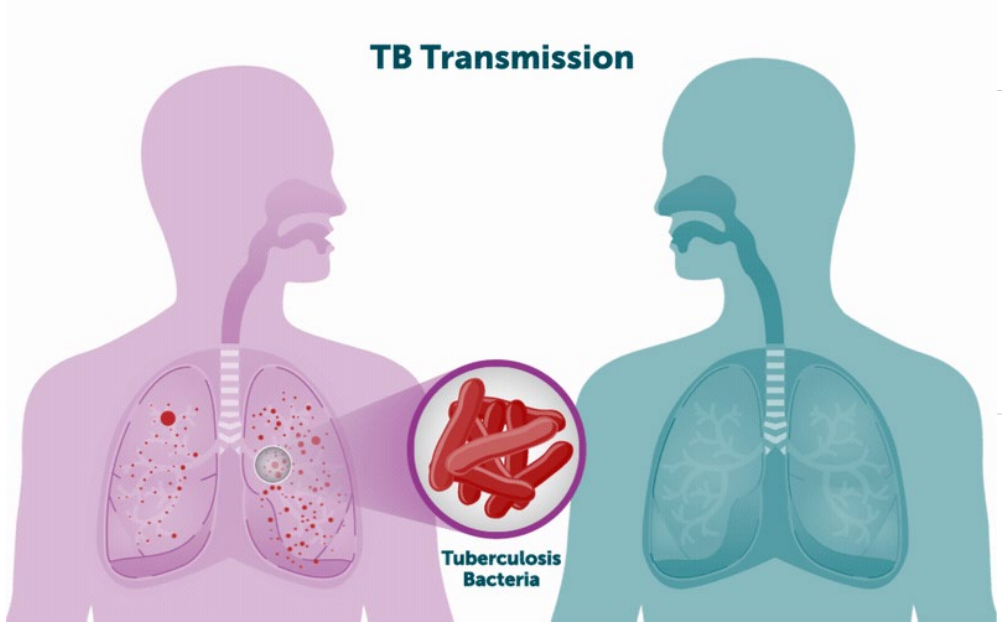
Tuberculosis Molecular Epidemiology: Deciphering Genotyping and Whole Genome Sequencing

Sarah Talarico, PhD, MPH

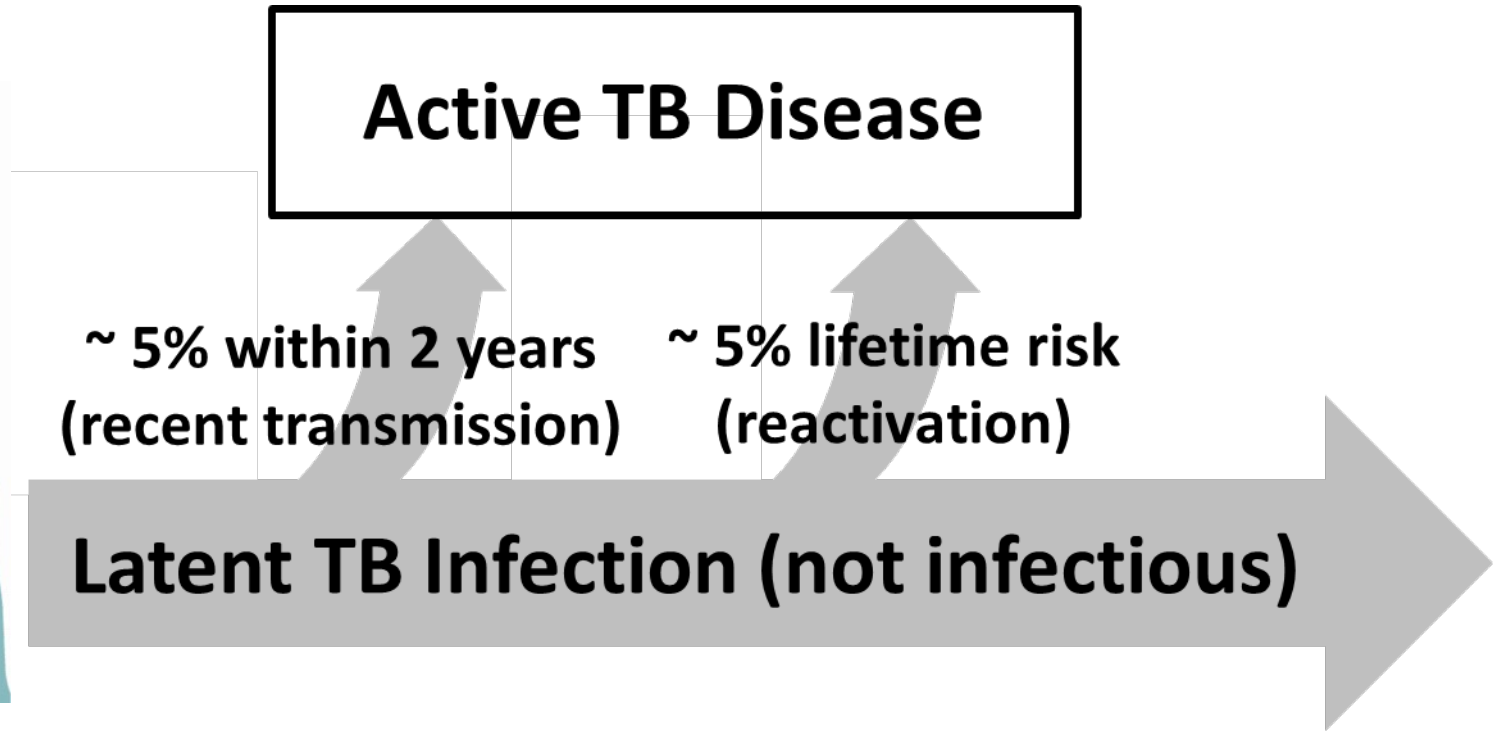
Molecular Epidemiology and Outbreak Investigations Team
Surveillance, Epidemiology, and Outbreak Investigations Branch
Division of Tuberculosis Elimination

October 23, 2024

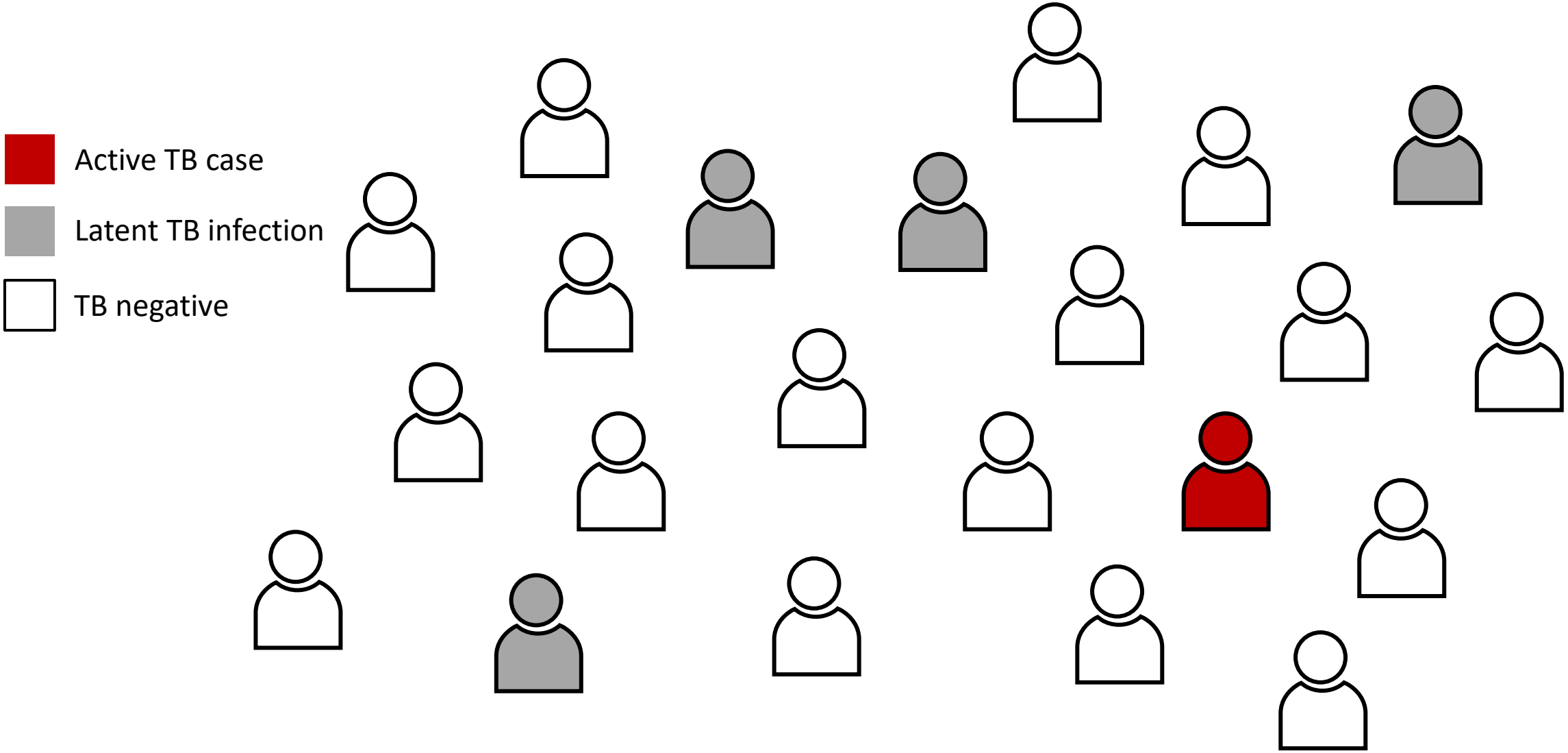
TB course of infection: latent TB infection and active TB disease



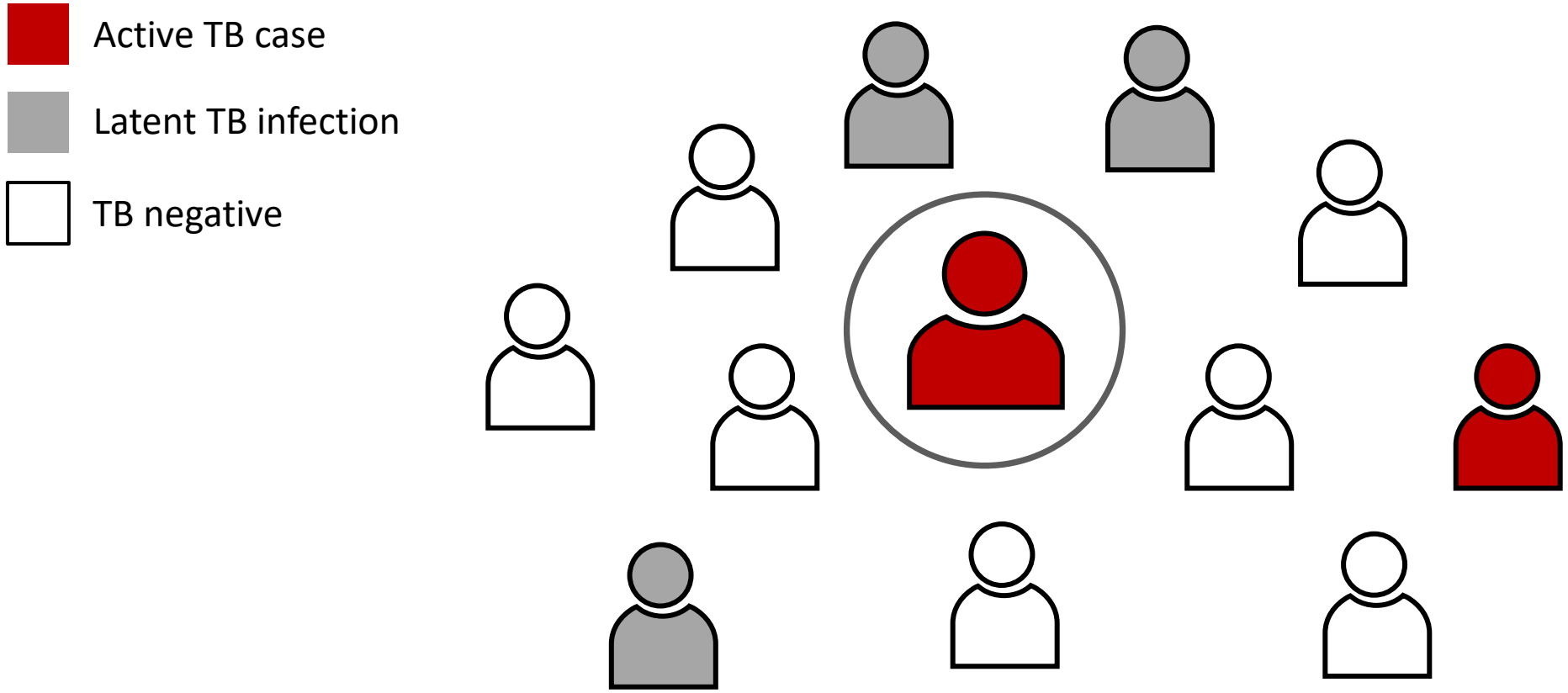
Person with active pulmonary TB disease (infectious)



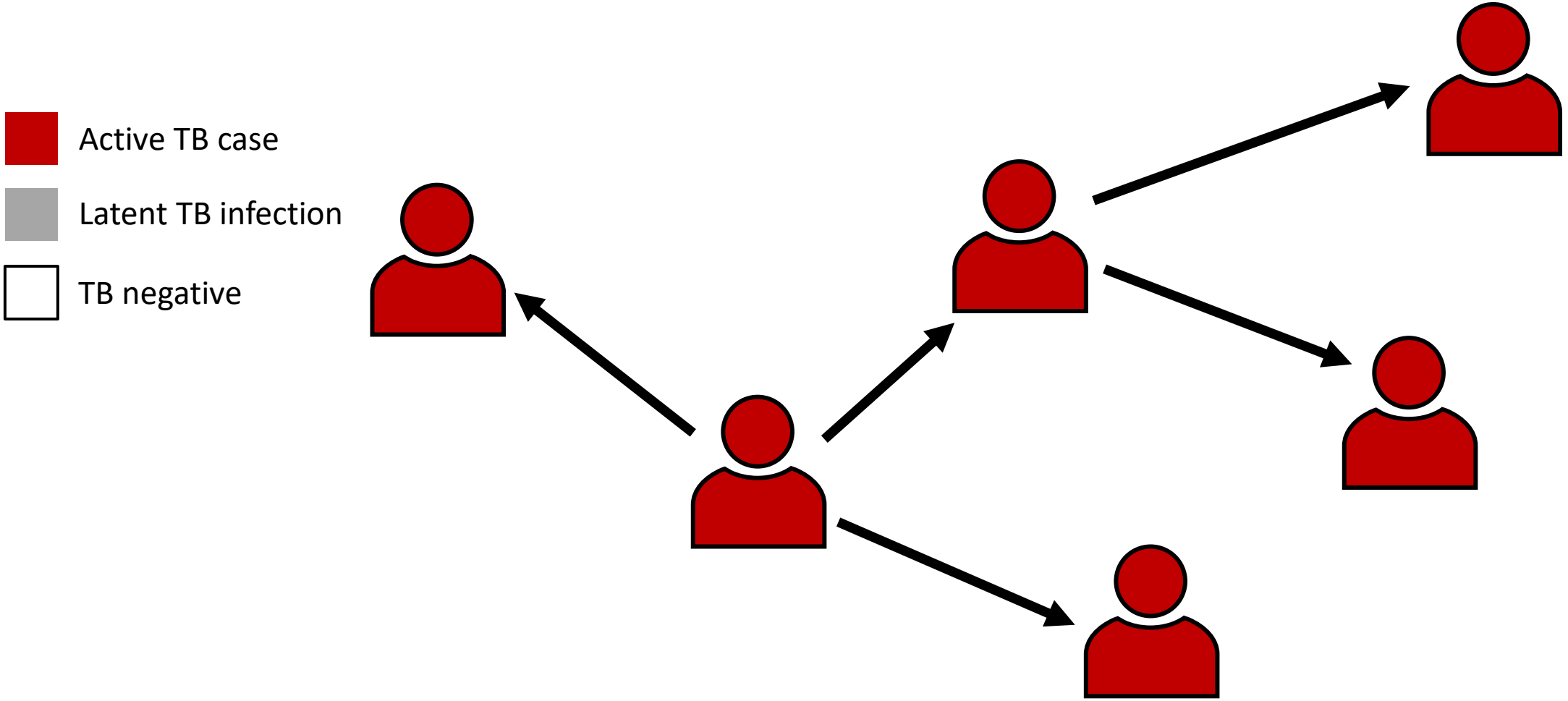
TB screening



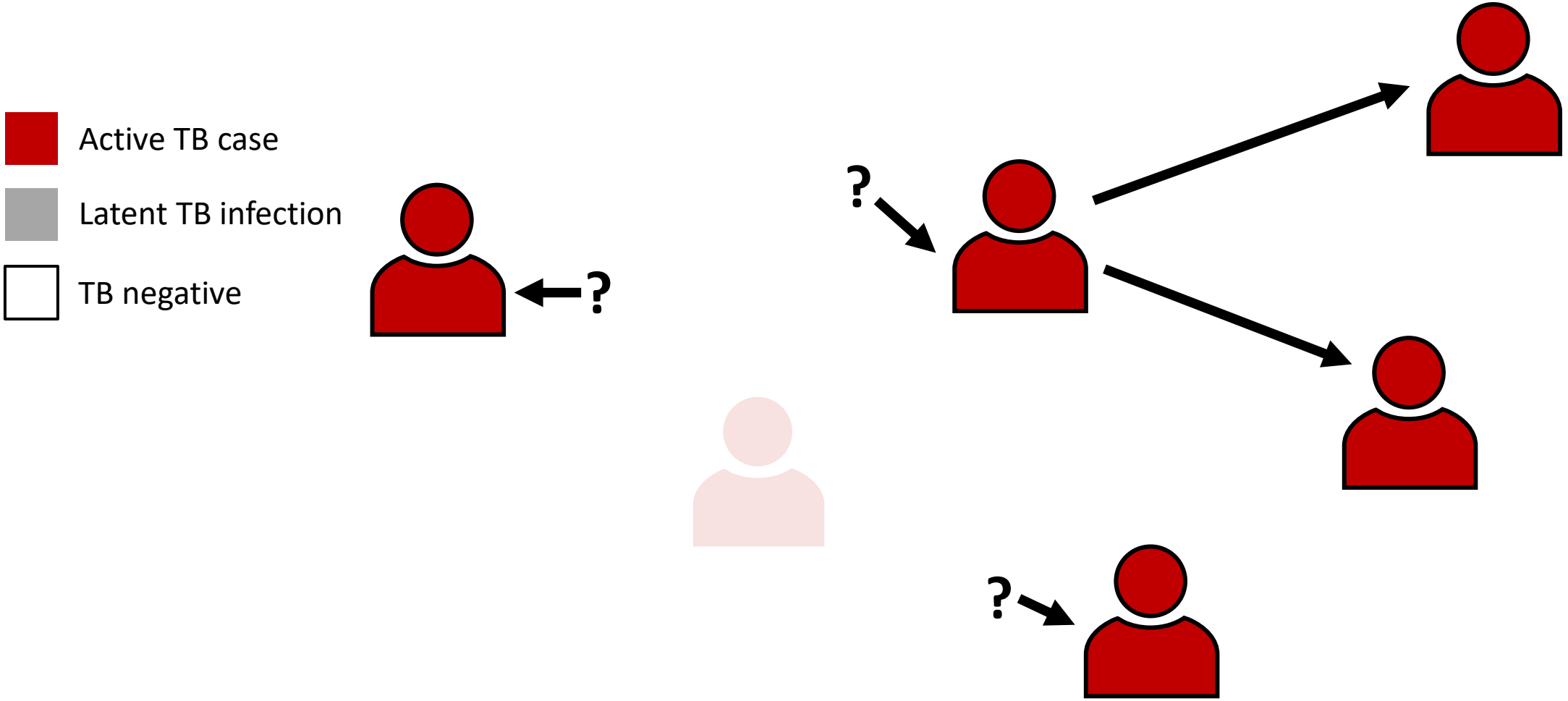
TB contact investigation



TB cluster or outbreak investigation



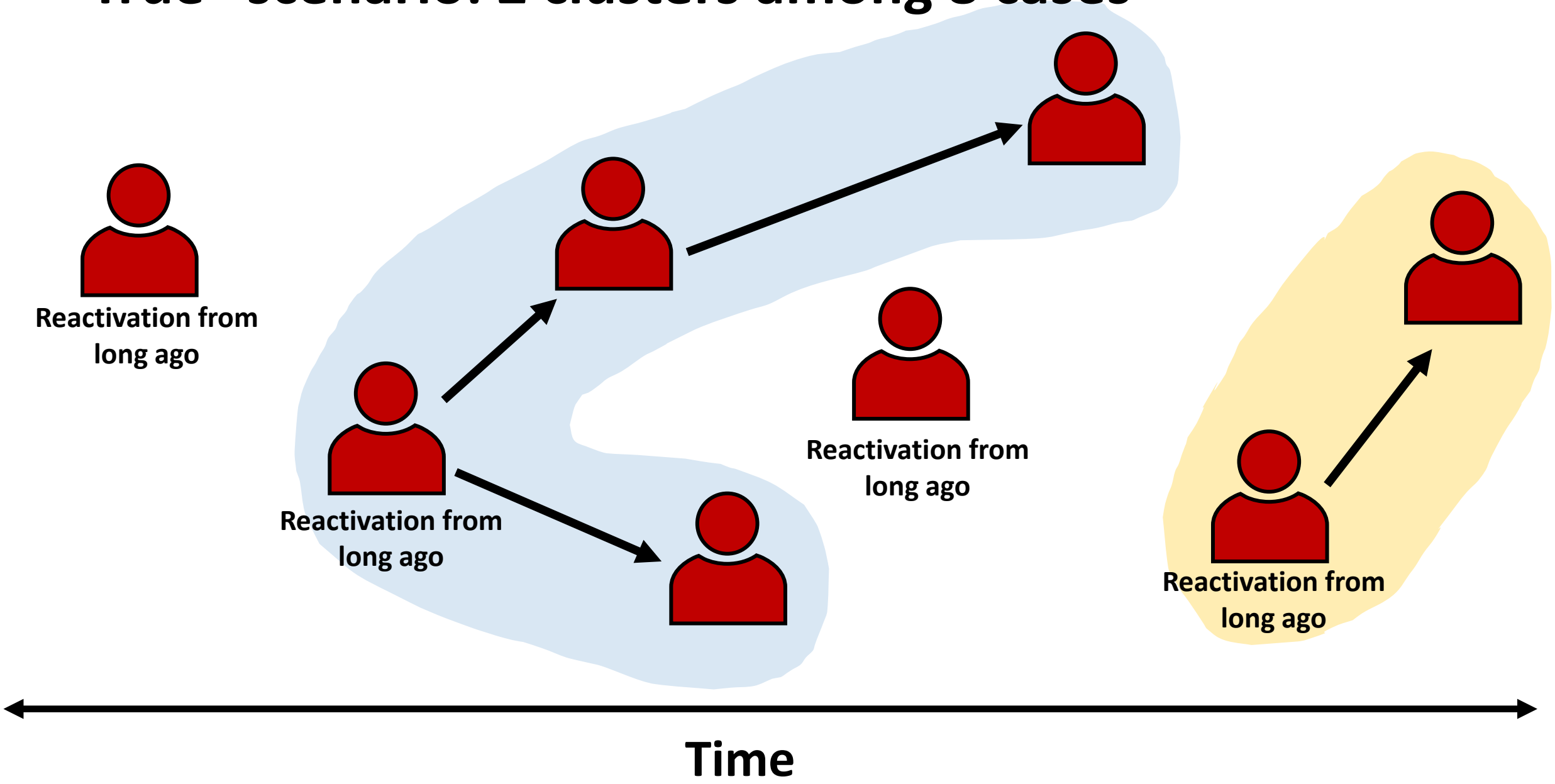
TB cluster or outbreak investigation



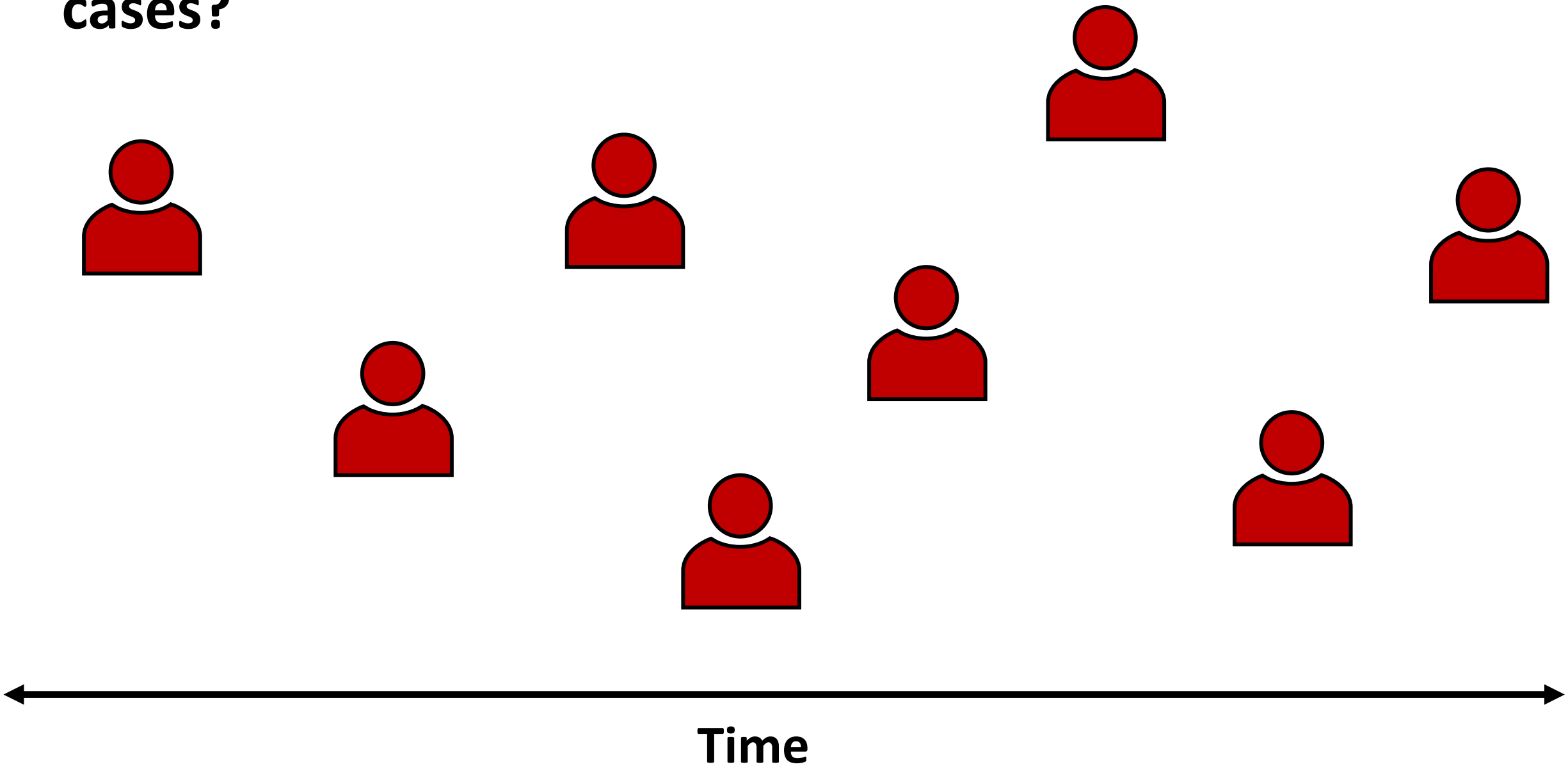
How do we know that there is a TB cluster or outbreak occurring?

- Goal
  - Reduce the burden of TB by identifying where transmission is currently occurring and interrupting it
- Challenge
  - Distinguish recent transmission from cases infected long ago

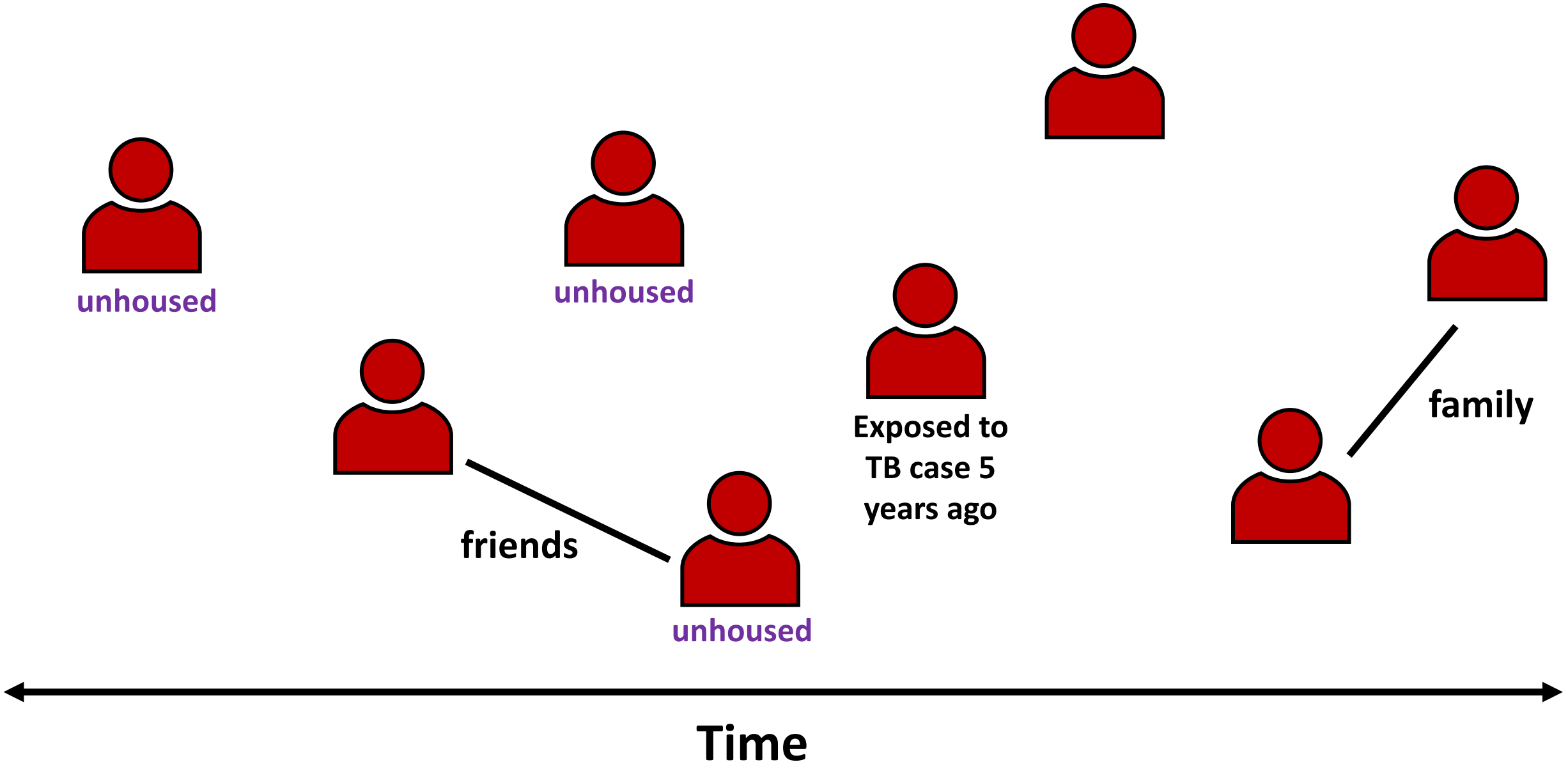
“True” scenario: 2 clusters among 8 cases



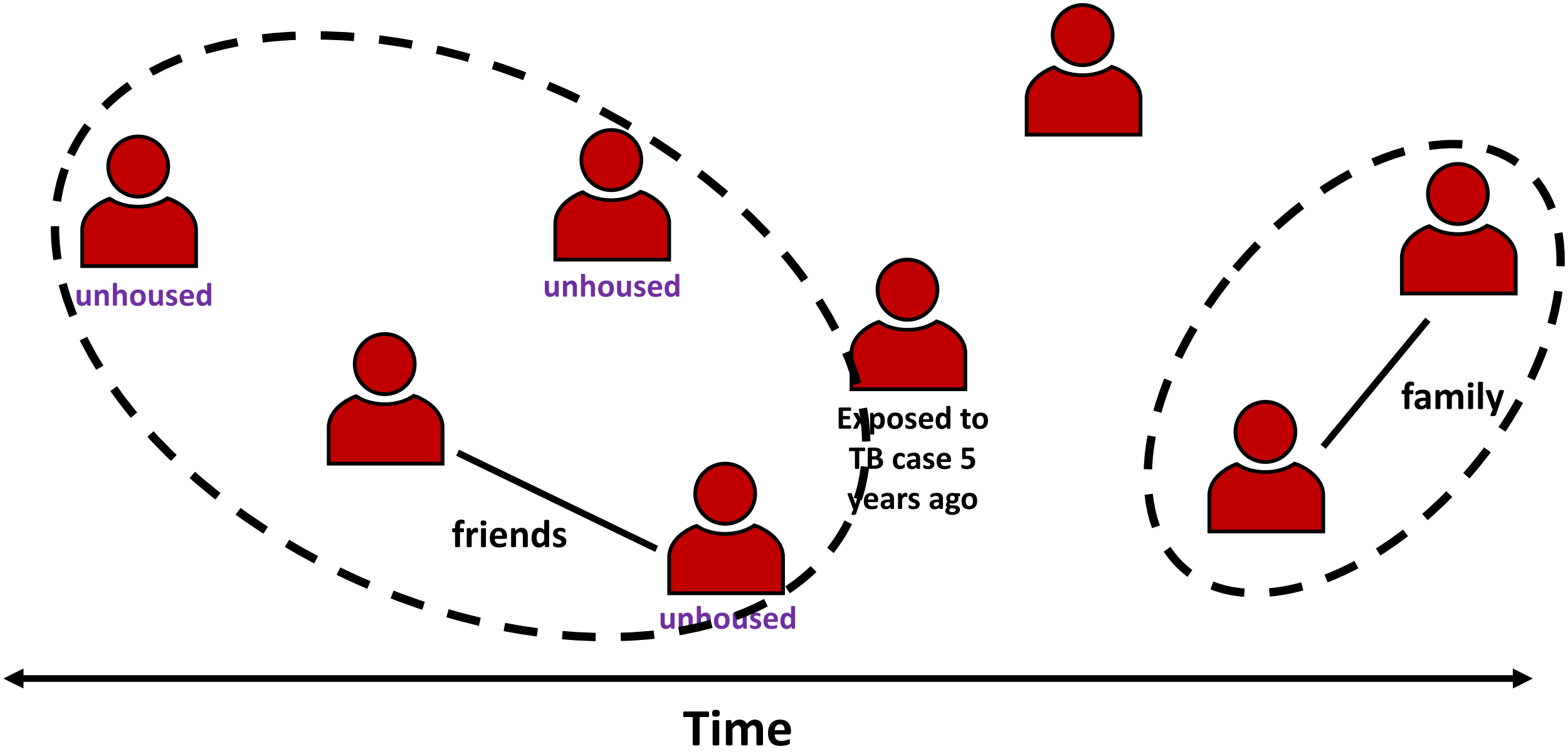
How to identify the clusters when all we can see is cases?



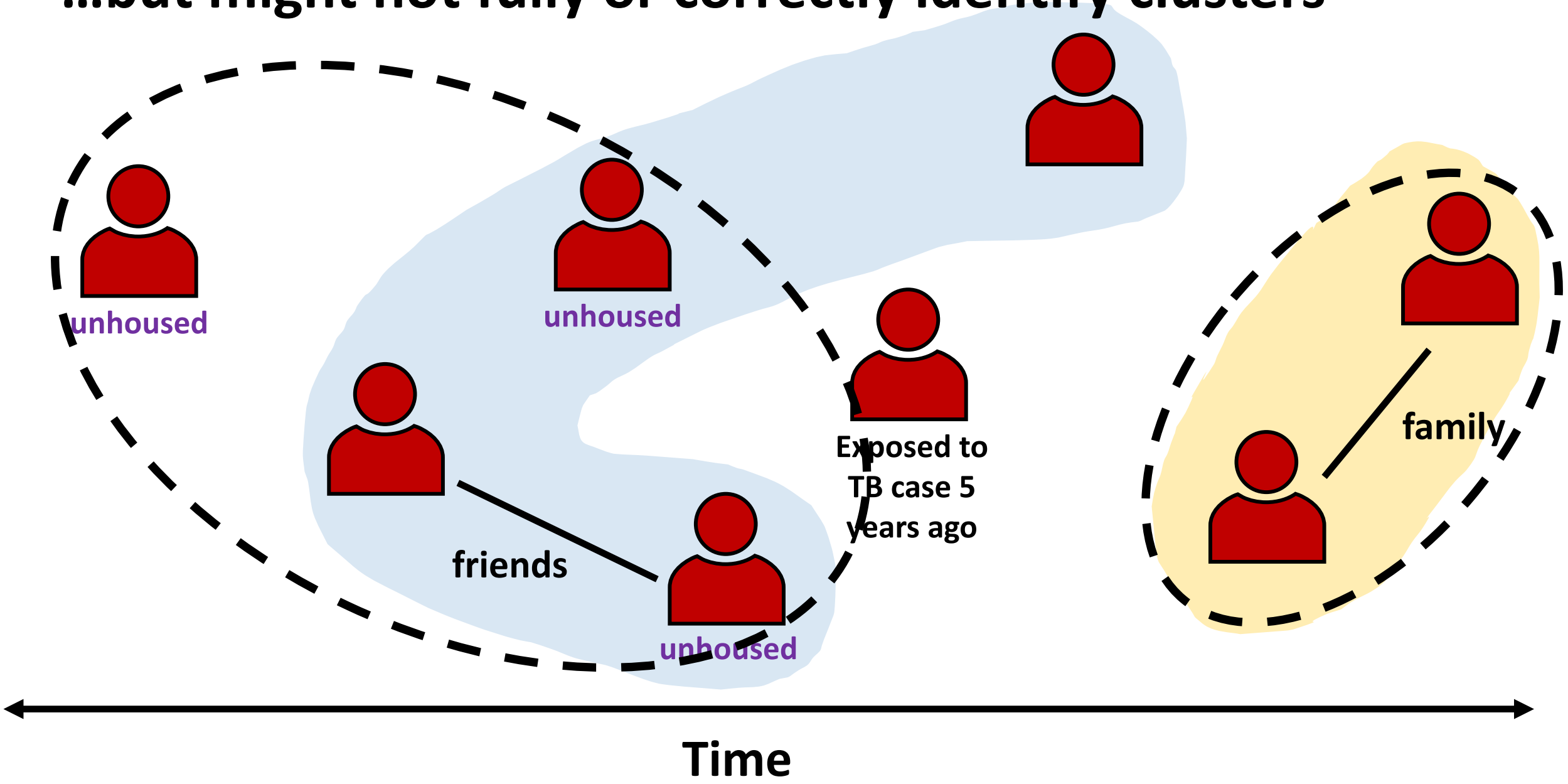
Having some epi data can help…



Having some epi data can help…



…but might not fully or correctly identify clusters



Challenges to relying exclusively on epidemiologic investigation

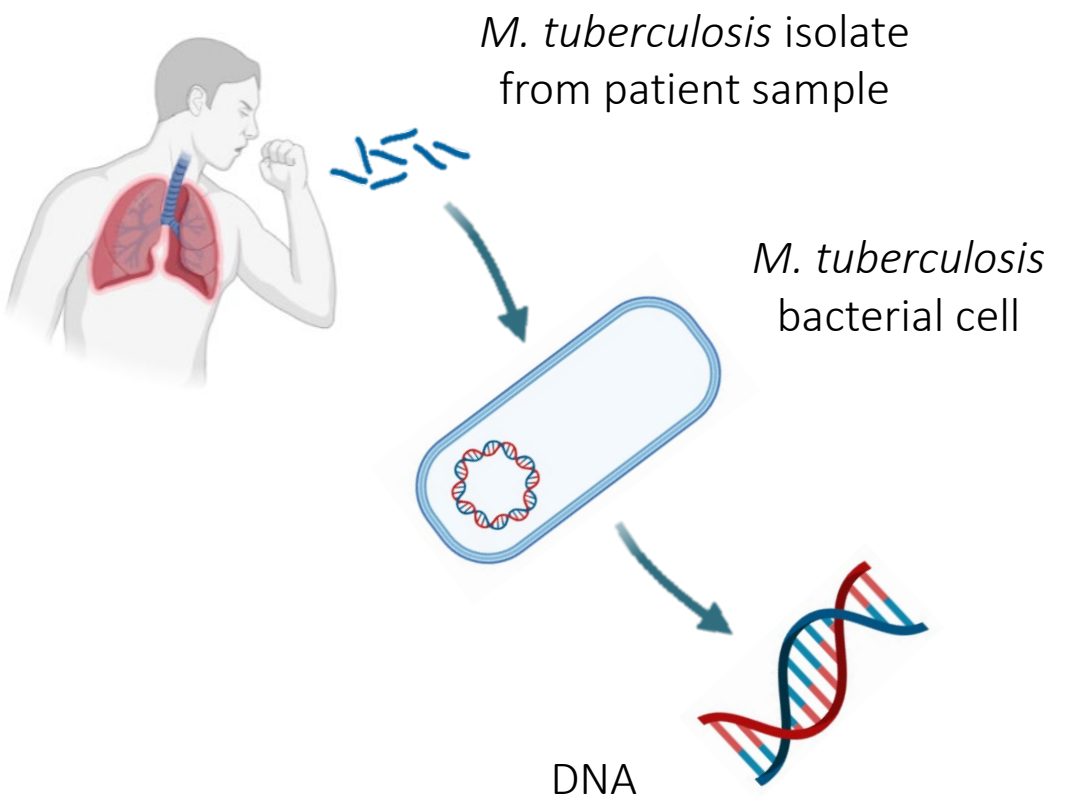
- Airborne transmission
- Exposure in congregate settings
- Long infectious periods
- Patient recall may be incomplete or unreliable
- Often in impoverished or marginalized communities

How do we know that there is a TB cluster or outbreak occurring?

- Goal
  - Reduce the burden of TB by identifying where transmission is currently occurring and interrupting it
- Challenge
  - Distinguish recent transmission from cases infected long ago
- Approach
  - Molecular epidemiology!
  - Use molecular genotyping data, combined with clinical and epidemiologic data, to detect, investigate, and monitor recent TB transmission

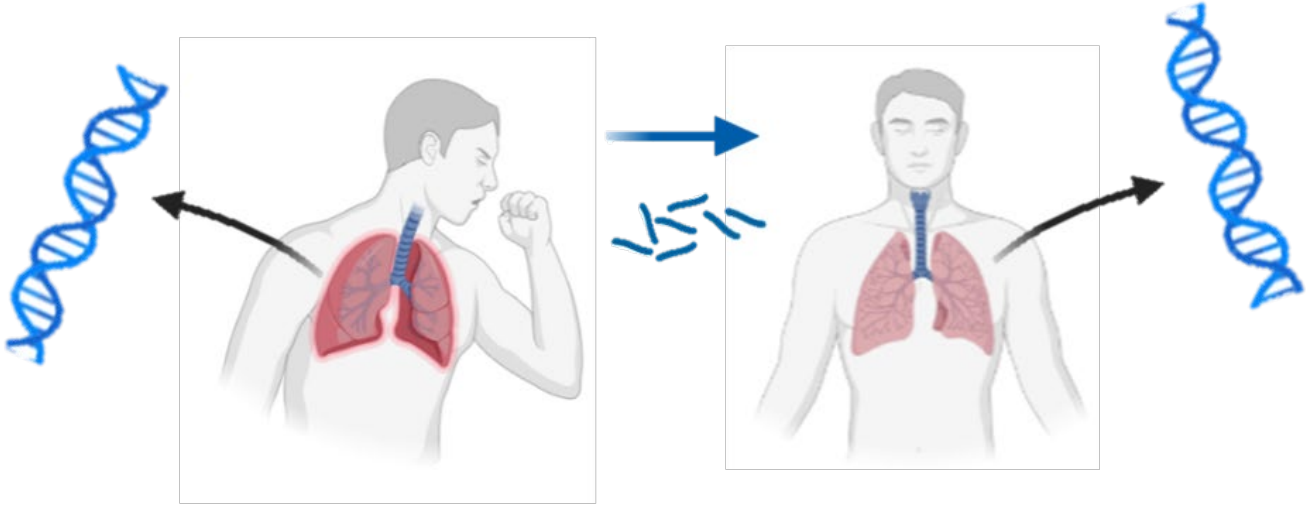
TB genotyping examines the DNA of M. tuberculosis isolates from TB patients

- The *M. tuberculosis* bacteria cultured from a TB patient is called the patient's isolate
- Bacteria, including *M. tuberculosis*, have DNA called a genome
- DNA is made up of four different nucleotides (abbreviated A, T, C, and G)
- The order of these nucleotides in the genome is the DNA sequence
- The genome of *M. tuberculosis* is over 4.4 million nucleotides long

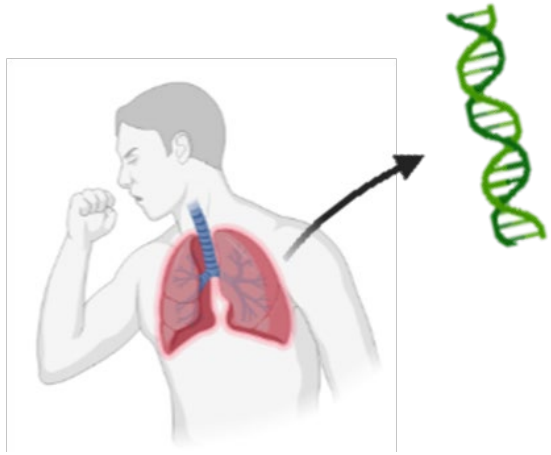


DNA sequence

Genotyping analyzes DNA to identify TB patients with similar *M. tuberculosis* genomes who are more likely to be linked by recent transmission

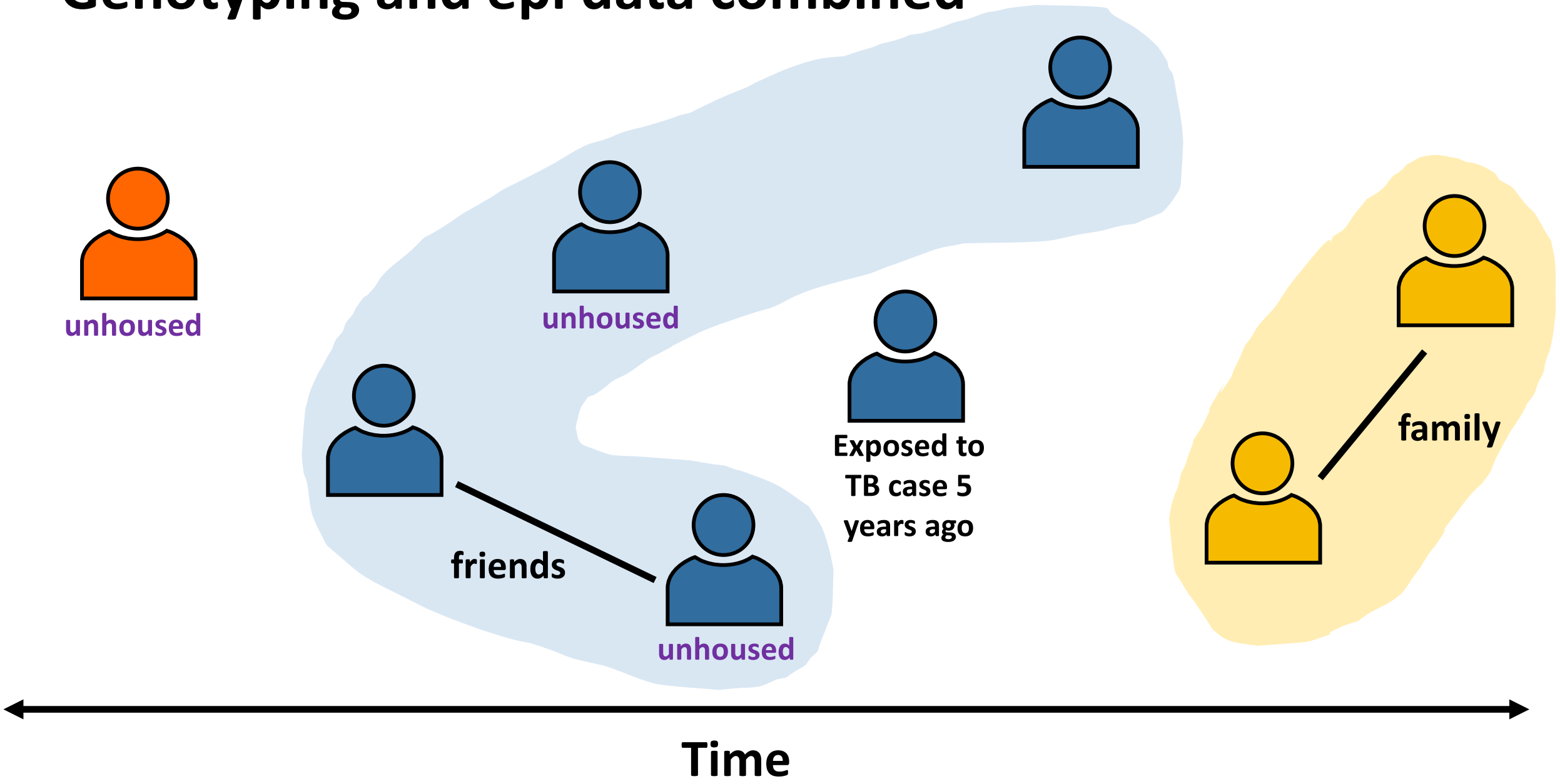


TB patients linked by recent transmission have isolates with the same genotype (blue)



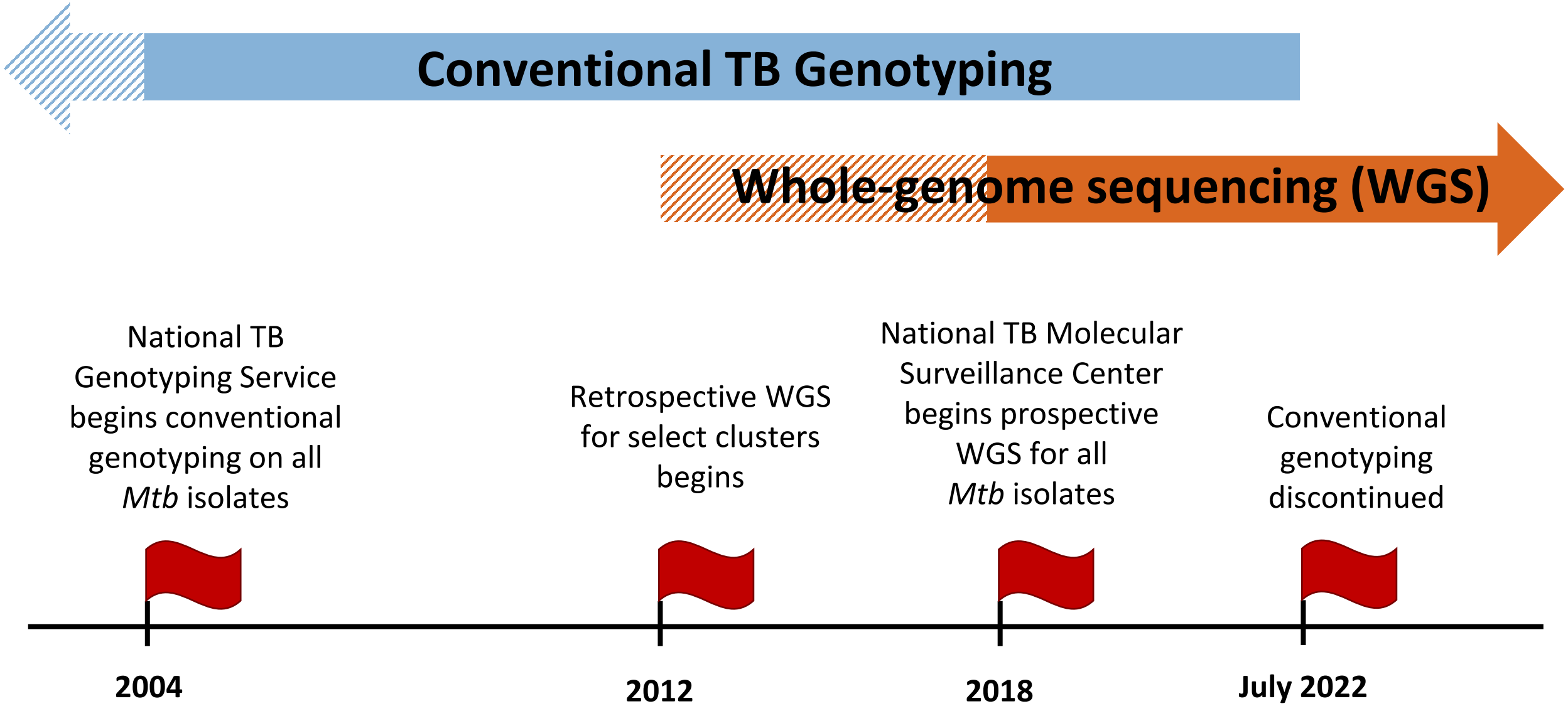
TB patient not linked by recent transmission has an isolate with a different genotype (green)

Genotyping and epi data combined

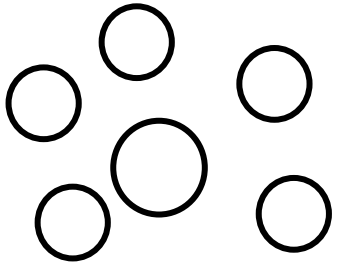


TB Genotyping Methods

TB genotyping through the years



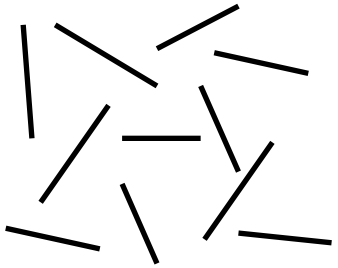
Whole-genome sequencing (WGS) of *M. tuberculosis*



Genomic DNA

~ 4.4 million basepairs

Shear DNA



Genomic DNA fragments

~ 500 basepairs

Create library and sequence

```
@NB551186:40:H5TN5AFXY:1:1110
1:21172:1116 1:N:0:
GGACTCCT+TATGCAGTTACGGAACC
CAATCAGGTCCAAAGGTCTTCATCAA
GGCGTCGGAAAGCACGTCGATAACA
GCGTCGCTCTGTTCGTTGGTTGGCTT
+
A/AAAEAEEEEEEEEE6/EEEEEAAE/A/
/E/AEEE/EEAAE/EEEEEE/AEEAE/EEE
//AEE6EAEEA//<</<<E/EEE<<//
```

File of sequence reads

Data transfer to CDC

WGS data analysis

WGS data analysis focuses on a type of genetic variation called a single nucleotide polymorphism (SNP)

A T G G C G T C A C G G T C A G

Example: mutation that changes C to T in DNA sequence

A T G G C G T T A C G G T C A G

Single nucleotide polymorphisms (SNPs) throughout the genome

WGS data can be used for many different types of analyses

```
@NB551186:40:H5TN5AFXY:1:1110
1:21172:1116 1:N:0:
GGACTCCT+TATGCAGTTACGGAACC
CAATCAGGTCCAAAGGTCTTCATCAA
GGCGTCGGAAAGCACGTCGATAACA
GCGTCGCTCTGTTCGTTGGTTGGCTT
+
A/AAAEAEEEEEEEE6/EEEEEAAE/A/
/E/AEEE/EEAAE/EEEEEE/AEEAE/EEE
//AEE6EAEEA//<</<<E/EEE<<//
```

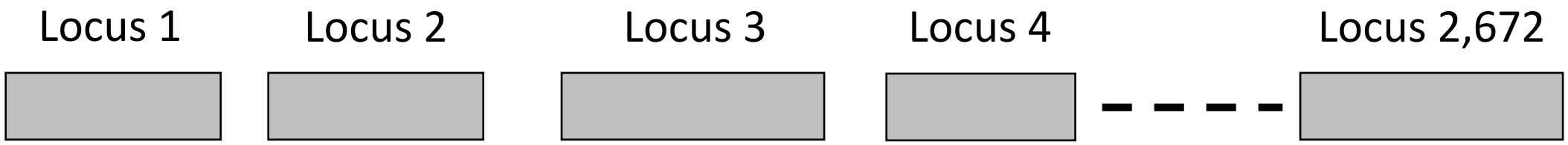
WGS data

Whole-genome SNP comparison (2012)

Detection of drug resistance (*rpoB* alerts, 2019)

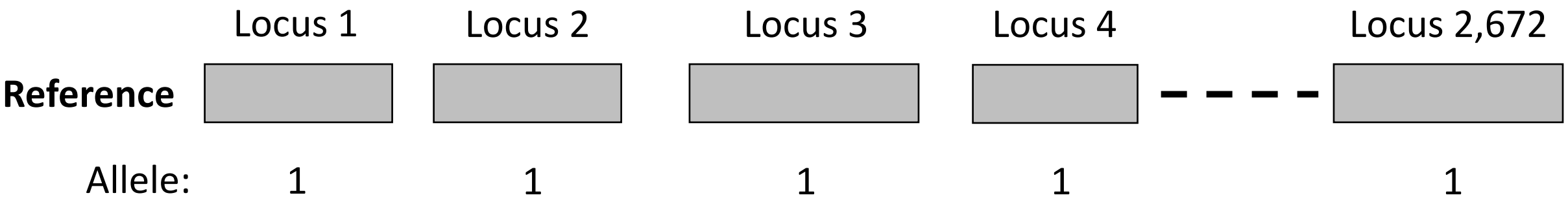
Whole-genome multilocus sequence typing (2021)

Whole-genome multilocus sequence typing (wgMLST)



- **Compares sequence at 2,672 loci throughout the genome**
  - Covers ~70% of the genome
- **Locus: location in the genome**
  - In this case, each locus is an individual gene

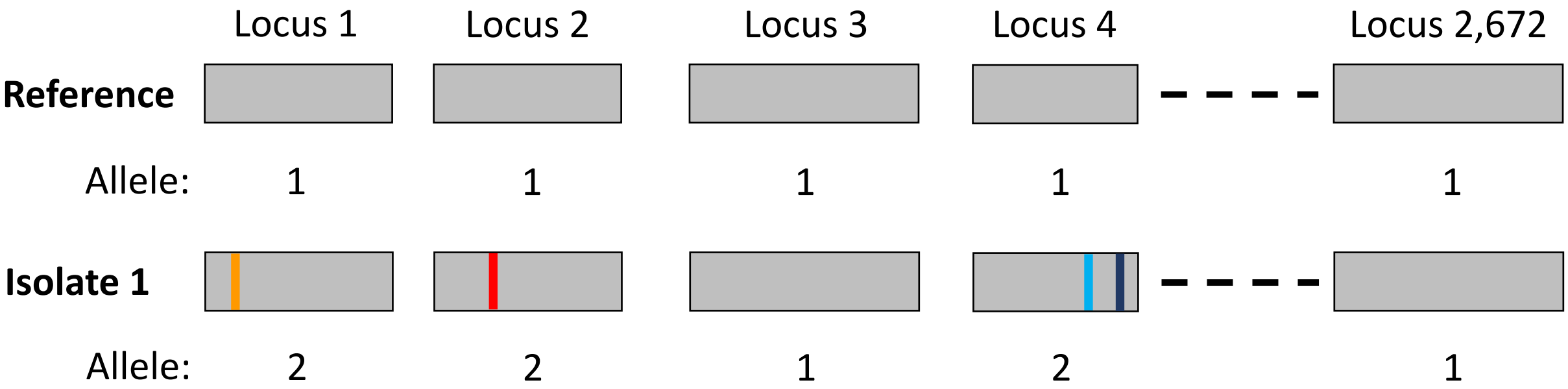
Whole-genome multilocus sequence typing (wgMLST)



Locus: location in the genome; in this case, each locus is an individual gene

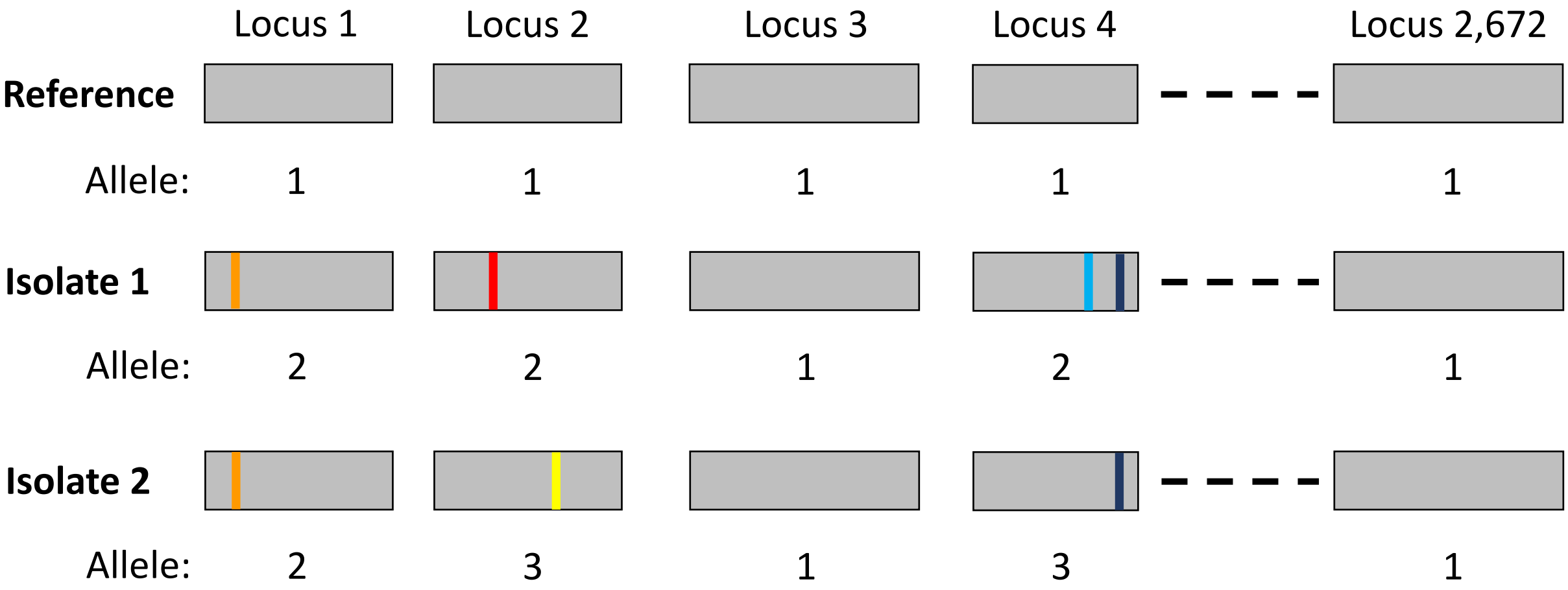
Allele: variant form of a gene

Whole-genome multilocus sequence typing (wgMLST)



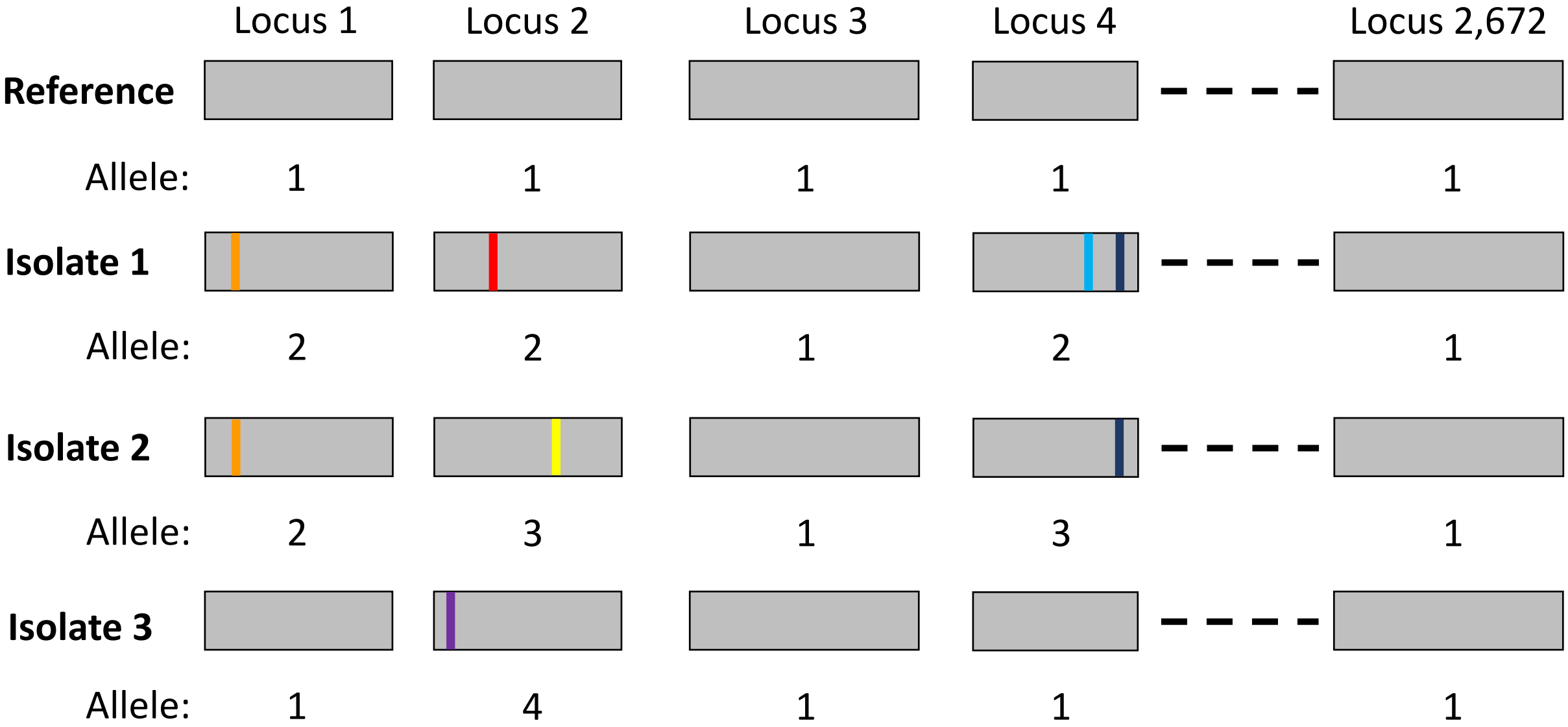
Colored lines indicate a single nucleotide difference compared to the reference

Whole-genome multilocus sequence typing (wgMLST)



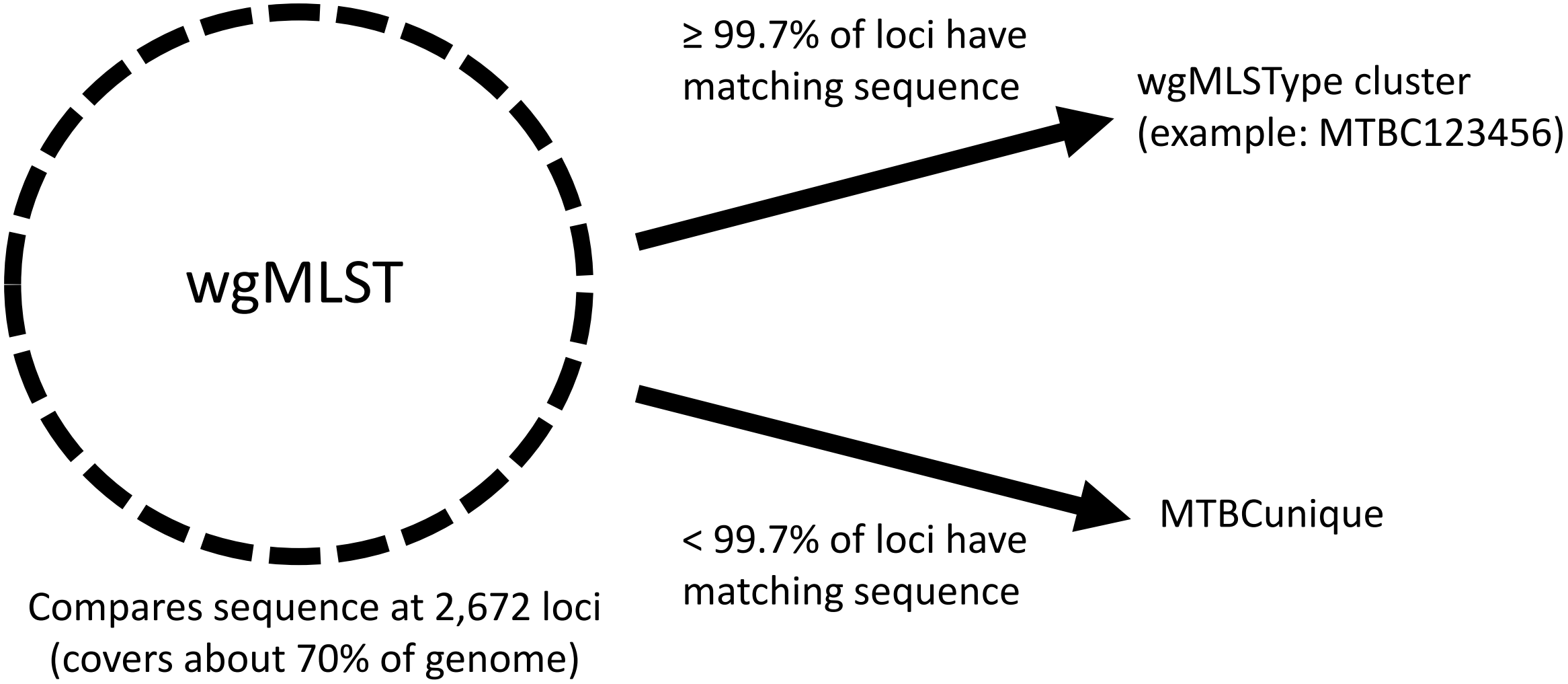
Colored lines indicate a single nucleotide difference compared to the reference

Whole-genome multilocus sequence typing (wgMLST)



Colored lines indicate a single nucleotide difference compared to the reference

Whole-genome multilocus sequence typing (wgMLST)



wgMLSType results

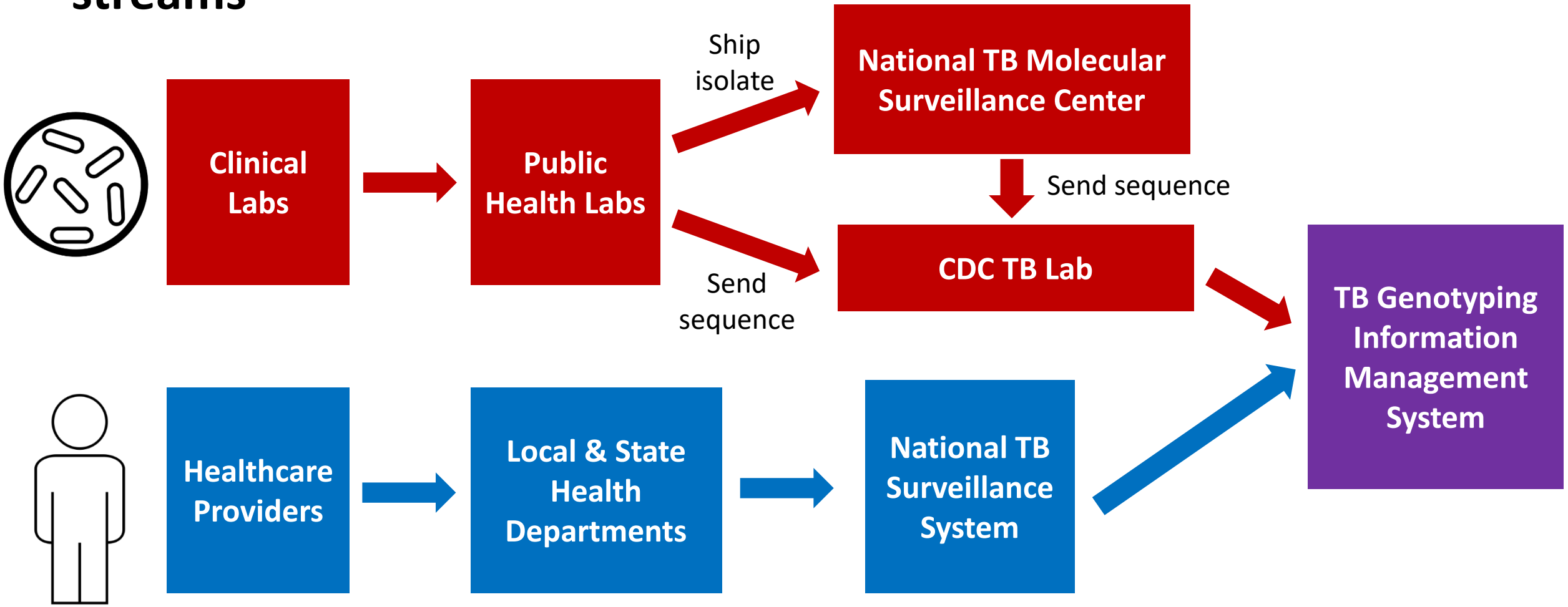
Mtb Isolate	wgMLSType
Isolate 1	MTBC099909
Isolate 2	MTBC099909
Isolate 3	MTBC099909
Isolate 4	MTBC099909
Isolate 5	MTBC099909
Isolate 6	MTBCunique
Isolate 7	MTBC000015
Isolate 8	MTBC000015

wgMLSType results

Mtb Isolate	wgMLSType	
Isolate 1	MTBC099909	Cluster 1: MTBC099909
Isolate 2	MTBC099909	
Isolate 3	MTBC099909	
Isolate 4	MTBC099909	
Isolate 5	MTBC099909	
Isolate 6	MTBCunique	> 7 SNPs from any other isolate nationally
Isolate 7	MTBC000015	Cluster 2: MTBC000015
Isolate 8	MTBC000015	

Combining TB genotyping data with TB surveillance data

TB Genotyping Information Management System (TB GIMS): combines genotyping and patient data flowing in from two streams



Search
Genotype Results
Patient Results
Records
Submitted Isolates
Submitted SRS
New Isolates
Edit Isolates
Submit Isolates
Find Duplicates
Import Data
wgMLSType Change
Reports and Tools
Watch List
Cluster Snapshot
Generate Reports
Templates
Export Data
Recent Transmission
Custom Cluster List
Alert Tracking List
LITT Analysis Tools
Build LITT Case Data File
Run LITT

Tuberculosis Genotyping Information Management System

The last TB GIMS Surveillance Upload includes data transmitted to CDC through: 02/21/2023

Searches and reports will only include data reported to CDC by the state and included in the latest TB GIMS surveillance upload.

Announcements:

No New Announcements.

State: ALL Submit

Genotyping Surveillance Coverage

Year	2020	2021	2022	2023 *
National (%)	98.0	97.8	88.9	19.7

*Year to date. NA-Not Available. Source: NTIP Data Processed Weekly

Timeliness of Genotyping - by Isolate

Time From → To	Median number of days National 2022	Median number of days National 2023 *	Goal (days)
Specimen collection → Isolate shipped to genotyping lab	54	65	NA
Receipt at genotyping lab → Genotype create date	9	12	14
Genotype create date → State Case No. entered	0	1	56
Genotype create date → Isolate Linked	2	5	90
Specimen collection → Isolate Linked	79	97	NA

TB GIMS Home - Version 2.1.5

- Search
  - Genotype Results
  - **Patient Results**
- Records
  - Submitted Isolates
  - Submitted SRS
  - New Isolates
  - Edit Isolates
  - Submit Isolates
  - Find Duplicates
  - Import Data
  - wgMLSType Change
- Reports and Tools
  - Watch List
  - Cluster Snapshot
  - Generate Reports
  - Templates
  - Export Data
  - Recent Transmission
  - Custom Cluster List
  - Alert Tracking List
- LITT Analysis Tools

Patient Results

To view surveillance data on TB patients, enter the search criteria then click **Find**

For more information, refer to **Patient Results** in the online help.

Information * Required

Basic Options

State *

County

Region

wgMLSType

GENType

PCRType

Cluster Name

State Case #

Submitter #

Accession #

Cluster Name2

Spoligotype

MIRU

MIRU2

Date Type

Start Date MM/DD/YYYY

End Date MM/DD/YYYY

Advanced Options

Find | Clear | Create Watch List Item

Search Results

Displaying 1 to 7 of 7 Records

CCID	STATE	STATE CASE #	Accession #	RPTDATE	CNTDATE	SEX	RACEHISP	AGE3	CITY	COUNTY	ZIPCODE	ORIGIN	CNTRYLN	YRSIN_US2
	ST	2023ST000000001	23RF0001	01/26/2023	01/14/2023	F	WHITE	65+	ANYTOWN	COUNTY A		USBORN		
	ST	2023ST000000002	23RF0002	01/11/2023	01/14/2023	M	HISP	65+	ANYTOWN	COUNTY A		USBORN		
	ST	2023ST000000003	23RF0003	01/10/2023	01/14/2023	M	WHITE	45-64	TOWNSVILLE	COUNTY A		USBORN		
	ST	2023ST000000004	23RF0004	01/10/2023	02/04/2023	F	BLACK	65+	METRO HEIGHTS	COUNTY A		USBORN		
	ST	2023ST000000005	23RF0005	01/09/2023	01/14/2023	F	BLACK	25-44	ANYTOWN	COUNTY A		USBORN		
	ST	2023ST000000006	23RF0006	01/03/2023	01/07/2023	M	ASIAN	45-64	TOWNSVILLE	COUNTY A		USBORN		
	ST	2023ST000000007	23RF0007	12/20/2022	01/14/2023	F	MULT	45-64	ANYTOWN	COUNTY A		USBORN		

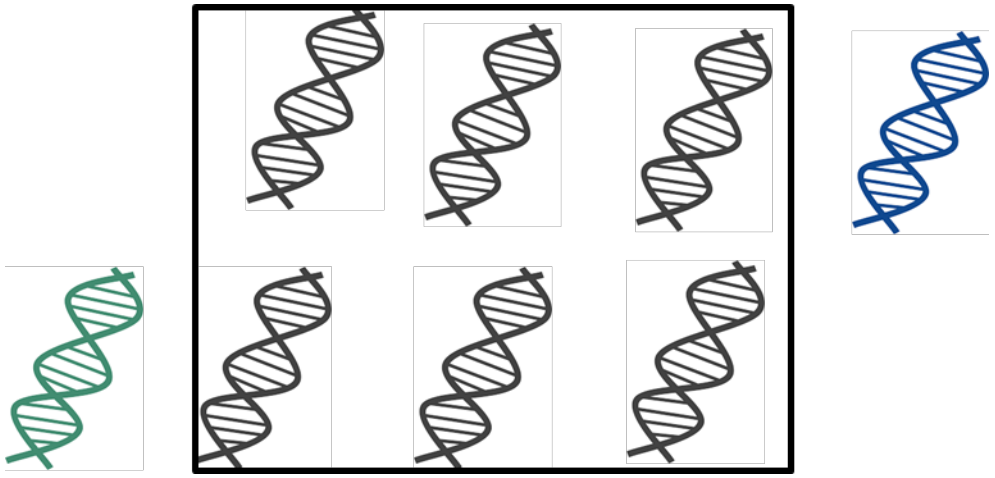
Some of the functions in TB GIMS

- Create isolate records before shipping to genotyping lab
- Search and view genotyping and patient results
- Link isolate to patient surveillance record
- Run reports and summaries of clusters
- Create watch lists to be notified of new cases in a cluster
- Run and track molecular surveillance algorithms
  - Weekly cluster alerts
  - Large outbreak surveillance
  - Yearly estimates of recent transmission
- Request additional WGS analysis (whole-genome SNP comparison)

Whole-genome SNP comparison

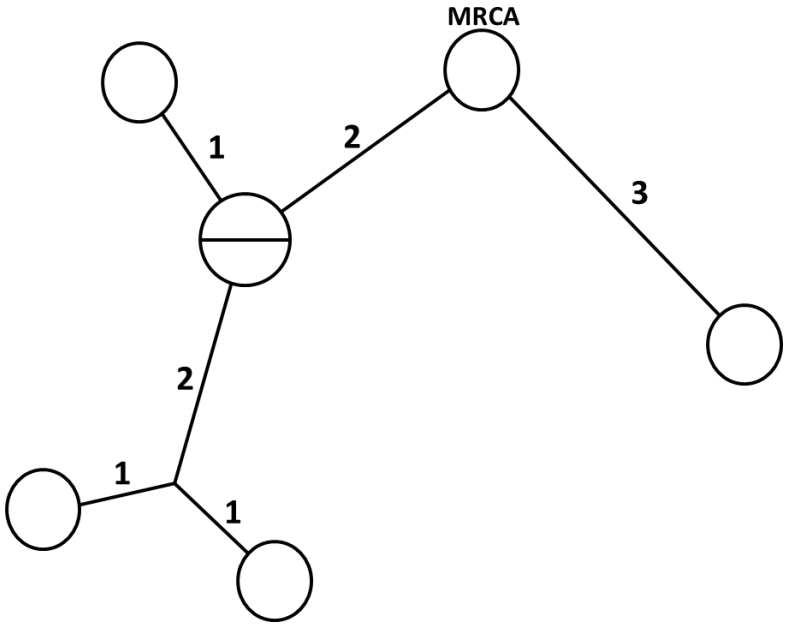
Isolates in a genotype cluster can be further differentiated by whole-genome SNP (wgSNP) comparison

Step 1. Detect a genotype cluster



Whole-genome multilocus sequence typing

Step 2. Examine genetic relationship among isolates in the cluster



Whole-genome single nucleotide polymorphism comparison

Example: wgMLSType versus wgSNP comparison

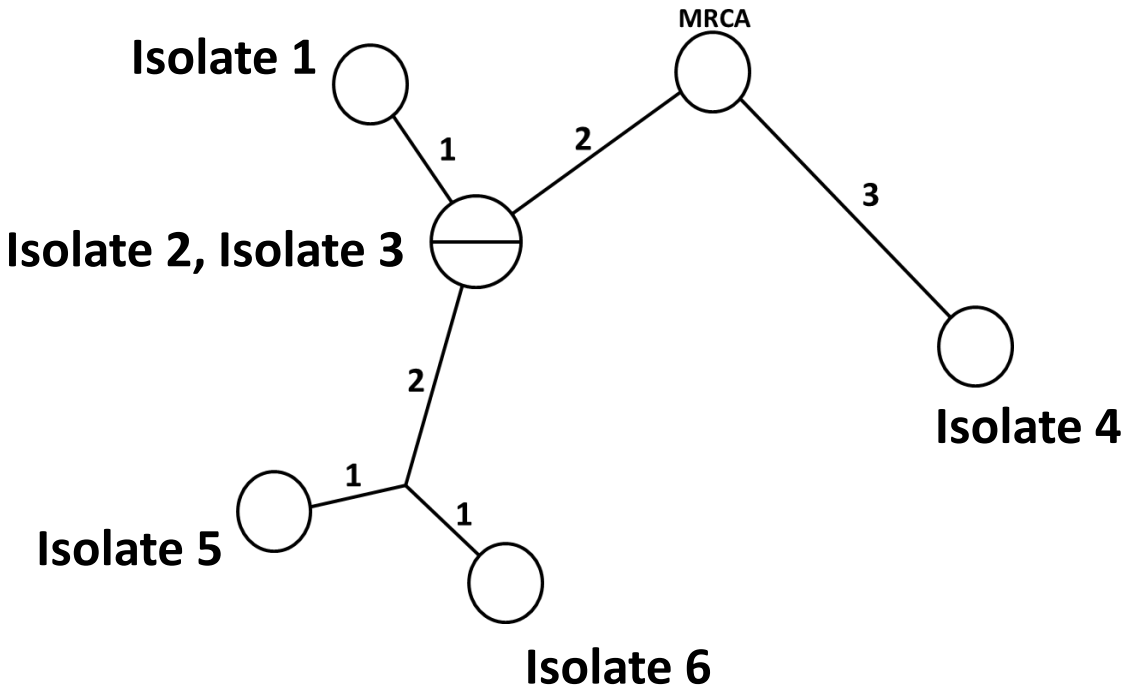
Step 1. wgMLSType

Cluster of six isolates with the same genotype MTBC001089 is detected

Isolate 1	MTBC001089
Isolate 2	MTBC001089
Isolate 3	MTBC001089
Isolate 4	MTBC001089
Isolate 5	MTBC001089
Isolate 6	MTBC001089

Step 2. wgSNP

Six MTBC001089 isolates are further differentiated and genetic relationships are examined

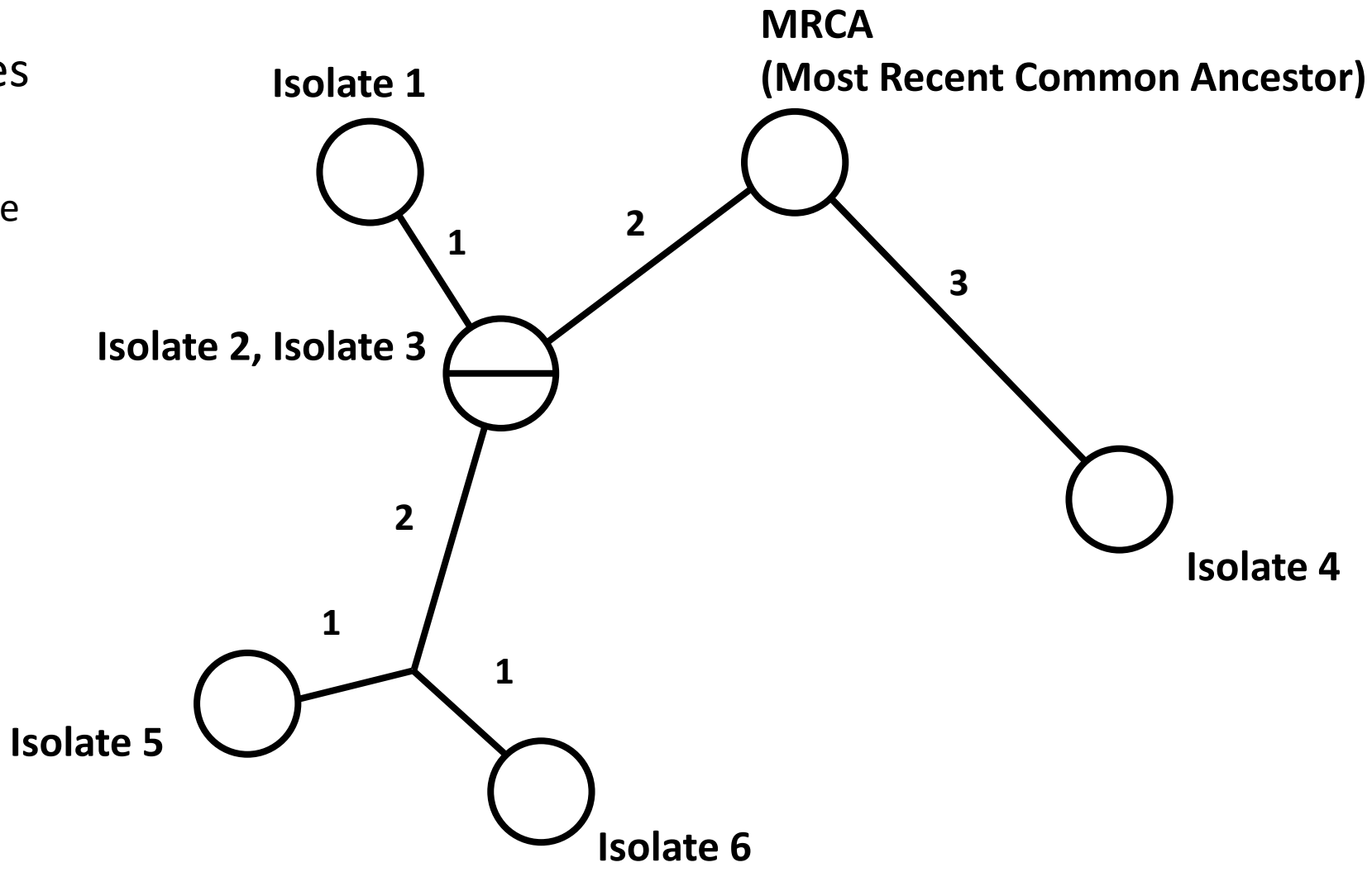


Whole-genome SNP comparison

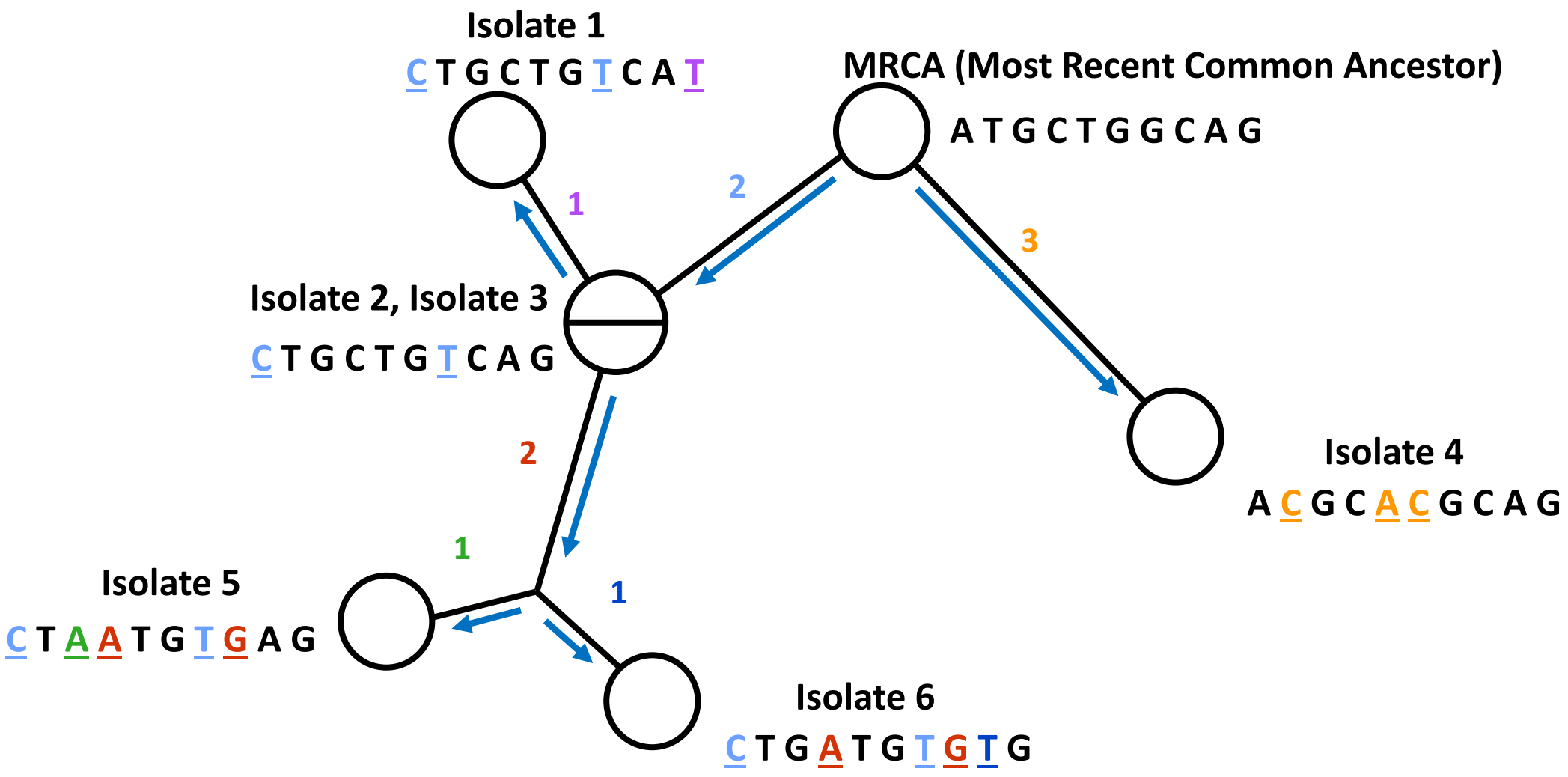
Reference	A	T	G	C	T	G	G	C	A	G
Isolate 1	C	T	G	C	T	G	T	C	A	T
Isolate 2	C	T	G	C	T	G	T	C	A	G
Isolate 3	C	T	G	C	T	G	T	C	A	G
Isolate 4	A	C	G	C	A	C	G	C	A	G
Isolate 5	C	T	A	A	T	G	T	G	A	G
Isolate 6	C	T	G	A	T	G	T	G	T	G

Whole-genome SNP comparison

- Isolates are shown as circles (called nodes)
  - Isolates that differ by 0 SNPs are displayed together in one node
- Lines are proportional in length to the number of SNPs that differ between the isolates
- Lines are labeled with the number of SNPs

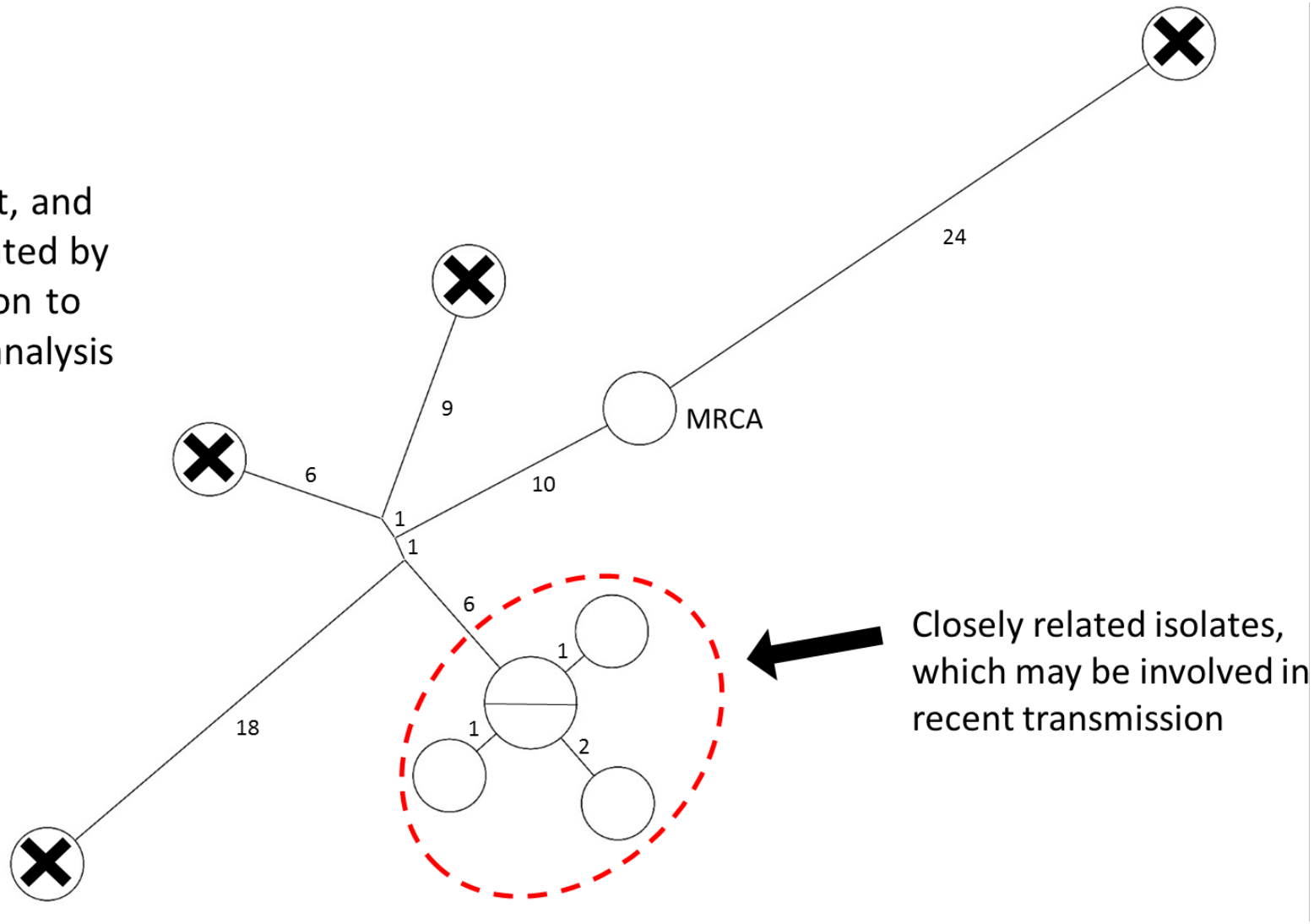


Whole-genome SNP comparison



Phylogenetic trees can be used to inform epidemiologic investigations

✕ = genetically distant, and unlikely to be related by recent transmission to other isolates in analysis



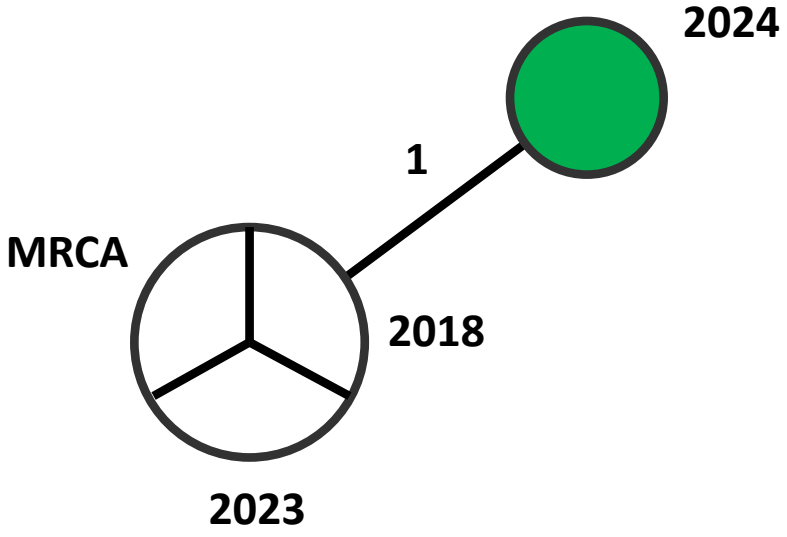
Closely related isolates, which may be involved in recent transmission

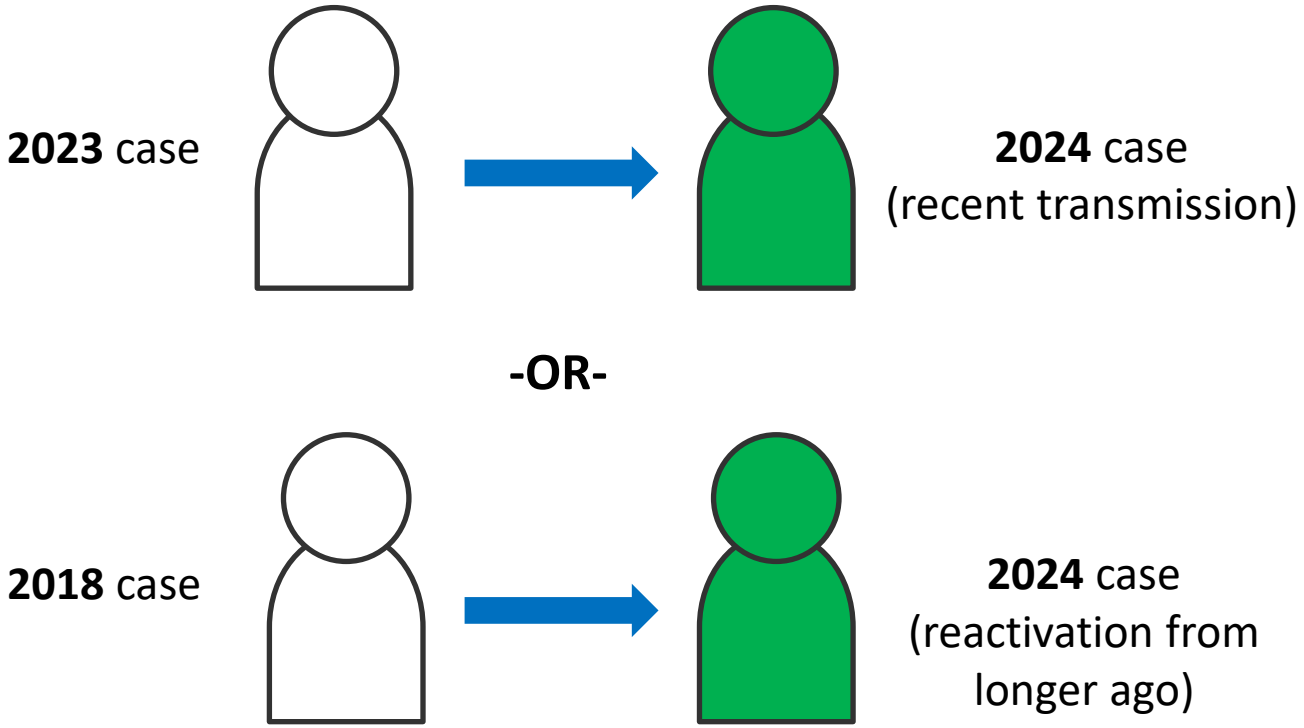
Limitations of whole-genome SNP comparison for understanding TB transmission

- Recent transmission is easier to rule out than to confirm with WGS
  - Even isolates that are closely related or identical by wgSNP can be due to reactivation
  - This is because mutations may not occur as frequently during latent infection and therefore SNPs may not accumulate

Limitations of whole-genome SNP comparison for understanding TB transmission

- Recent transmission is easier to rule out than to confirm with WGS





Limitations of whole-genome SNP comparison for understanding TB transmission

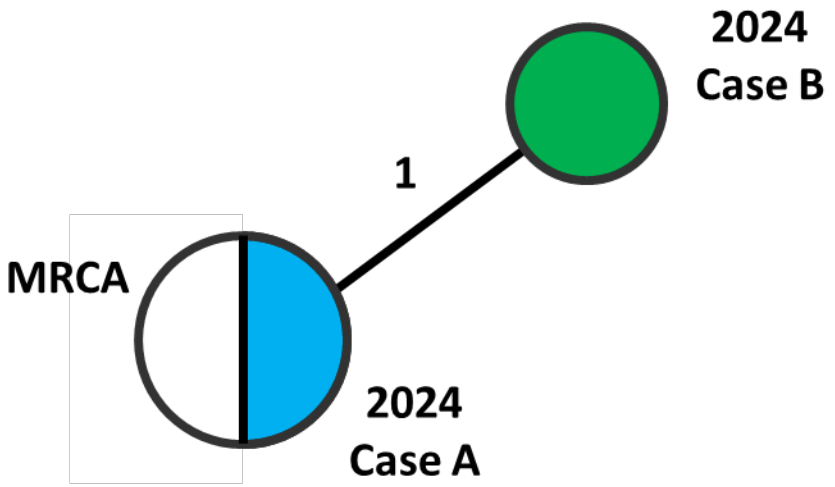
- A phylogenetic tree shows how isolates are genetically related to each other, but is not the same as a transmission diagram

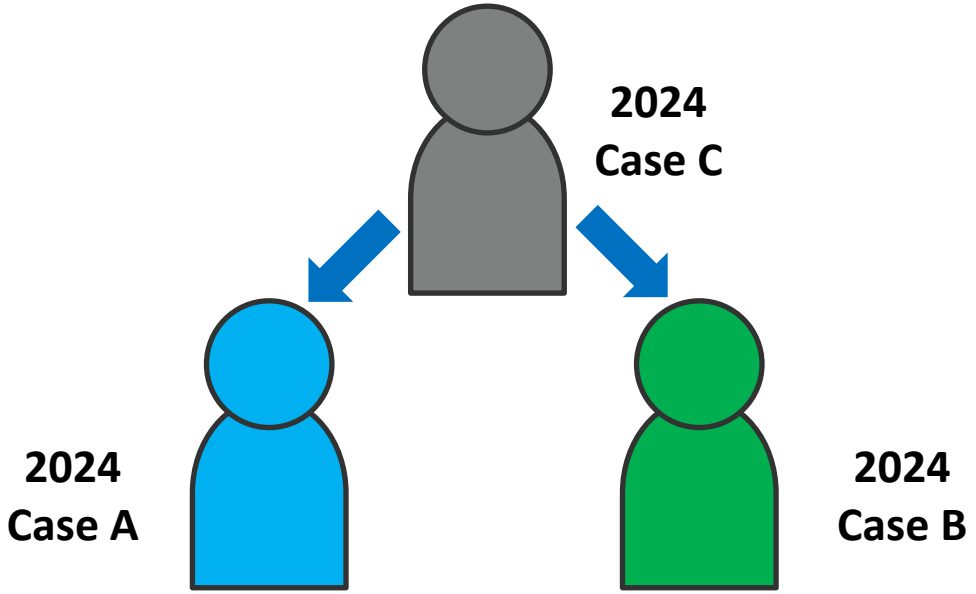
Limitations of whole-genome SNP comparison for understanding TB transmission

- A phylogenetic tree shows how isolates are genetically related to each other, but is not the same as a transmission diagram
  - Isolates might be missing from the analysis
    - Examples: Cases that are not yet diagnosed, not culture-confirmed, contaminated isolates, out-of-country cases

Limitations of whole-genome SNP comparison for understanding TB transmission

- A phylogenetic tree shows how isolates are genetically related to each other, but is not the same as a transmission diagram



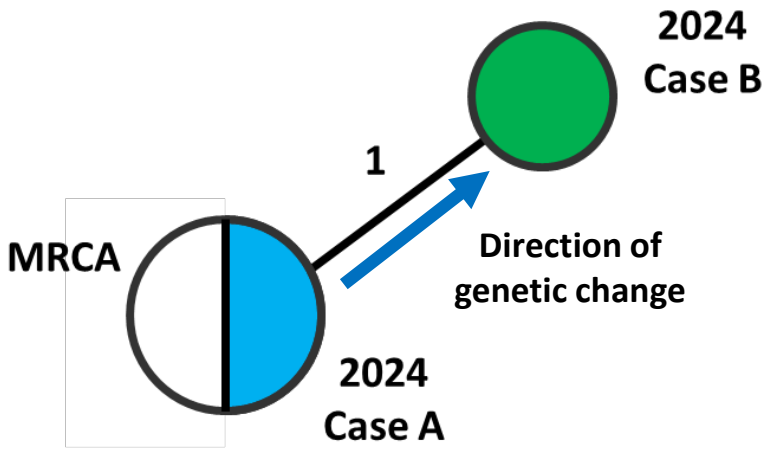


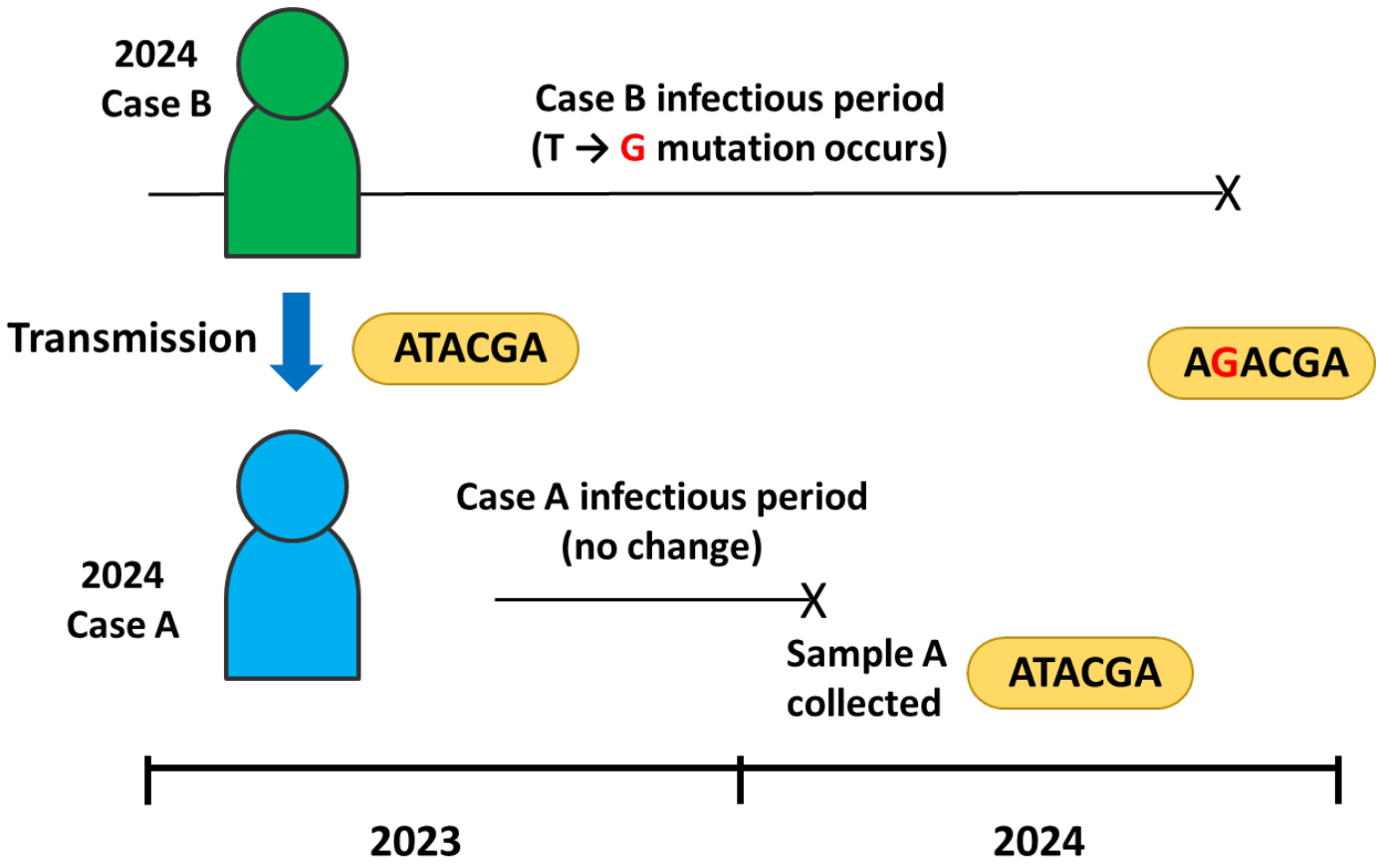
Limitations of whole-genome SNP comparison for understanding TB transmission

- A phylogenetic tree shows how isolates are genetically related to each other, but is not the same as a transmission diagram
  - Might be differences in sequence between bacterial population transmitted and bacterial population in sample that gets sequenced
    - Direction of genetic change depicted on tree might not be the same as direction of transmission

Limitations of whole-genome SNP comparison for understanding TB transmission

- A phylogenetic tree shows how isolates are genetically related to each other, but is not the same as a transmission diagram





Limitations of whole-genome SNP comparison for understanding TB transmission

- WGS alone should not be used to infer direction of TB transmission
- The phylogenetic tree should be used in conjunction with clinical and epidemiologic information to assess recent transmission and infer direction of TB transmission

WHOLE-GENOME SNP COMPARISON CASE STUDIES

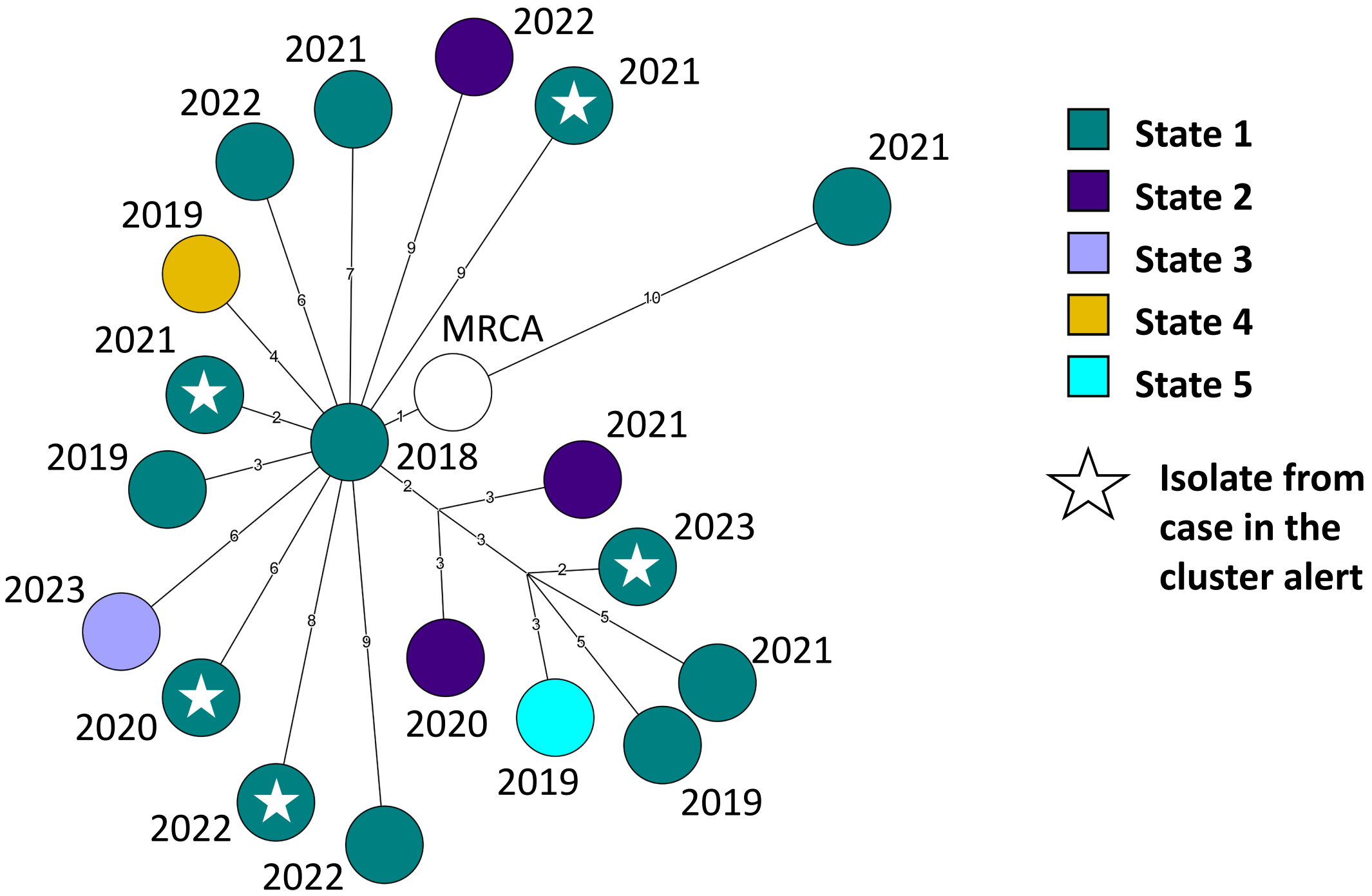
A Couple Case Studies

1. Refuting recent transmission among cases in a genotype-matched cluster
2. Separating cases into subclusters of possible recent transmission

Case Study 1: Refuting recent transmission among cases in a genotype matched cluster

Background

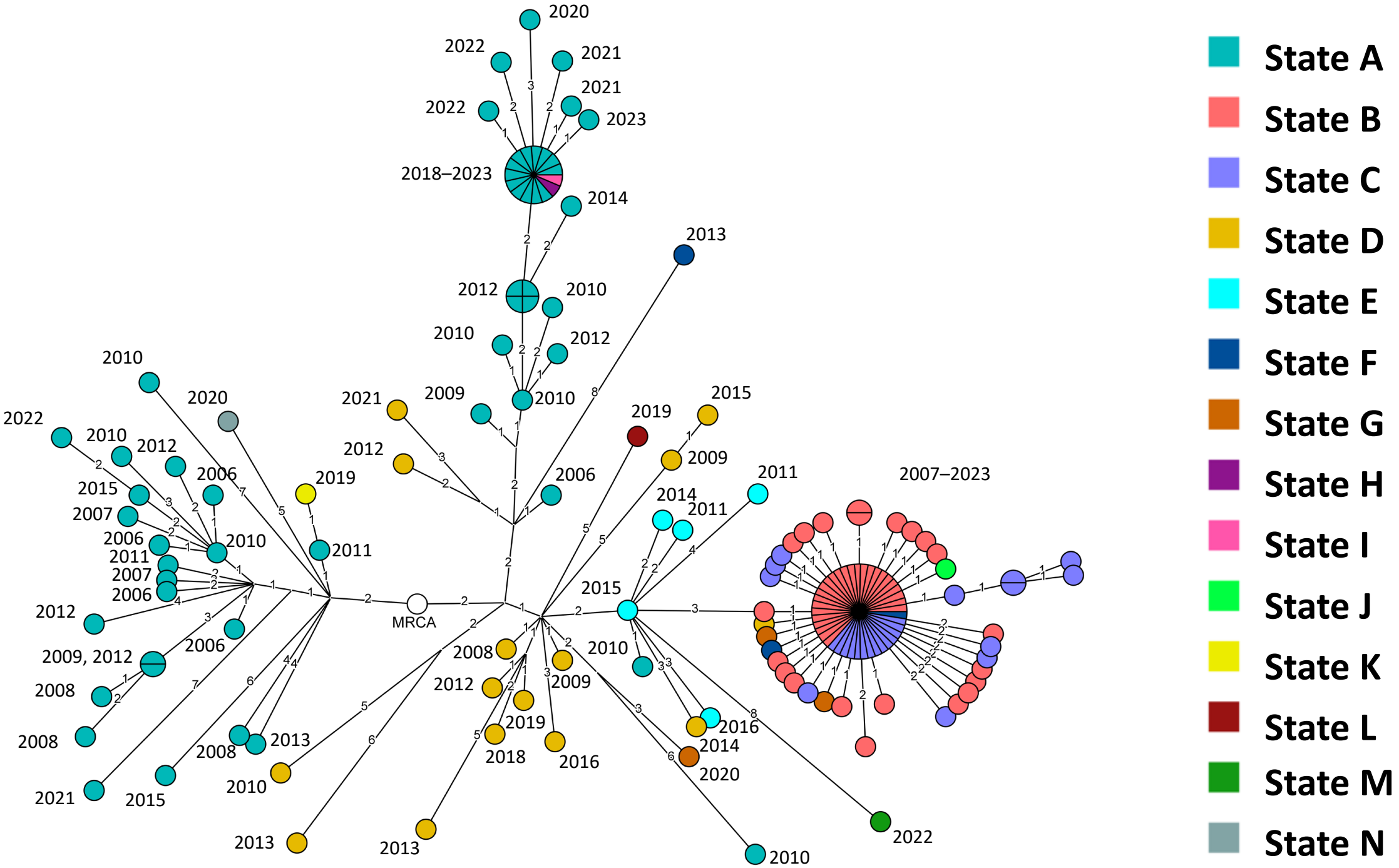
- **County cluster alert with 5 genotype matched cases**
- **Only 19 cases nationally with this wgMLSType since 2018**
  - 13 of these in the state with the cluster alert
- **All 19 cases in non-U.S.–born patients originating from the same country**

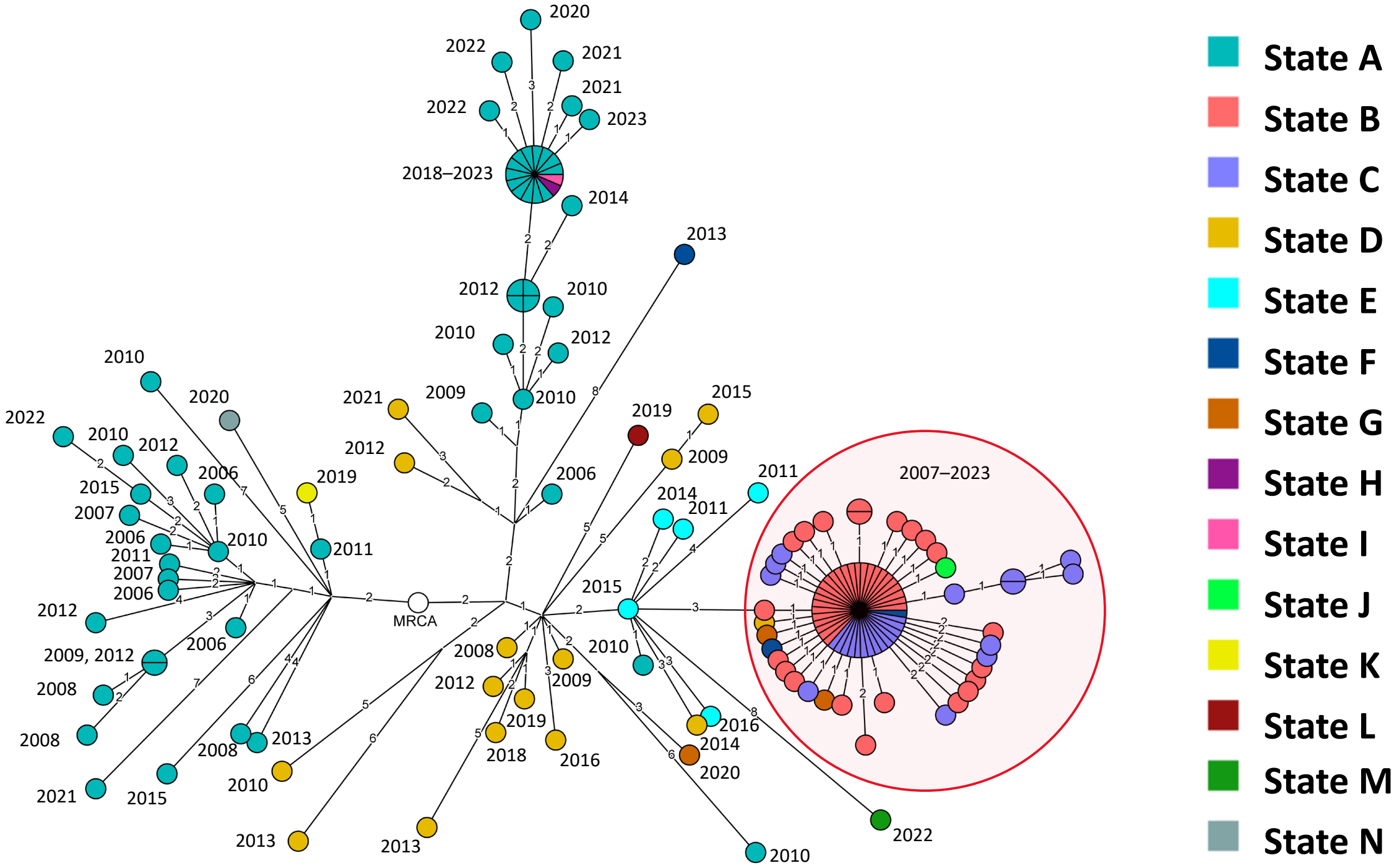

2022
2021
2021
2022
2021
2019
MRCA
2021
2018
2021
2019
2023
2023
2021
2020
2020
2019
2019
2022
2022
State 1
State 2
State 3
State 4
State 5
Isolate from case in the cluster alert

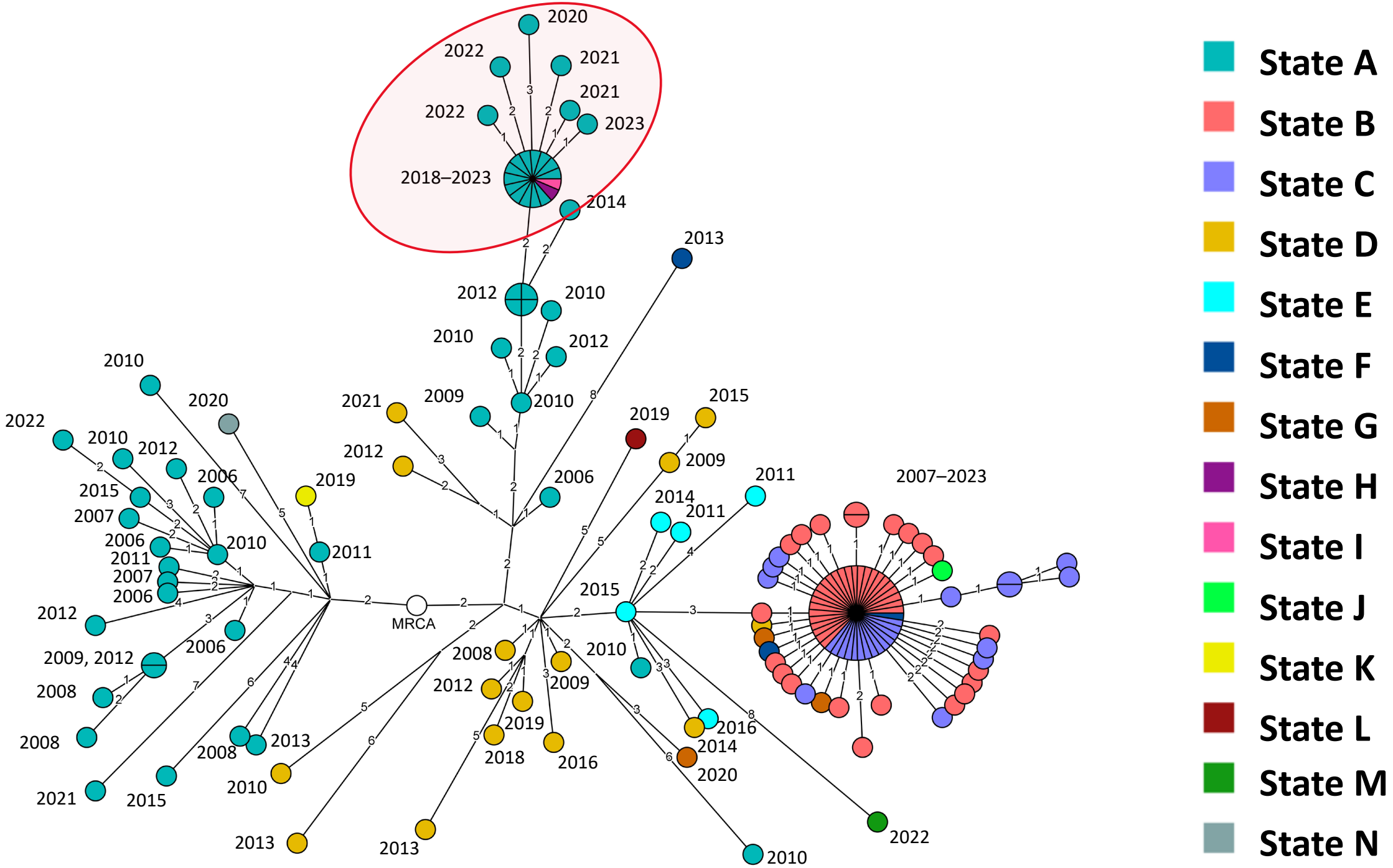
Case Study 2: Separating cases into subclusters of possible recent transmission

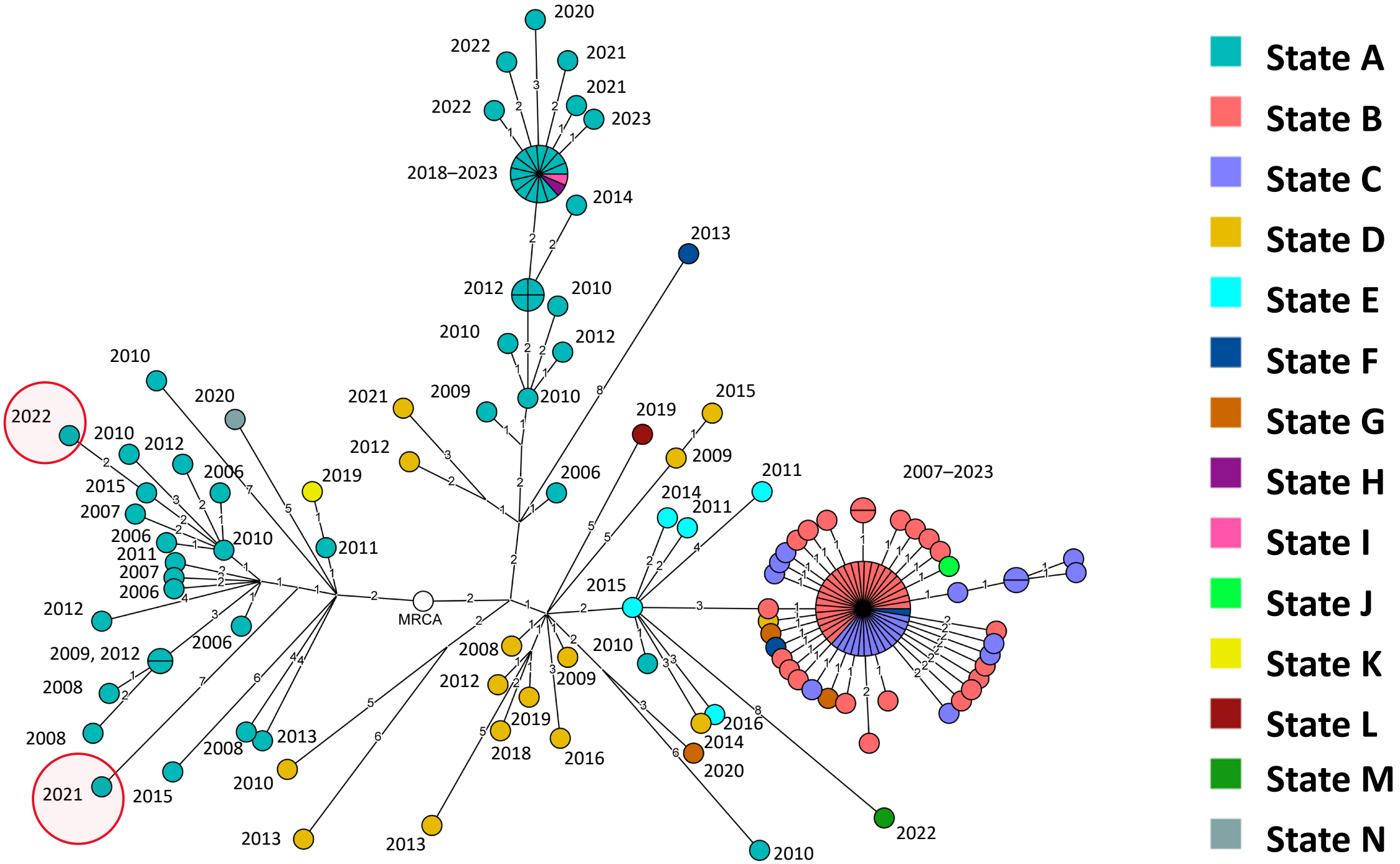
Background

- **Common genotype that has generated numerous county cluster and large outbreak alerts going back to 2012**
  - 15 county cluster alerts (4 counties in 3 states)
  - 2 separate large outbreak alerts
- **46 cases with this wgMLSType since 2018**
  - Additional 123 cases in this cluster prior to 2018
  - 18 states


2018–2023
2007–2023
MRCA
State A
State B
State C
State D
State E
State F
State G
State H
State I
State J
State K
State L
State M
State N


2018–2023
2007–2023
MRCA
State A
State B
State C
State D
State E
State F
State G
State H
State I
State J
State K
State L
State M
State N


2018–2023
2007–2023
MRCA
State A
State B
State C
State D
State E
State F
State G
State H
State I
State J
State K
State L
State M
State N


2020
2022
2021
2021
2022
2023
2018–2023
2014
2013
2012
2010
2010
2012
2010
2010
2020
2022
2010
2012
2006
2015
2007
2010
2006
2011
2007
2006
2012
2009, 2012
2006
2008
2008
2021
2015
2008
2013
2010
2013
2021
2012
2019
2011
2009
2006
MRCA
2019
2015
2009
2011
2014
2011
2015
2010
2008
2012
2009
2019
2018
2016
2013
2016
2014
2020
2010
2022
2007–2023
State A
State B
State C
State D
State E
State F
State G
State H
State I
State J
State K
State L
State M
State N

Summary

- We use the genotyping data to identify clusters of TB cases that might represent recent transmission or an outbreak
  - Opportunity for public health action to interrupt further transmission
- *M. tuberculosis* isolates are first assigned a wgMLSType that is used for cluster detection and in molecular surveillance algorithms
- Whole-genome SNP comparison is then used to further assess which cases in a genotype-matched cluster might be related by recent transmission
  - Can be used to inform cluster and outbreak investigation

QUESTIONS?

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 cdc.gov
Follow us on X (Twitter) @CDCgov & @CDCEnvironment