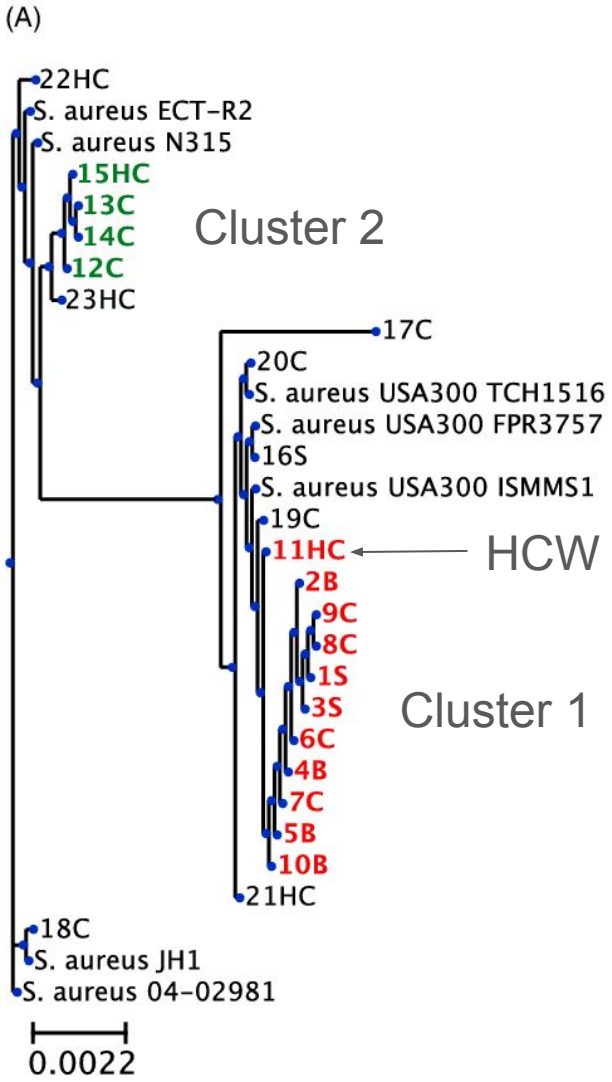




Madera, Sharline, et al. 2023

Genomics tells us HCW11 likely source of Cluster 1





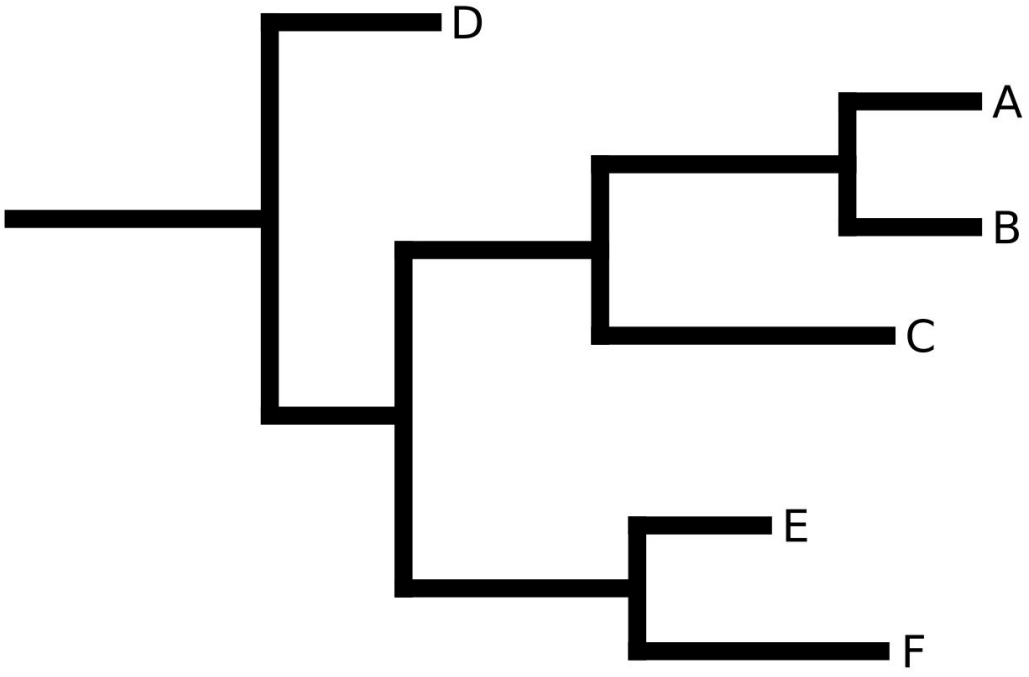
Madera, Sharline, et al. 2023

Conclusions

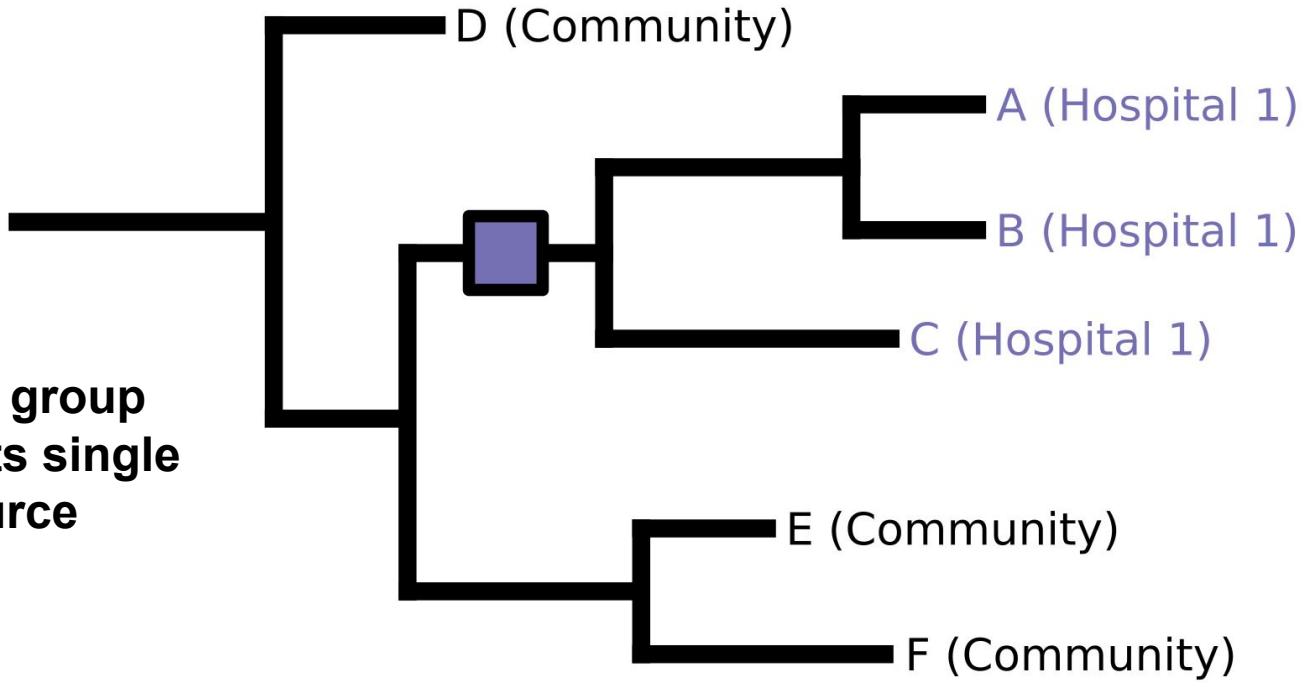
- Infection Prevention and Control (IPAC) increasingly incorporates genomics
- Clinical microbiology can involve differences in scale to normal lab work
- Sequencing involves randomly* sampling many short (2nd generation) or fewer long noisy (3rd generation) “reads” from DNA
- Sequencing is a physical process so involves measurement error
- Reference-based assembly: comparing reads to reference and trying to distinguish real changes from errors.
- *de novo* assembly: stitching together reads* which share sequences.
- Comparing genomes by alignment lets you find shared/different base pairs.
- Phylogenetic trees can be inferred from these patterns.
- Trees represent a sampling from the underlying evolving population.
- Genomes and trees can be used to track and stop outbreaks (among many other things!).

Extra Slides

So, tree tells us about possible outbreak source

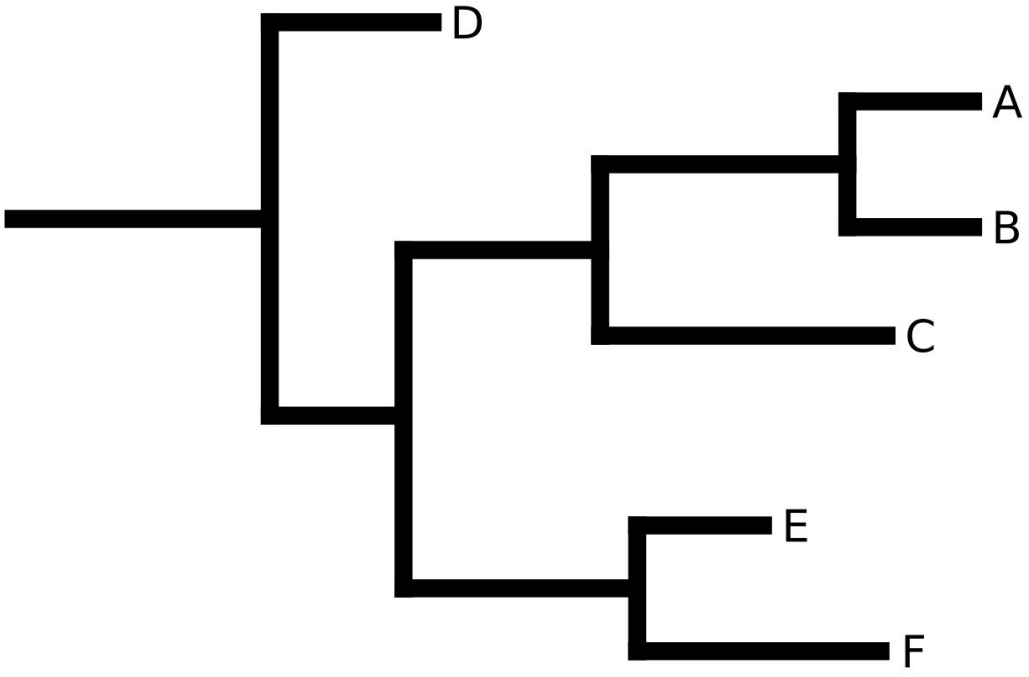


So, tree tells us about possible outbreak source

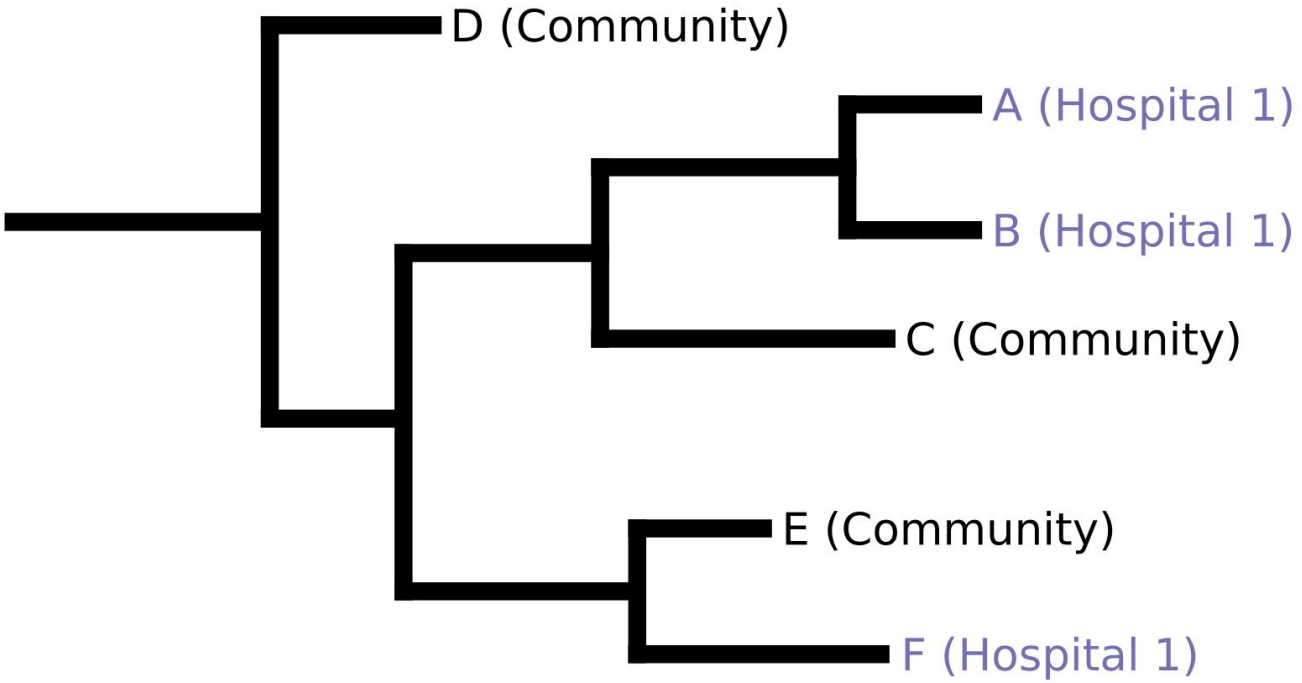


Single group suggests single source

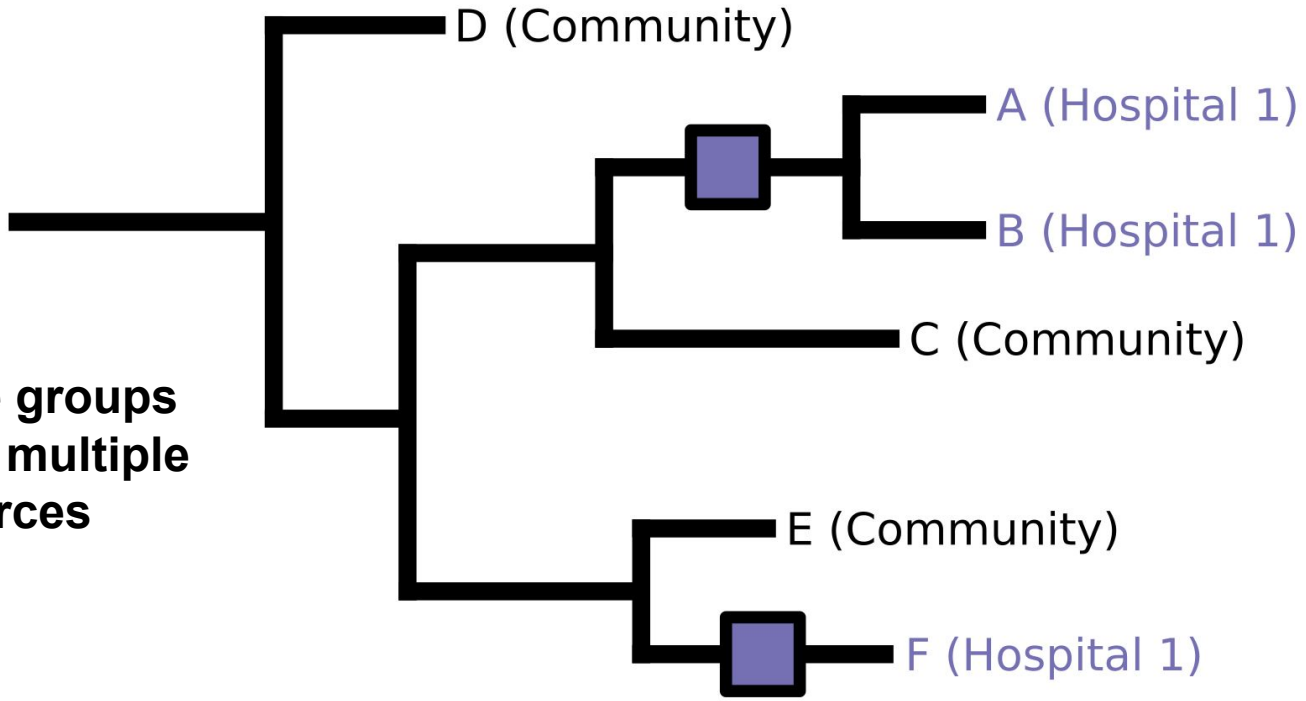
So, tree tells us about possible outbreak source



So, tree tells us about possible outbreak source

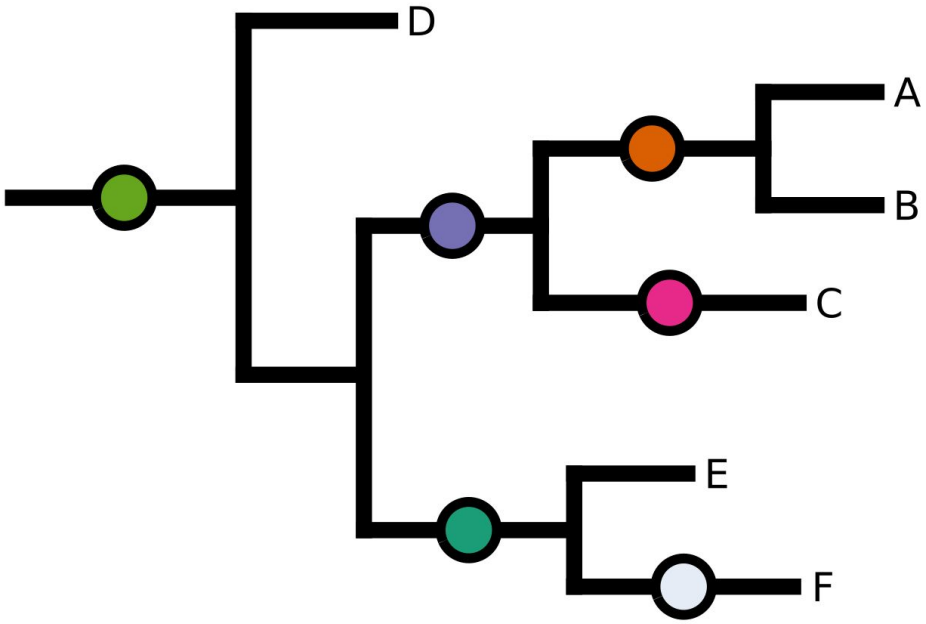


So, tree tells us about possible outbreak source

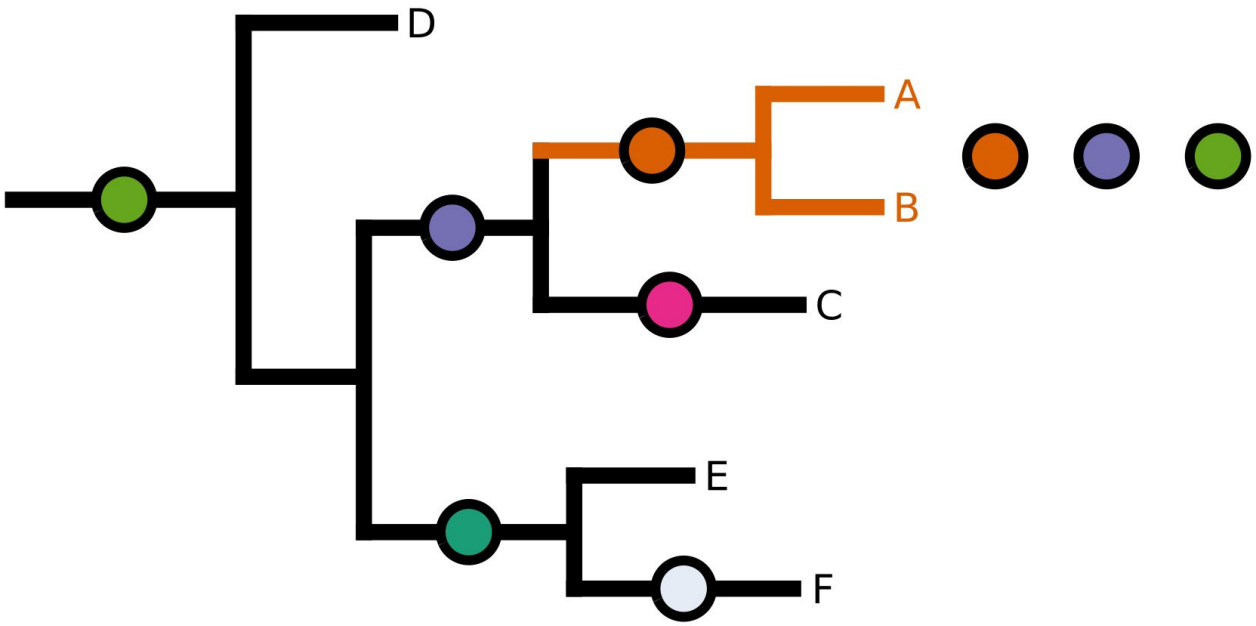


Multiple groups suggest multiple sources

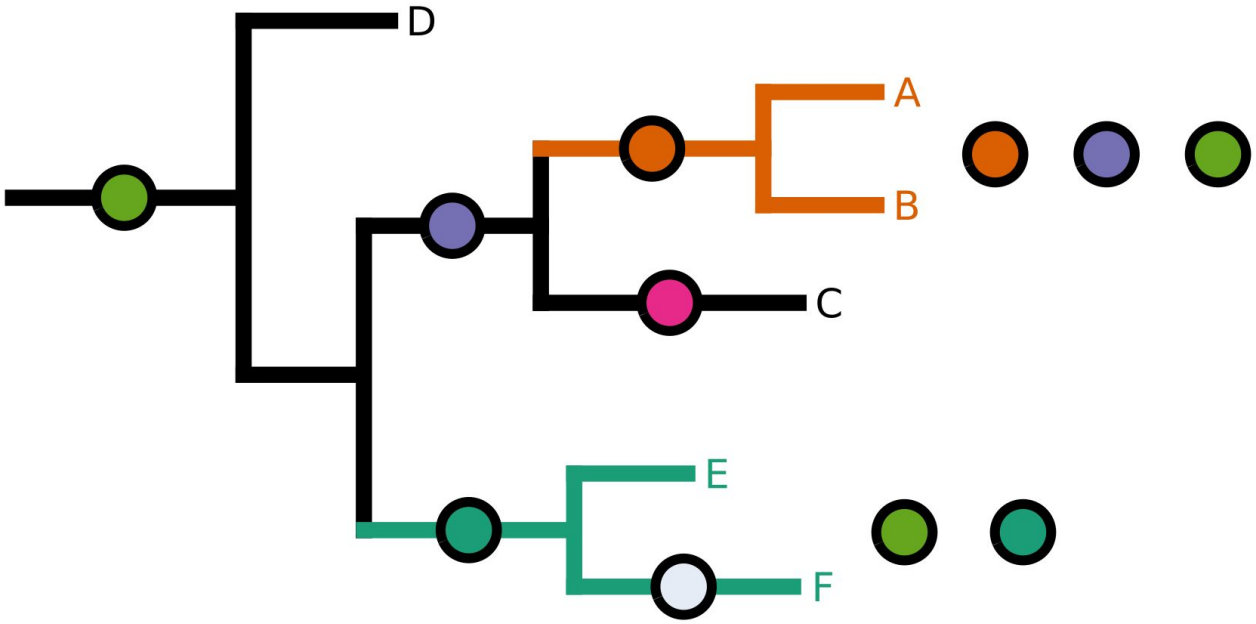
Define lineages (groups) of pathogens



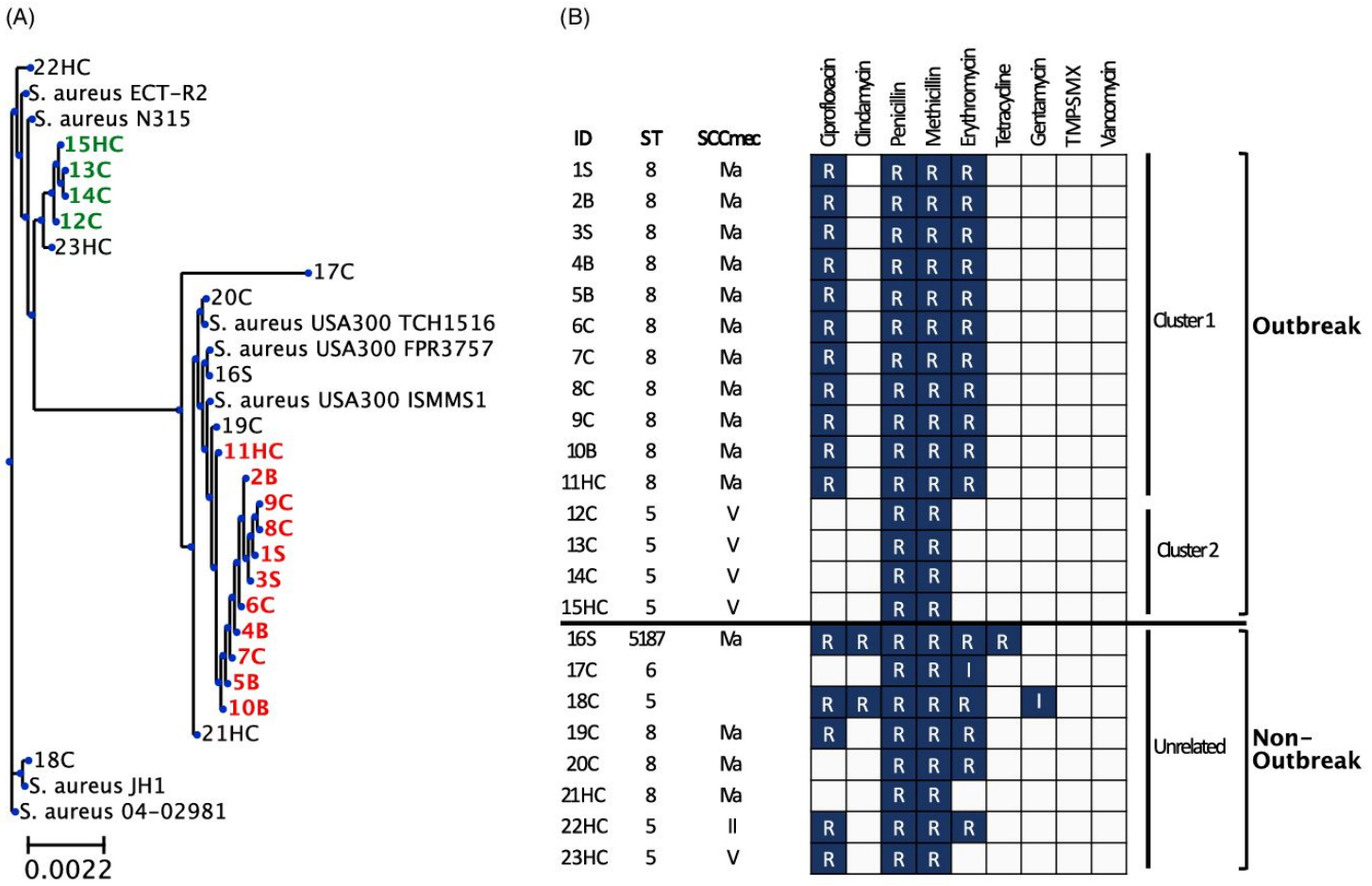
Can also define lineages (groups) of pathogens



Can also define lineages (groups) of pathogens



Lineages, typing, and phenotyping



Madera, Sharline, et al. 2023