Whole-genome sequencing for investigation of recent TB transmission in the United States: Current uses and future plans

Sarah Talarico, PhD and Benjamin Silk, PhD
Surveillance, Epidemiology, and Outbreak Investigations Branch

Tambi Shaw, MPH and Martin Cilnis, MPH, MS
TB Control Branch/California Department of Public Health

Hello, I'm Sarah Talarico. I'm an epidemiologist with the Molecular Epidemiology Activity in the Division of Tuberculosis Elimination at CDC

I will be presenting the following set of training slides focused on the use of whole-genome sequencing for investigating recent TB transmission along with colleagues, Tambi Shaw and Martin Cilnis, from the TB Control Branch at the California Department of Public Health

Learning objectives

At the end of this presentation, participants will be able to describe

- Key differences between conventional genotyping and WGS
- What is being represented on a phylogenetic tree
- How WGS is used to assess whether patients are potentially linked by recent transmission
- Why WGS alone cannot be used to infer direction of transmission
- How TB control programs can use WGS analysis in an investigation

At the end of this presentation, participants will be able to describe:
Key differences between conventional genotyping and WGS
What is being represented on a phylogenetic tree
How WGS is used to assess whether patients are potentially linked by recent transmission
Why WGS alone cannot be used to infer direction of transmission
How TB control programs can use WGS analysis in an investigation

Outline

- **Part 1: Introduction to using whole-genome sequencing (WGS) for detection and investigation of recent TB transmission**
  - Goals of TB molecular epidemiology
  - Current genotyping methods
  - Use of WGS for investigating TB transmission
  - Guide for interpreting results of WGS analysis
- **Part 2: Case studies using WGS to investigate TB cluster alerts in California**
  - 2 case studies with WGS and epidemiologic data
  - One confirmed outbreak and one refuted outbreak
- **Part 3: Plans for transition to universal prospective WGS**

This presentation is divided into three parts

Part 1 is an introduction to using whole-genome sequencing (or WGS) for detection and investigation of recent TB transmission

I will cover the goals of TB molecular epidemiology, current genotyping methods, use of WGS for investigating TB transmission, and a guide for interpreting results of WGS analysis

Part 2 will be case studies using WGS to investigate TB cluster alerts in California

Martin and Tambi will present two case studies with WGS and epi data, one is a confirmed outbreak and one is a refuted outbreak

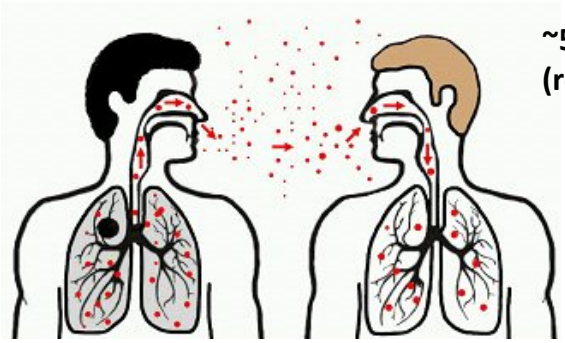
Then in Part 3, I will briefly describe the plans for transition to universal prospective WGS. A separate presentation covering the details of how universal prospective WGS will be implemented will be made available in the future

Part 1

Using WGS for detection and investigation of recent TB transmission

First I will go over some background for how to use WGS for detection and investigation of recent TB transmission

TB Transmission and Course of Infection



~5% within 2 yrs.
(recent transmission)

Active Disease
(infectious)

~5% lifetime risk
(reactivation)

~95%

Latent Infection
(not infectious)

TB is caused by Mycobacterium tuberculosis and is transmitted through the airborne route

5% of people who are infected develop active TB disease within 2 years. These people are infectious and can carry on the chain of transmission

The other 95% develop latent TB, which is not infectious

However, about 5% of people with latent TB will reactivate and develop active disease at some point in their life

So two main strategies for eliminating TB are detecting and treating latent TB infection and detecting and interrupting ongoing transmission, which is the focus for today's presentation

TB Molecular Epidemiology: Targeting Recent Transmission

- 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
  - Combine molecular, clinical, and epidemiologic data to detect, investigate, and monitor recent TB transmission

TB molecular epidemiology targets recent transmission with the goal of reducing the burden of TB by identifying where transmission is currently occurring and interrupting it

The challenge is we need to distinguish TB cases that are due to recent transmission from cases that were infected long ago and are just now developing active disease

We do this by combining molecular, clinical, and epidemiologic data to detect, investigate, and monitor recent transmission

Why combine molecular, clinical, and epidemiologic data to understand TB transmission?

- 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
- Molecular genotyping data can provide additional, complementary information to aid detection and investigation of transmission
  - Genotyping identifies cases with genetically similar *M. tuberculosis* isolates that are more likely to be linked by transmission

We combine molecular data with clinical and epi data because there are challenges when it comes to trying to rely exclusively on epi data to investigate TB transmission

The fact that transmission is airborne can make it difficult to assess exposure

Assessing exposure in congregate settings can be very complex as well

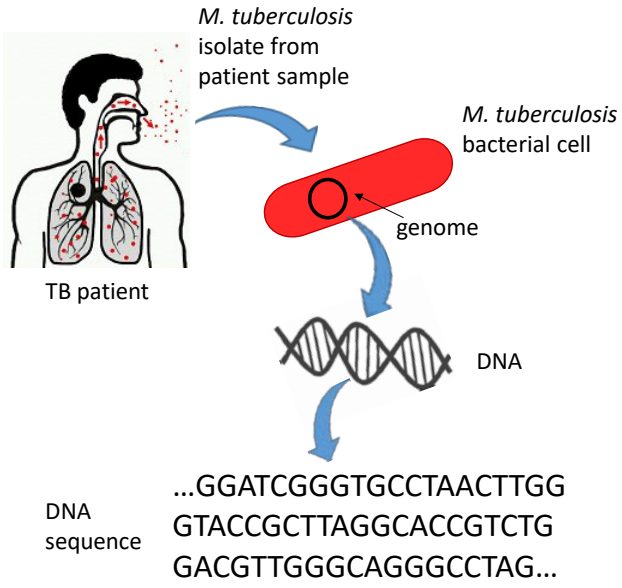
TB also can have infectious periods that span years and for that reason patient recall may be incomplete or unreliable

TB transmission often occurs in impoverished or marginalized communities who are difficult to access

For these reasons, it is helpful to use genotyping which can provide additional, complementary information to aid detection and investigation of transmission by identifying cases with genetically similar M. tuberculosis isolates that are more likely to be linked by transmission

Genotyping examines the DNA of *M. tuberculosis* isolates from TB patients

- The *M. tuberculosis* bacteria 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



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

The M. tuberculosis bacteria from a TB patient is called the patient's isolate

Bacteria, including M. tuberculosis, have DNA called a bacterial genome

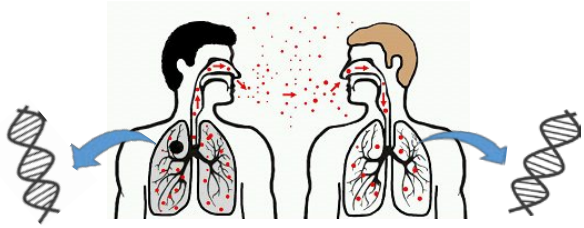
DNA is made up of four different nucleotides (abbreviated A, T, C and G)

The order of these nucleotides in the genome is the DNA sequence

The genome of M. tuberculosis is over 4.4 million nucleotides long

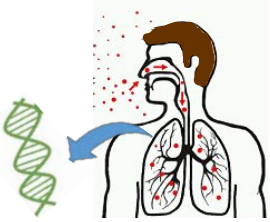
Genotyping can be used to identify TB patients who are more likely to be linked by recent transmission

- Changes in the DNA (mutations) occur over time, so *M. tuberculosis* bacteria don't all have the exact same DNA sequence
- At the time of transmission, the person transmitting the infection and the person acquiring the infection will have *M. tuberculosis* with identical 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 (black)

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



Genotyping can be used to identify TB patients who are more likely to be linked by recent transmission

Changes in the DNA, called mutations, occur over time so M. tuberculosis bacteria don't all have the exact same DNA sequence

At the time of infection, the person transmitting the infection and the person acquiring the infection will have M. tuberculosis with identical DNA sequence

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

In this schematic, transmission is occurring between these two people at the top and they have M. tuberculosis isolates with the same genotype (shown in black), but this person at the bottom is not part of that transmission chain and has an M. tuberculosis isolate with a different genotype (shown in green)

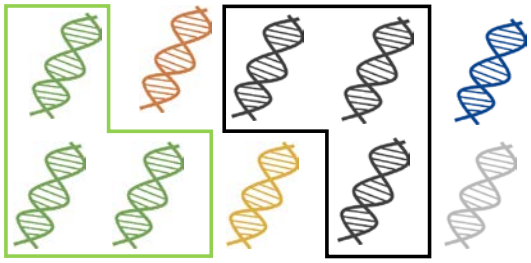
Detecting Clusters of Recent Transmission using Genotyping

- 2 or more isolates with the same genotype are clustered
- Algorithms that consider time and space are used to identify clustered cases that may be due to recent transmission

<u>CDC cluster detection methods</u>

- LLR cluster alerts: Unexpected increase in concentration of a genotype in a jurisdiction during a 3-year time period
- Large outbreak surveillance: 10 or more cases in a 3-year period related by recent transmission

<u>*M. tuberculosis* genotypes in a community</u>



CDC uses M. tuberculosis genotyping data to detect clusters of possible recent transmission

2 or more isolates with the same genotype are considered clustered

This schematic on the right is showing M. tuberculosis genotypes in a community, and we can identify a green cluster and a black cluster

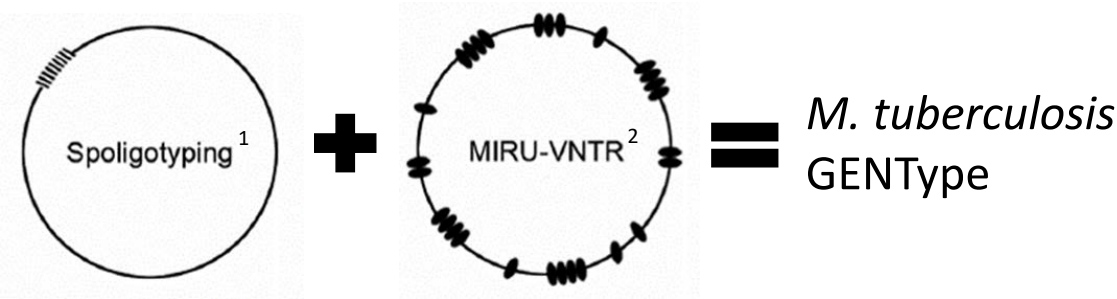
Algorithms that consider time and space are then used to identify clustered cases that may be due to recent transmission

And CDC has developed cluster detection methods for this purpose

One method is the LLR cluster alert that detects an unexpected increase in concentration of a genotype in a jurisdiction during a 3-year time period

Another type of alert is for surveillance of large outbreaks, defined as 10 or more cases in a 3-year period related by recent transmission

Current *M. tuberculosis* genotyping is based on only ~1% of the genome



Two assays to detect differences in repetitive regions of the genome

1. Spacer Oligonucleotide Typing
2. Mycobacterial Interspersed Repetitive Units-Variable Number of Tandem Repeats

Adapted from: Guthrie JL, Gardy JL. Ann N Y Acad Sci. 2016 Dec 23. doi: 10.1111/nyas.13273

Current M. tuberculosis genotyping methods cover only about 1% of the genome and are based on differences in repetitive regions within the M. tuberculosis genome

The current method combines the results of two assays, spoligotyping and MIRU-VNTR, to give an M. tuberculosis GENType

Specifically, spoligotyping is based on the presence or absence of 43 spacer sequences in a direct repeat region of the genome

And MIRU-VNTR is based on differences in the number of copies of tandem repeats at 24 regions or loci of the genome

Isolates that have the exact same spoligotype and 24 locus MIRU-VNTR pattern are assigned the same GENType

GENTyping provides low resolution for examining genetic relatedness of isolates

- Examines only a small portion (~1%) of the genome
- Regions examined may not change within a timeframe that is useful for understanding recent transmission
- Substantial past transmission of a GENType in a community makes it harder to distinguish:
  - Cases due to reactivation of infection that was acquired during the past transmission versus cases due to recent transmission
  - Separate chains of recent transmission among cases with the same GENType

However, GENTyping provides relatively low resolution for examining the genetic relatedness of isolates because it only examines a small portion, about 1%, of the genome

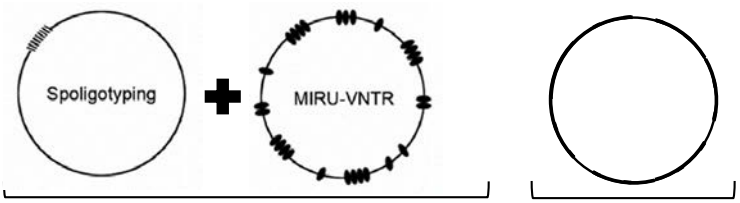
The regions of the genome examined by GENTyping may not change within a timeframe that is useful for understanding recent transmission

This makes interpretation difficult when there has been substantial past transmission of a GENType in a community because it is harder to distinguish cases due to reactivation of infection that was acquired during the past transmission versus cases due to recent transmission

And it is harder to distinguish separate chains of recent transmission among cases with the same GENType

WGS can provide added resolution for examining genetic relatedness of isolates

- **Expands coverage of the genome to ~90%**
  - Captures much more of the genetic changes that occur



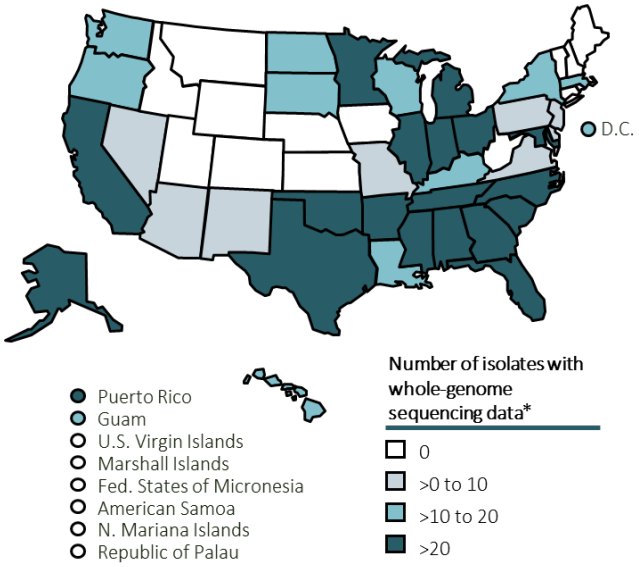
GENTyping (1% coverage) **WGS (90% coverage)**

Adapted from: Guthrie JL, Gardy JL. Ann N Y Acad Sci. 2016 Dec 23. doi: 10.1111/nyas.13273

Whole-genome sequencing (or WGS) can provide added resolution for examining genetic relatedness of isolates by expanding coverage of the genome to about 90%, compared to the 1% that is covered by GENTyping
This captures much more of the genetic changes that occur

Retrospective WGS for Select GENType clusters

- WGS and phylogenetic analysis of >100 clusters nationally to date
  - 2012: first WGS of a GENType cluster
  - 2014: WGS performed for all GENType clusters that alerted for large outbreak surveillance
  - 2016: WGS expanded to include other select GENType clusters that could inform public health action



* N = 2,776 isolates, data current as of Aug. 2017

CDC has been performing WGS and phylogenetic analysis retrospectively for select GENType clusters and has analyzed more than 100 clusters nationally to date

2012 is when we first did WGS of a GENType cluster

In 2014, we started performing WGS for all GENType clusters that alerted for large outbreak surveillance

In 2016, we expanded WGS capacity to include other select GENType clusters that could inform public health

This map is showing the number of isolates with whole-genome sequencing data for each state or territory, with a total of 2,776 isolates having been sequenced as of August 2017, but retrospective sequencing is still ongoing

Whole-genome single nucleotide polymorphism (wgSNP) analysis

- A single nucleotide polymorphism (SNP) is a mutation at a single position (A,T,C, or G) in the DNA sequence
- wgSNP analysis uses WGS data to identify SNPs that are useful for examining the genetic relationship among isolates
- SNPs that are identified in the wgSNP analysis are mapped on to a phylogenetic tree to diagram the genetic relationship among isolates
- The phylogenetic tree can be used to target and inform epidemiologic investigation of these cases

The whole-genome sequencing data is used to perform a whole-genome single nucleotide polymorphism analysis (or wgSNP analysis)

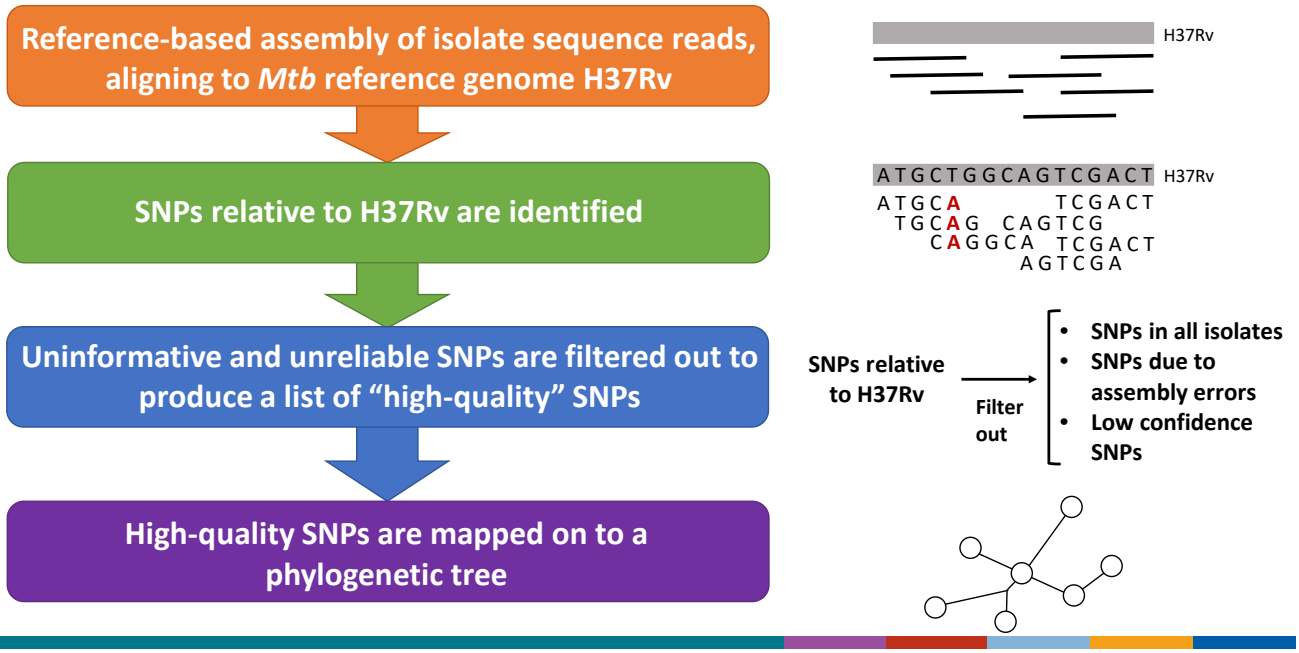
A single nucleotide polymorphism (or SNP as we call them) is a mutation at a single position (an A, T, C, or G) in the DNA sequence

wgSNP analysis uses WGS data to identify SNPs that are useful for examining the genetic relationship among isolates

SNPs that are identified in the wgSNP analysis are mapped on to a phylogenetic tree to diagram the genetic relationship among isolates

The phylogenetic tree can be used to focus and inform epidemiologic investigation of the cases

wgSNP analysis



wgSNP analysis is done by first aligning the isolate sequence reads to a reference genome, we use M. tuberculosis H37Rv

This is shown on the right here where the sequence reads for the isolate are in black and they are being matched up to the sequence of the reference genome shown in grey

Then, SNPs relative to the reference genome H37Rv are identified

This is shown here where the isolate sequence has an A at this position where H37Rv has a T

Then the next step is that uninformative and unreliable SNPs are filtered out to produce a list of high-quality SNPs

What we filter out are SNPs that are present in all the isolates in the analysis (because if the SNP is present in all the isolates in the cluster being analyzed, the SNP is not informative for understanding the genetic relationships among the isolates)

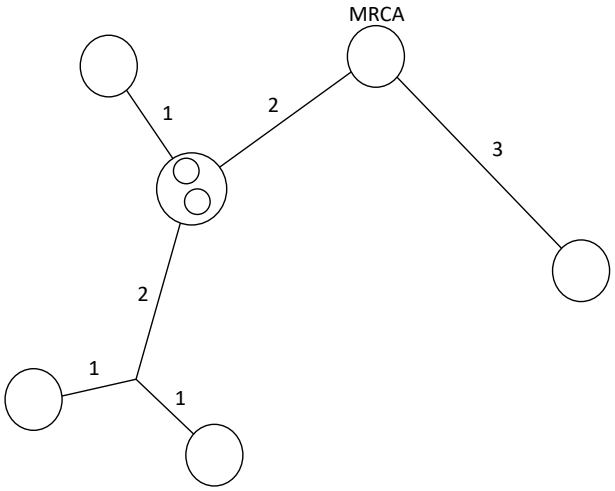
We also filter out SNPs that are identified because of assembly errors (which means the sequence read wasn't aligned to the correct place of the reference genome)

And low confidence SNPs (for example if there are very few sequence reads that cover the SNP position)

Then lastly the high-quality SNPs are mapped on to a phylogenetic tree to diagram the genetic relationships between the isolates

Guide for interpreting the phylogenetic tree

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



Here is a guide for interpreting the phylogenetic tree

On this tree, the isolates are shown as circles, called nodes (these would usually be labeled with the isolate's accession number)

Isolates that have the same genome type are displayed together in one node

Nodes are connected by lines that are proportional in length to the number of SNPs that differ between the isolates

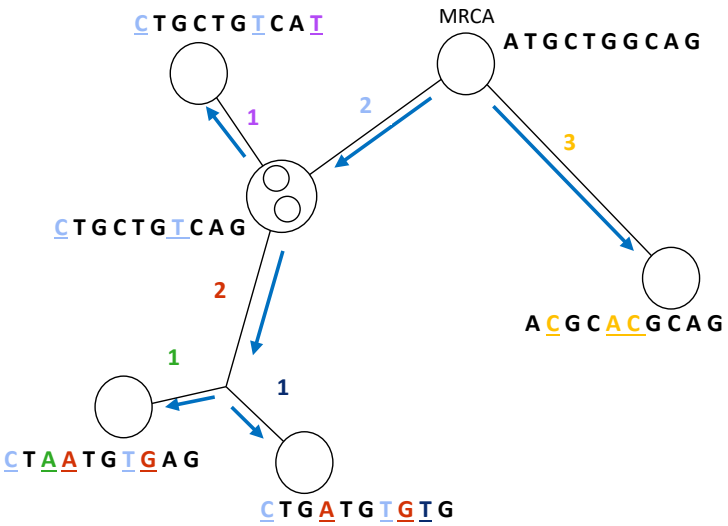
And the lines are labeled with the number of SNPs

You will see that the tree also has a node labeled MRCA

Guide for interpreting the phylogenetic tree

MRCA = Most Recent Common Ancestor

- Hypothetical genome type (not an actual isolate)
- All isolates on the tree are descended from this hypothetical genome type
- Serves as a reference point for examining the direction of genetic change (⟶)



MRCA stands for most recent common ancestor

The MRCA is not an actual isolate but a hypothetical genome type from which all isolates on the tree are descended

And so the MRCA serves as a reference point for examining the direction of genetic change which is shown with these blue arrows

On this tree, if we start up here at the MRCA and move down this branch, there are three SNPs shown in yellow and this isolate has those three SNPs

But if we were to move down this other branch, these two isolates in this node don't have those three yellow SNPs but they have two different SNPs shown in blue

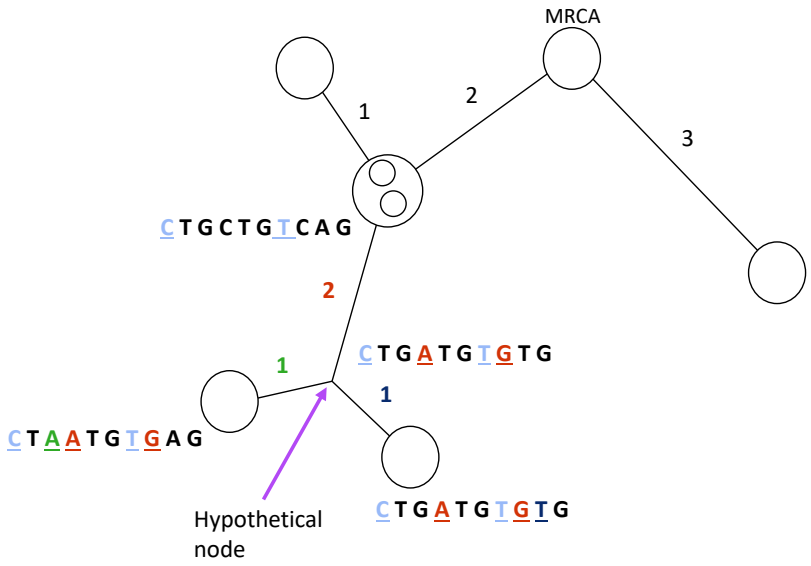
And then from there, we can move up this branch, and this isolate has those same two blue SNPs plus one more SNP shown in purple

If we move down this way from the node with the two isolates, then these isolates down here have the two blue SNPs plus two red SNPs and then each one has one additional SNP as well – this one has a green SNP and this one has a dark blue SNP

Guide for interpreting the phylogenetic tree

Hypothetical Node

- Branching point with no circle
- Represents a hypothetical genome type
- No actual isolate with this genome type in the analysis



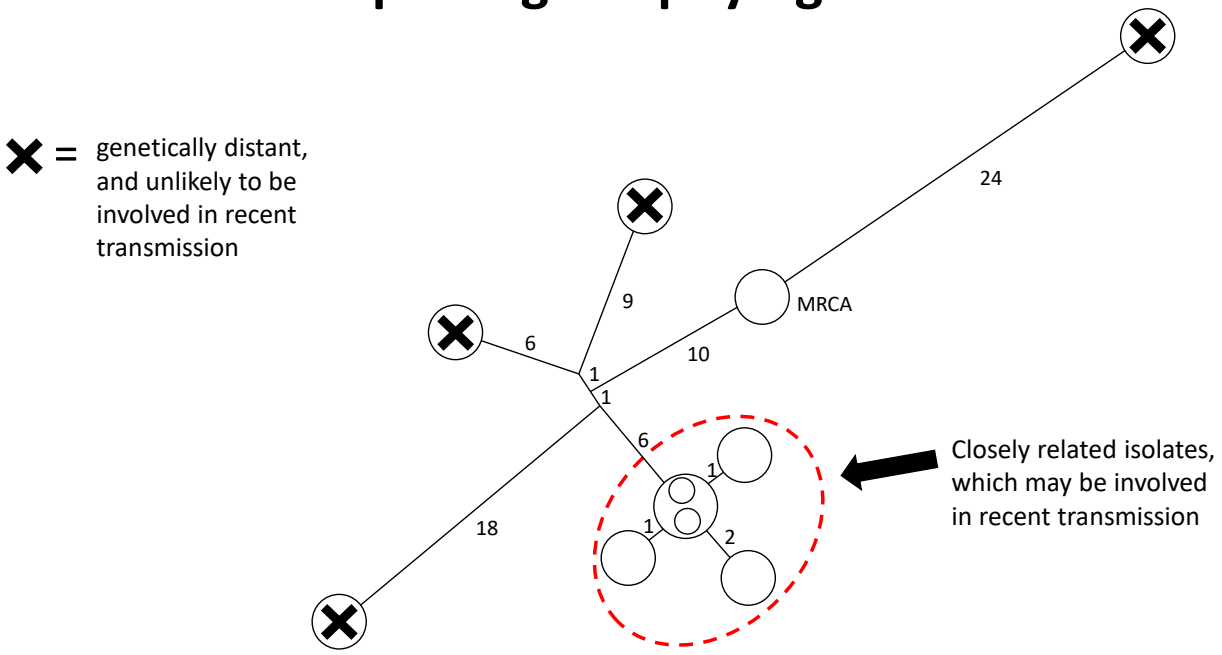
And you will see at the bottom here that trees also sometimes have a branching point with no circle

This is called a hypothetical node

The hypothetical node represents a hypothetical genome type but there is no actual isolate with this genome type in the analysis

So in this example, if we start here at this node with the two blue SNPs and move down this branch to the hypothetical node, you see that the genome type at the hypothetical node has the two blue SNPs and the two red SNPs but we don't actually have an isolate with that genome type in the analysis

Guide for interpreting the phylogenetic tree



We use the tree to examine the number of SNPs that differ between the isolates and identify groups of closely related isolates that may be involved in recent transmission

As well as to identify genetically distant isolates that are unlikely to be involved in recent transmission

This helps programs prioritize cases for focus of their epi investigation

Guide for interpreting the phylogenetic tree

- SNP thresholds for categorizing *M. tuberculosis* isolates as genetically distant or closely related have not been formally established for CDC's wgSNP analysis yet
- Based on CDC's general experiences using wgSNP analysis for investigating recent transmission:
  - Isolates with 0 – 5 SNP differences are considered closely related
  - Isolates with 6 or more SNP differences are considered genetically distant
- SNP thresholds will vary depending on the methods used for the wgSNP analysis, and cannot be compared to thresholds used by other groups with different analysis methods
- These recommended SNP thresholds may change as CDC's wgSNP analysis methods are further developed or based on results of a formal validation analysis of SNP thresholds

SNP thresholds for categorizing M. tuberculosis isolates as genetically distant or closely related have not been formally established for CDC's wgSNP analysis yet

Based on CDC's general experiences using wgSNP analysis for investigating recent transmission:

Isolates with 0-5 SNP differences are considered closely related

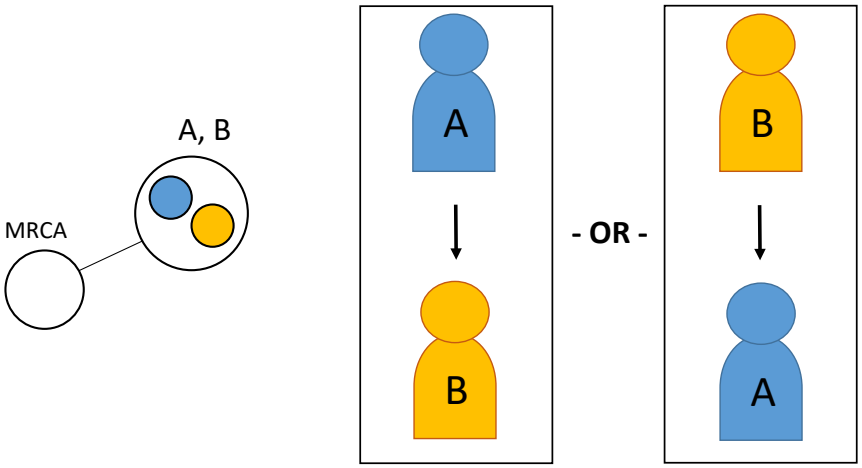
And isolates with 6 or more SNP differences are considered genetically distant

SNP thresholds will vary depending on the methods used for the wgSNP analysis, and cannot be compared to thresholds used by other groups with different analysis methods

These recommended SNP thresholds may change as CDC's wgSNP analysis methods are further developed or based on results of a formal validation analysis of SNP thresholds

Phylogenetic tree is not the same as a transmission diagram

Directionality of transmission cannot be inferred from wgSNP analysis alone

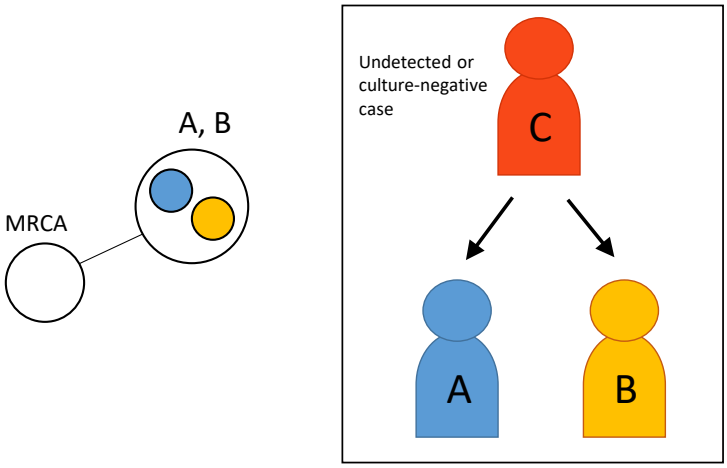


An important consideration to keep in mind when looking at these trees is that the phylogenetic trees are not the same as transmission diagrams because directionality of transmission cannot be inferred from wgSNP analysis alone

If we look at this simple tree on the left, isolates A and B are identical and it's possible that patient A could have transmitted to patient B or patient B could have transmitted to patient A

Phylogenetic tree is not the same as a transmission diagram

Consideration #1: Directionality cannot be inferred because cases involved in transmission may not be included on tree

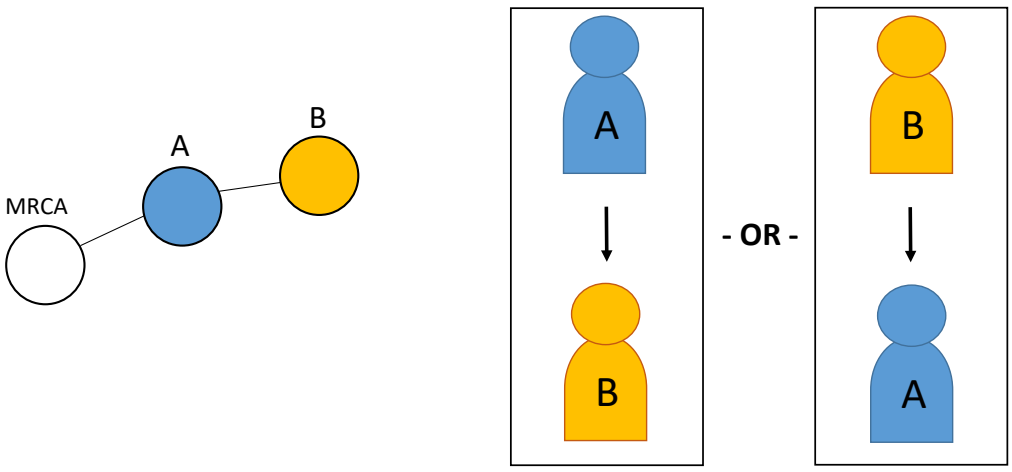


Directionality of transmission also cannot be inferred because there could be cases involved in transmission that are not included in the WGS analysis

For this same tree, it is possible that there is no transmission between patient A and B and transmission was through a third case that does not have an isolate on the tree because they are an undetected case or were culture-negative

Phylogenetic tree is not the same as a transmission diagram

Consideration #2: Directionality cannot be inferred because genetic changes could occur between the time of transmission and collection of the patient's sample



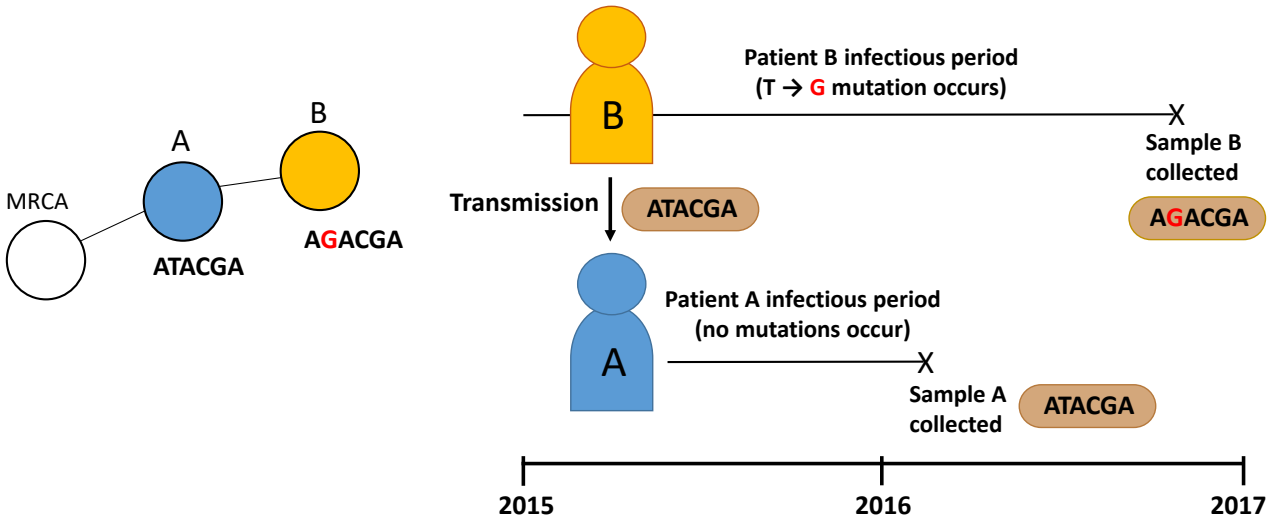
Another consideration is there could be genetic changes that occur between the time of transmission and collection of the patient's sample

With this tree, it is tempting to think that patient A transmitted to patient B since isolate B is shown to have evolved from isolate A

That could be true but it is actually still possible that patient B could have transmitted to patient A

Phylogenetic tree is not the same as a transmission diagram

Consideration #2: Directionality cannot be inferred because genetic changes could occur between the time of transmission and collection of the patient's sample



To illustrate this concept with an example, patient B could transmit to patient A in 2015, so both patients have this same genome type in 2015

Patient A's sample is collected about one year later and no mutations have occurred during that time

Patient B's sample is collected even later and during the time period between transmission and collection of the sample, patient B could have a mutation (here a T to G) in his infecting strain, which would result in a tree that looks like this

For these reasons, we don't use the trees to try to infer transmission between patients. We use them to identify clusters of cases that may be due to recent transmission

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

- Even isolates that are closely related or identical by WGS can be due to reactivation
  - This is because mutations may not occur as frequently during latent infection and therefore SNPs may not accumulate
- The phylogenetic tree should be used in conjunction with clinical and epidemiologic information to assess recent transmission

It is also important to remember that it is easier to rule out recent transmission than confirm it in using WGS

Even isolates that are closely related or identical by WGS can be due to reactivation

This is because mutations may not occur as frequently during latent infection and therefore SNPs may not accumulate

So the phylogenetic tree should be used in conjunction with clinical and epidemiologic information to assess recent transmission

Summary

- WGS can provide greater resolution than GENType for investigating recent TB transmission
- CDC has used WGS retrospectively to examine genetic relatedness of isolates clustered by GENType
- wgSNP analysis is performed to produce a phylogenetic tree for examining genetic relationships between isolates in a GENType cluster
- The phylogenetic tree should be used in tandem with epidemiologic data to identify clusters of closely related isolates that may indicate recent transmission, not to draw conclusions about direction of transmission among individual patients

In summary, WGS can provide greater resolution than GENType for investigating recent TB transmission

CDC has used WGS retrospectively to examine genetic relatedness of isolates clustered by GENType

wgSNP analysis is performed to produce a phylogenetic tree for examining genetic relationships between isolates in a GENType cluster

The phylogenetic tree should be used in tandem with epidemiologic data to identify clusters of closely related isolates that may indicate recent transmission, not to draw conclusions about direction of transmission among individual patients

Part 2

Case Studies: Using WGS to investigate TB cluster alerts in California

Now with that background, I will turn it over to Martin and Tambi who will present the case studies using WGS to investigate TB cluster alerts in California

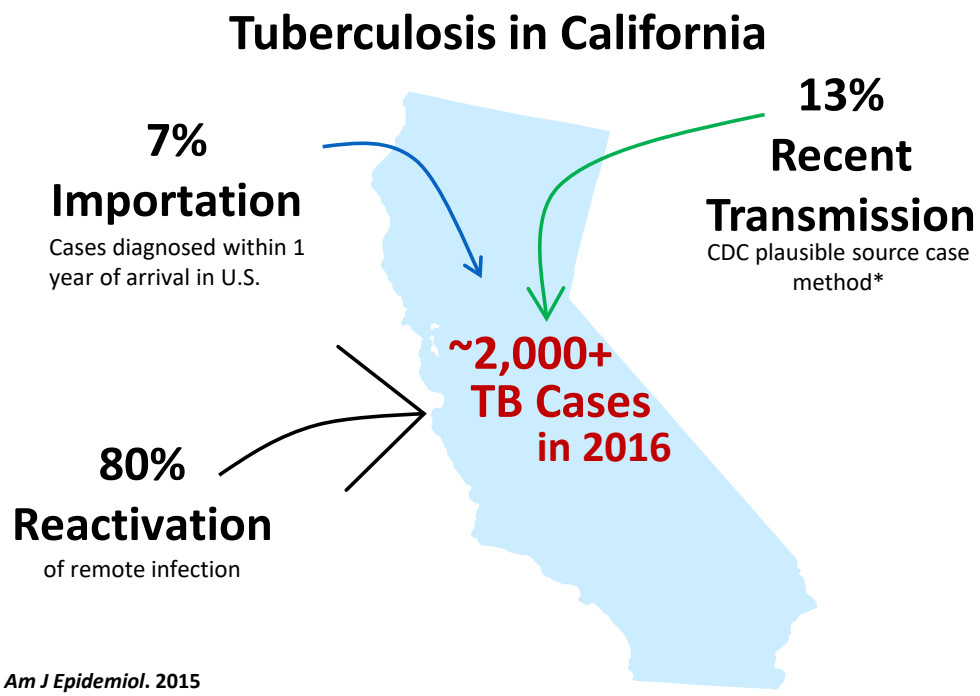
Image: http://www.business2community.com/public-relations/make-sure-youre-barking-up-the-right-tree-0320344#!bPmnsw

Conventional genotyping methods are an important tool to identify and investigate genotyping clusters that might be outbreaks or transmission events.

Sometimes our cluster investigations lead us to previously unknown and important outbreaks where intervention is feasible and warranted. We find the red cat, if you will, because conventional genotyping leads us to bark up the right tree.

But, there are limitations of conventional TB genotyping and some of our cluster investigations lead to barking up the wrong tree…like the dog in this image…when we investigate genotype clusters of cases that might not be linked in a chain of transmission.

Our talk will describe how WGS can help us make sure we're barking up the right tree.



*** France et al, *Am J Epidemiol*. 2015**

Some background about TB in CA to provide some context for our talk

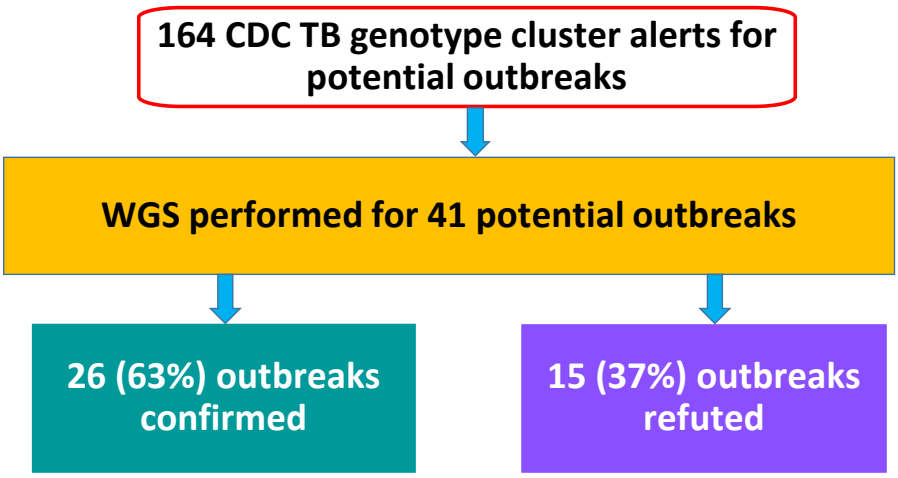
CA reported more than 2,000 cases of TB in 2016, more than any other state in the U.S.

When we look at why cases occur, there are 3 primary drivers of morbidity.

- The reactivation of infection acquired a long time ago is responsible for about 80% of cases.
- Importation of TB represents about 7% of the cases.
- Recent TB transmission within CA is estimated to be responsible for about 13% of cases

Our focus today is on recent transmission

Impact of WGS on TB Outbreak Classifications, California, 2013 – 2016



Before we dive into our case studies, let's step back and quickly review how WGS has impacted our TB outbreak work in California.

So, in the past 4 years, there have been 164 CDC TB genotype cluster alerts, which represent potential outbreaks.

We analyzed data from all of our TB outbreak investigations from 2013 to 2016 and for which we had WGS results.

WGS was performed on genotype clusters associated with 41 potential TB outbreaks.

After considering WGS results, along with clinical and epi data, 63% or almost two-thirds of the clusters were confirmed as outbreaks. In many of these confirmed outbreaks, WGS helped identify cases that could be excluded from further investigation because their TB isolates were genetically distant from the outbreak case isolates and unlikely to be linked by recent transmission to the outbreak.

Importantly, more than 1/3 were refuted as outbreaks. All 15 of the refuted outbreaks started as suspected outbreaks. These refuted outbreaks are examples where reliance on conventional TB genotyping would have led us to bark up the wrong tree—we and local partners would have spent precious time and resources looking for epi links between cases that did not exist.

Case Study 1

Confirmed outbreak in a high TB incidence jurisdiction

Next, I will talk about the first case study which was a confirmed outbreak in a high TB incidence jurisdiction.

Background

- CDC alert for a TB GENType cluster in County A
  - 8 of the 13 cases in California with the GENType lived in County A
- 6 cases in County A had known epi links: a confirmed outbreak involving a high school and 2 households
- Unknowns
  - Are the 2 remaining cases in County A also part of the outbreak?
  - Are the 5 California cases outside of County A part of the outbreak?
  - Are any of the 7 cases not part of the outbreak linked to each other in a separate chain of transmission?
  - Where to focus further work to interrupt TB transmission?
- Requested CDC perform WGS

We received a CDC TB genotype cluster alert for a genotype cluster in County A

When we examined this cluster more closely, we found that there were 13 CA cases with the genotype and 8 of them lived in County A. The remaining 5 cases were scattered across 4 other counties.

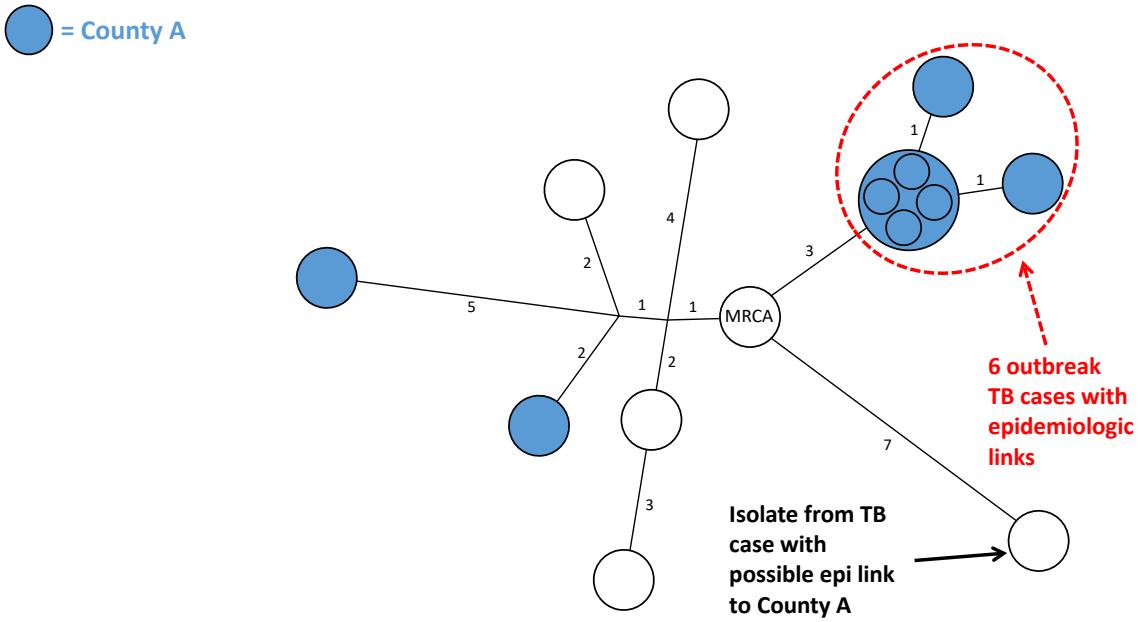
We notified the County A about the CDC alert. They were already aware that 6 of these cases in their County had known epi links. This was actually a confirmed outbreak involving a high school and 2 households.

But, we and County A were left with some unknowns:

- Are the 2 other cases in County A part of the outbreak?
- Are the 5 California cases outside of County A part of the outbreak?
- Are any of the 7 cases not part of the outbreak linked to each other in a separate chain of transmission?
- Where should we focus further work to interrupt TB transmission for this outbreak?

To help answer these questions we asked CDC to perform WGS on all 13 California cases in the genotype cluster.

Phylogenetic Tree + Epi Data



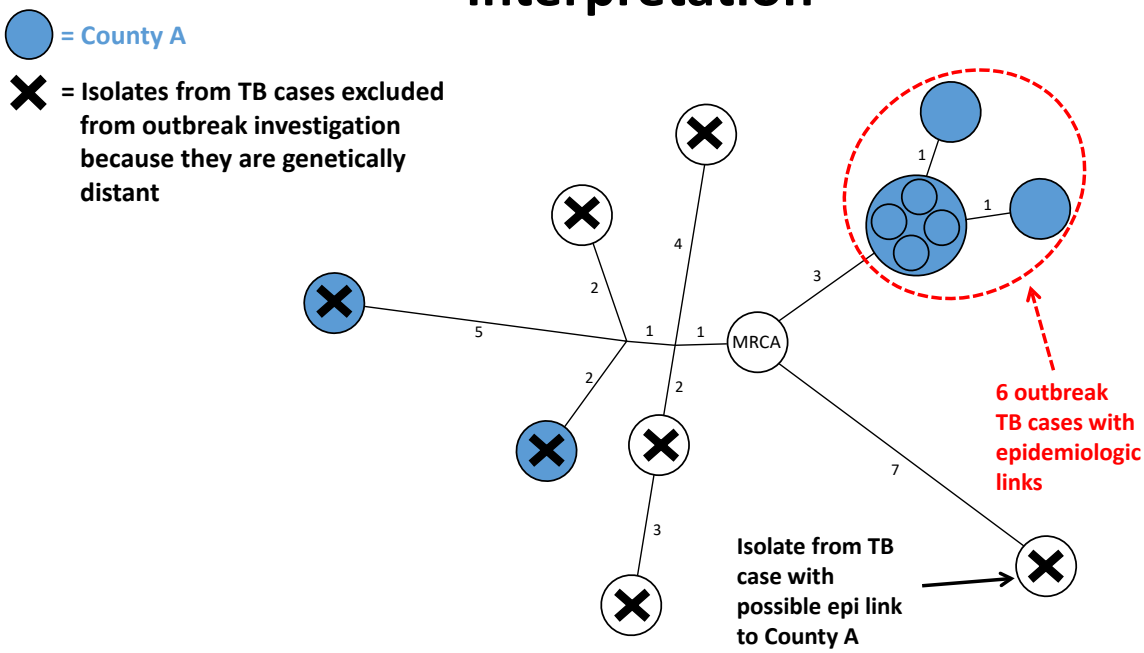
Here is the phylogenetic tree for all 13 cases in the genotype cluster. The circles represent isolates from the TB cases and lines connect the most-related isolates. The numbers on the lines indicate the number of nucleotide differences between isolates. MRCA is the most recent common ancestor which is a theoretical isolate from which all of the isolates in the cluster are directly descended. TB cases diagnosed in County A are shaded in blue.

The most notable feature of the tree is a group of isolates to the right of the MRCA which is circled by a red-dashed line. These are isolates from 6 TB outbreak cases with already-known epidemiologic links. The large circle represents 4 isolates with identical sequences (that is, there were no SNPs among the 4 isolates). Protruding from the large circle are 2 isolates that are closely-related, with only one SNP difference.

Also to the right of the MRCA is an isolate from a TB case with possible epi links to County A. The case reportedly stayed at a relative's house in County A for a considerable amount of time. However, the isolate is quite genetically distant from the outbreak cases, with 10 or more SNPs from the outbreak cases, which suggests that the case is not part of the outbreak.

To the left of the MRCA are 6 isolates, 2 of which are from County A TB cases. You can see from the tree that the sequences of the 6 isolates are different from the outbreak cases and are quite different from each other.

Interpretation

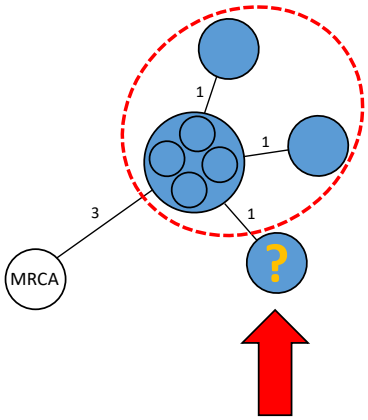


After we add clinical and epi data to the tree, we interpret the results. So, let's think back to the unknowns we listed at the start of this investigation:

- Are the 2 other cases in County A also part of the outbreak? No, they are genetically distant from the nearest outbreak cases by 7 and 10 SNPs, respectively.
- Are the 5 California cases outside of County A part of the outbreak? No, they are genetically distant from the nearest outbreak case by 6 to 10 SNPs?
- Are any of the cases that aren't part of the outbreak linked to each other in a separate chain of recent transmission? No, they are 3 to 14 SNPs from each other and there is no strong epi or clinical data to suggest they are linked.

New Clustered TB Case

= County A



Isolate from a recently diagnosed patient in County A

But the story does not end there. Several months after the initial genotype cluster alert, another case from County A genotyped into the cluster.

We notified County A about the new case in the alerted cluster. The county TB program did not know of any epi links the new case had to any of the other cases in the genotype cluster.

The case was a health care worker whose TB skin test recently converted to positive after years of negative TB tests.

WGS was performed on the HCW's TB isolate and it grouped together with the outbreak cases where the red arrow is pointed.

The county is investigating if/how the case is linked to the outbreak. The concern is that this HCW might have acquired TB on the job and that other HCWs and patients might have also been exposed and are now at risk for TB.

Case Study 1: Public Health Outcomes

- Avoided unnecessary investigation of 7 cases, including 5 residing in different counties outside of County A
- WGS results enabled continued focus on 6 cases linked by recent transmission
- County A intensified work to identify, evaluate, and treat contacts to outbreak cases
- County A also investigating the new patient whose TB is genetically closely related to the outbreak to determine if/how linked to outbreak

The public health outcomes for this investigation were:

- We avoided unnecessary investigation of seven cases, including five residing in different counties outside of County A
- WGS results enabled continued focus on six cases linked by recent transmission
- County A intensified work to identify, evaluate, and treat contacts to outbreak cases
- County A also was investigating the new patient whose TB is genetically closely related to the outbreak to determine if and how the patient is linked to the outbreak

Case Study 2

Refuted outbreak in a low TB incidence jurisdiction

In the second case study, I will present a refuted outbreak in a low TB incidence jurisdiction

Background

- In an 8-month period, 4 GENType-matched TB cases were initially detected in a rural county
- The county typically has about 10 TB cases per year
- Few GENType clusters in the county
- GENType is relatively uncommon in the state
- Local staff are relatively new to TB control; have responsibilities other than TB

We received a CDC TB genotype cluster alert for 4 cases with matching TB genotypes diagnosed in an 8-month period in a rural county with low TB morbidity—usually this county reports about 10 cases per year. GENType clusters are rare in the county.

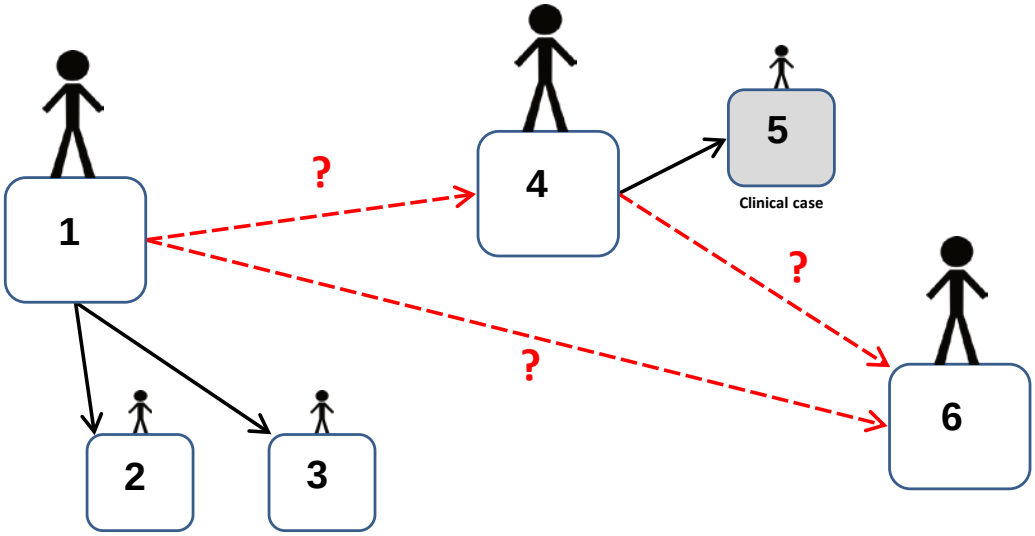
In addition, the genotype that alerted in the county is rare in CA and in the US. In the 3 years preceding the alert there were only about 17 other cases in the entire US with this genotype. The more uncommon a TB genotype is in the US, the more likely it is that clusters of cases with that genotype are linked by recent transmission. So, the rarity of this TB genotype plus the clustering in time and place in a low morbidity TB jurisdiction were concerning for a possible outbreak

When we looked more closely at the surveillance data for the 4 cases that generated the cluster alert, we saw that 2 of the patients were adult men with sputum smear positive TB and 1 had a chest x-ray showing cavitation—characteristics associated with more infectious forms of TB. Both also were reported to be drug users, and we know that cases with that risk factor can present challenges for contact investigations.

We also noted that 2 of the culture-confirmed cases were in young, US-born children—one was an infant and the other was 5 years old. By definition, TB in young children is a red flag for recent TB transmission.

And the county public health staff were all relatively new to TB and were juggling multiple responsibilities in addition to TB.

Investigation



When we contacted the county TB program, we found their contact investigations of the adult cases had identified some epi links.

County staff knew that Case 1, an adult male, was the likely source case to 2 pediatric cases: one infant and a 5 year old child. The dark lines represent definite epi & transmission links.

They also knew that Case 4, an adult male, was the likely source case for a clinically-diagnosed case in a child. We hadn't been aware of this pediatric case being part of the cluster until we spoke with the county.

Two groups of cases with epi links but no known links between the 2 groups:

1 adult and 2 children ≤ 5 years old (Cases 1, 2, and 3)

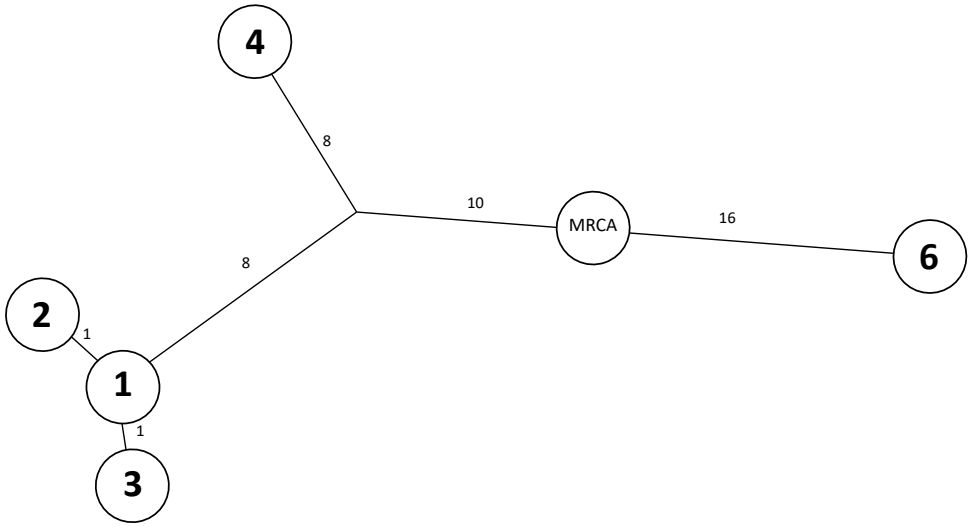
1 adult and a clinically-diagnosed case in a child (Cases 4 & 5)

So, at that point we had a 5-person suspected TB outbreak. We also learned that the contact investigations for the infectious adult cases were challenging and the contact follow-up and investigations were incomplete.

We offered onsite field investigation assistance and the county accepted. We deployed a 2-person team to the field to assist the county. Despite intensified onsite investigation over a period of about 3 weeks, no connections could be found between the 2 groups.

We asked CDC to perform WGS. A few months later, the genotype for another adult case in the county (Case 6) matched to the cluster. No epi links could be found for Case 6.

Phylogenetic Tree

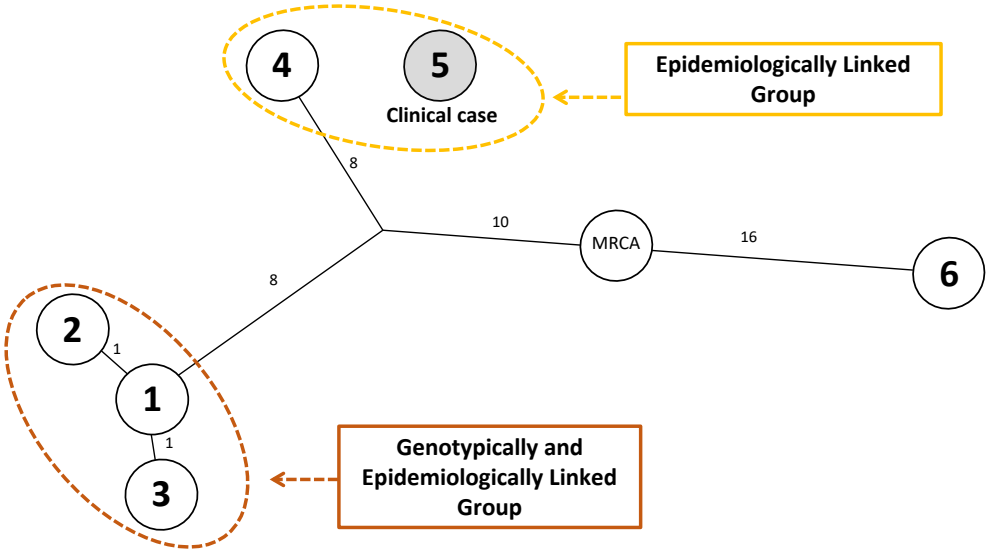


This phylogenetic tree shows the analysis of the WGS results we received back from CDC.

You can see that isolates from Cases 1, 2, and 3 are genetically closely related and are separated by only 1-2 SNPs.

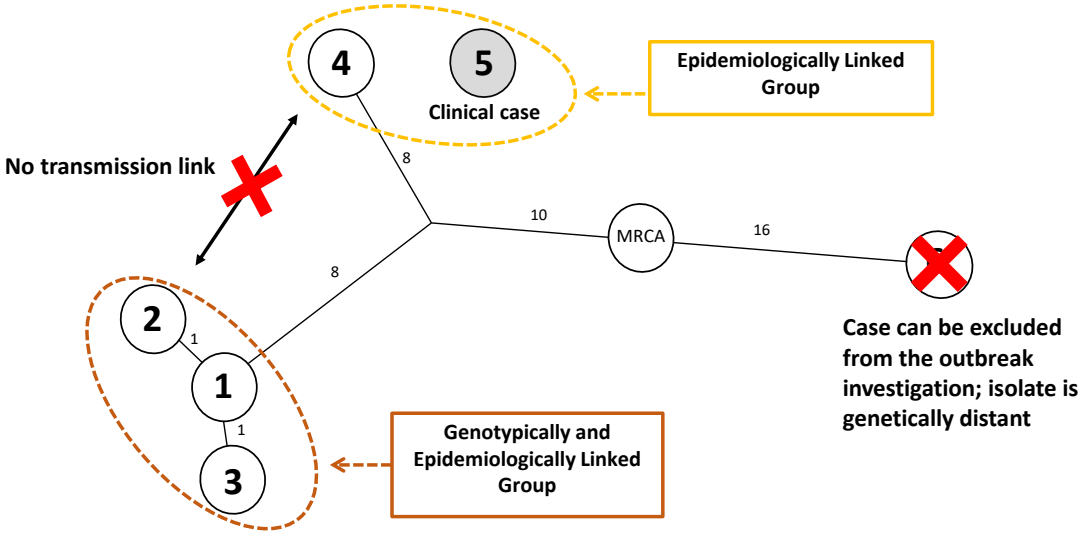
The isolates from Cases 4 and 6 are genetically distant from each other; they are separated by 34 SNPs (8+10+16). They are also genetically distant from the isolates of Case 1, 2, and 3.

Phylogenetic Tree + Epi Data



When we received the phylogenetic tree from CDC, we added the epi link results from the investigation we had so far by circling the nodes representing cases with epi links. To help us interpret the results, we also added to the tree Case 5 -- the clinically-diagnosed pediatric case that had no TB isolate to sequence.

Interpretation



After overlaying epi data to the tree, we analyzed and interpreted all available data to understand transmission dynamics in this suspected outbreak.

As you can see in the bottom left portion of the tree, we found that WGS results corroborated the already-known transmission links between Adult Case 1 and Pediatric Cases 2 & 3. You can see that the TB genome from the two pediatric cases differed by only one SNP from the TB genome of the adult source case. This tells us these cases have TB that is genetically closely-related and consistent with recent transmission.

Importantly, the WGS results showed substantial genetic distance of 16 SNPs between these 2 groups of cases—the ones at the lower left and upper part of the tree. We had looked hard for an epi link between the 2 adult cases 1 and 4 and we were concerned we had missed a connection—potentially a drug-related connection--and potentially had failed to identify an important transmission site or other contacts who were at risk for TB.

We and county colleagues were relieved to learn that WGS results corroborated the epi investigation finding that there was no link identified between those 2 adult cases; we could stop looking for a recent transmission link because one didn't exist.

We were also happy to learn that Case 6 on the far right side had TB that was genetically distant by 34 SNPs from the other adult cases. Case 6 is also on the other side of the MRCA meaning that case had distinct SNPs that other cases in the genotype cluster did not share, further indicating Case 6 is unlikely to be related to the other cases by recent transmission. We were able to exclude Case 6 from the investigation.

Case Study 2: Summary

- WGS results showed two separate chains of limited recent transmission
- WGS corroborated initial finding that Case 1 was the source case to two pediatric cases
- TB program can focus efforts on ensuring that each adult case had a complete contact investigation

To summarize, WGS results showed two separate chains of limited recent transmission, WGS corroborated initial finding that Case 1 was the source case to two pediatric cases, and the TB program can focus efforts on ensuring that each adult case had a complete contact investigation

Case Study 2: Public Health Outcomes

- Identified additional high-priority contacts during the GENType cluster investigation
- Intensified follow-up of contacts to ensure evaluation and treatment
- Developed local protocol for using new short-course regimen for treating TB infection
- Provided contact investigation training of local health department staff

As for public health outcomes for this investigation, we identified additional high-priority contacts during the GENType cluster investigation, intensified follow-up of contacts to ensure evaluation and treatment, developed local protocol for using new short-course regimen for treating TB infection, and provided contact investigation training of local health department staff

WGS Limitations

- No SNP thresholds have been formally validated
- WGS results were generally not available early in investigations but turnaround times will improve as lab capacity expands

It is important to recognize the limitations of our analysis. These include but aren't limited to:

- No SNP threshold has been formally validated as a gold standard for identifying cases likely linked by recent transmission. While our experience to date suggests the SNP thresholds we used were very concordant with epi links (or lack thereof), a more formal analysis and reporting of those concordance data would be another study.
- WGS results were generally not available early in investigations; most WGS results analyzed mid-course in investigations or retrospectively. But, we're already observing that turn-around-times are improving as lab capacity for TB sequencing and phylogenetic analysis continues to expand. Analysis methods are also becoming more automated which will speed turn-around-times.

Conclusions

- Combined analysis of clinical, epidemiologic, and phylogenetic data can help focus TB investigations
- WGS results can:
  - More precisely identify outbreaks and outbreak cases
  - Avoid unnecessary investigations of clusters with cases not linked by recent transmission

I hope we have helped describe how the combined analysis of clinical, epi, and phylogenetic data can help focus TB investigations.

WGS results can help us make sure we're barking up the right tree by enabling us to more precisely identify outbreaks and outbreak cases. Importantly, these data can also help us avoid unnecessary investigations of clusters with cases not linked by recent transmission.

Part 3

Plans for transition to universal prospective WGS

Thanks, Martin and Tambi

Now in part 3, I will briefly describe the plans for transition to universal prospective WGS

A separate presentation covering the details of how universal prospective WGS will be implemented will be made available in the future

Universal Prospective WGS begins in 2018

- WGS of isolates from all new culture-confirmed cases of TB
- GENType will continue to be analyzed during an initial 3 year transition period (2018 – 2020)
  - GENType will be reported in TB GIMS
  - Cluster alerts will be based on GENType
- In 2021, WGS will become the standard method for genotyping
- WGS data will be used for two separate analyses to examine transmission
  - wgMLST (whole-genome multi-locus sequence typing)
  - wgSNP (whole-genome single nucleotide polymorphism analysis)

Universal prospective WGS begins in 2018 and WGS of isolates from all new culture-confirmed cases of TB will be performed

But GENType will continue to be analyzed during an initial 3-year transition period, which will be 2018 through 2020

During this time, GENType will continue to be reported in TB GIMS and cluster alerts will be based on GENType

In 2021, WGS will become the standard method for TB genotyping

WGS data will be used for two separate analyses to examine transmission: wgMLST (which is whole-genome multi-locus sequence typing) and wgSNP (which is whole-genome single nucleotide polymorphism analysis)

And I will explain these two analyses in more detail in the following slides

Analysis of clustering using WGS data: wgMLST vs. wgSNP

	wgMLST (whole-genome multi-locus sequence typing)	wgSNP (whole-genome single nucleotide polymorphism)
Level of analysis	all isolates	isolates in a cluster
Use	assigning isolates to a wgMLSType that can be used for cluster alerting	examining genetic relationships among isolates
Output	wgMLSType (short string of numbers similar to a GENType)	phylogenetic tree

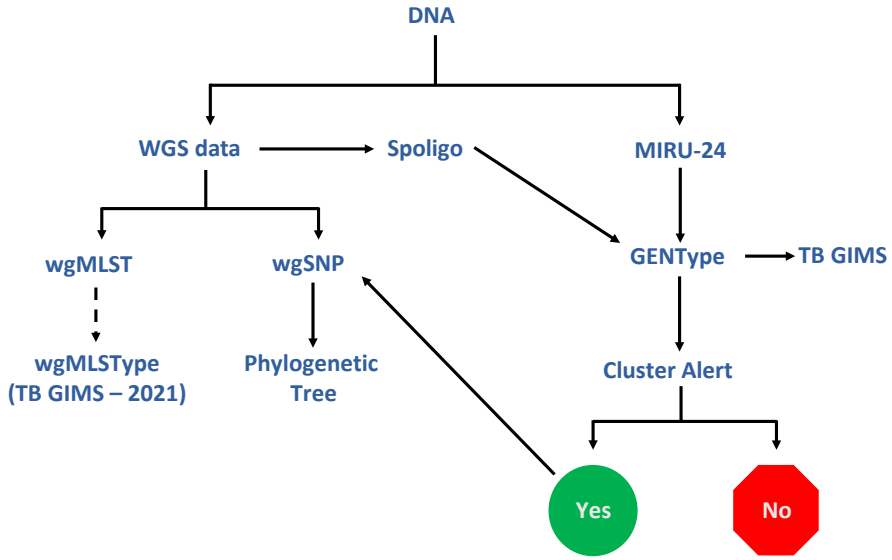
Although wgMLST and wgSNP both use WGS data, they differ because wgMLST is a scheme that will be applied to all isolates to compare the genomic sequences of the isolates and assign them to a wgMLSType that can be used for cluster alerting

The wgMLSType is a standard naming scheme that will be a short string of numbers similar to a GENType

On the other hand, we use wgSNP to examine the genetic relationships among isolates in a cluster in more detail (and the cluster can be based either on GENType or wgMLSType) and the output is a phylogenetic tree that diagrams these genetic relationships

Universal prospective WGS begins in 2018

TB Genotyping Methods and Data Flow (2018 – 2020)



This is what the TB genotyping methods and data flow will look like for the 3-year transition period

During this time, we will be performing whole-genome sequencing as well as conventional genotyping with spoligotype and 24 locus MIRU for each isolate

GENType will continue to be analyzed and reported in TB GIMS and GENType will be used for cluster alerting

The WGS data will be used for wgMLST and wgSNP analysis

However, wgMLSType won't be displayed in TB GIMS until after the initial three year overlap period in 2021

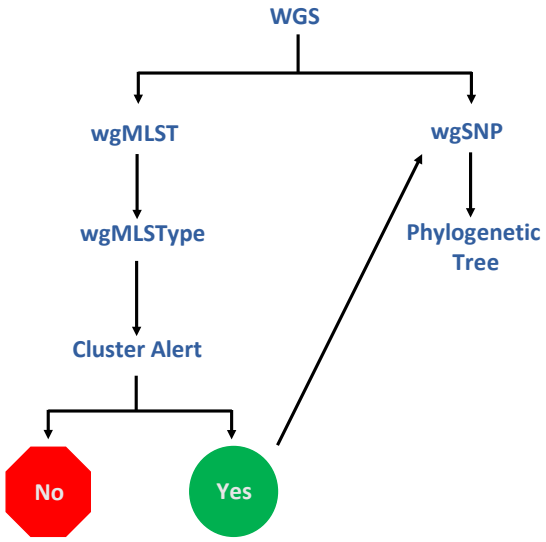
This will allow for the necessary time to adapt algorithms for using wgMLST to define clusters

For clusters that alert during this time based on GENType, wgSNP analysis will be performed to assess the potential of recent transmission and this information will be communicated to the state program

For clusters that have had previous retrospective WGS performed, any new isolates in the cluster will be added to the analysis and the updated results will be reported to the state program

wgMLSType will replace GENType for cluster alerting in 2021

TB Genotyping Methods and Data Flow (2021)



In 2021, WGS will become the standard method performed to identify TB clusters and wgMLSType will be used to generate cluster alerts

For clusters that alert, the genetic similarities among the clustered isolates can be examined in more detail with wgSNP analysis

Acknowledgments

- TB Outbreak Prevention and Control Section, California Department of Public Health
- Applied Research Team and Molecular Epidemiology Activity, DTBE
- Microbial Diseases Lab, California Department of Public Health
- Association of Public Health Laboratories

We would like to thank the TB Outbreak Prevention and Control Section at the California Department of Public Health, the Applied Research Team and Molecular Epidemiology Activity at DTBE, the Microbial Diseases Lab at the California Department of Public Health, and the Association of Public Health Laboratories

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.