Centers for Disease Control and Prevention
National Center for HIV, Viral Hepatitis, STD, and TB Prevention

Transition to whole-genome multilocus sequence typing (wgMLST) for TB cluster detection

Sarah Talarico, PhD, MPH

Laboratory Branch and Surveillance, Epidemiology, and Outbreak Investigations Branch

Outline

- Background on TB genotyping and use of whole-genome sequence data for examining TB transmission
- Methods behind using whole-genome multilocus sequence typing (wgMLST) for cluster detection
- How clustering with wgMLSType compares to clustering with GENType
- Where we are in the transition

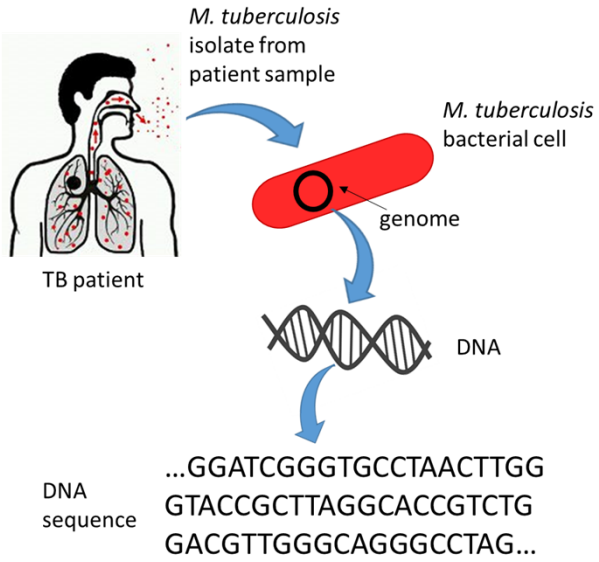
- My presentation will first go over some background on TB genotyping and use of whole-genome sequence data for examining TB transmission
- Then I will explain the methods behind using whole-genome multi-locus sequence typing for cluster detection
- After that, I will present some data showing how clustering with wgMLSType compares to clustering with our current GENType
- And then will finish with a brief overview of where we are in this transition to using whole-genome sequence data for TB cluster detection

TB genotyping and use of whole-genome sequence data for examining TB transmission

- First, some background on TB genotyping and use of whole-genome sequence data for examining TB transmission

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

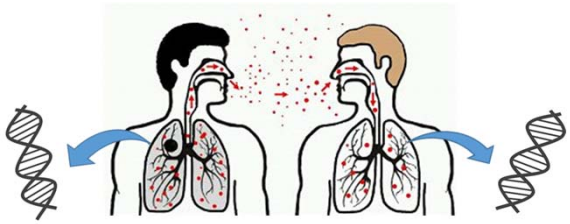
- The *M. tuberculosis* bacteria cultured from a TB patient is called the patient's isolate
- Bacteria, including *M. tuberculosis*, have DNA called a genome
- DNA is made up of four different nucleotides (abbreviated A, T, C, and G)
- The order of these nucleotides in the genome is the DNA sequence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



- 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 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

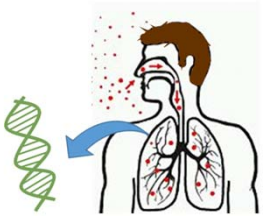
Genotyping can be used to identify TB patients who are more likely to be linked by recent transmission

- 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)

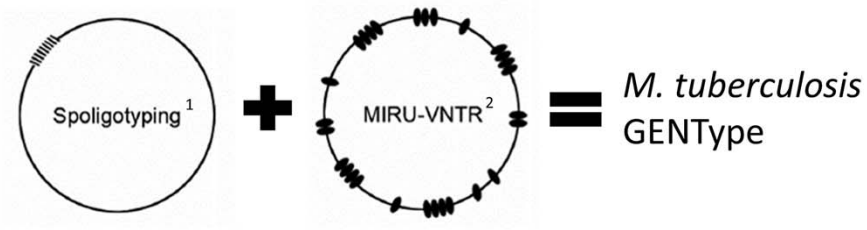
TB patient not linked by recent transmission has isolate with different genotype (green)



- 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)

National TB Genotyping Service

- Established in 2004
- Genotype one isolate from each culture-confirmed case
- ~ 9,000 isolates are genotyped each year



Presence or absence of 43 spacer sequences in direct repeat region

Differences in copy number of tandem repeats at 24 loci

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

- Back in 2004, the national TB genotyping service was established with the goal of genotyping one isolate from each culture-confirmed TB case in the U.S.
- Approx. 9,000 isolates are genotyped each year
- The current method combines the results of two assays, spoligotyping and MIRU-VNTR, to give an M. tuberculosis GENType
- 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

National TB Genotyping Service

	Spoligotype	MIRU1 (12 loci)	MIRU2 (12 loci)	GENType
Isolate 1	000000000003771	223325173533	444534423428	G00010
Isolate 2	000000000003771	223325173533	444534423428	G00010

- The presence or absence of the 43 spacer sequences is boiled down to a 15-digit number that is the spoligotype using octal code
- The number of repeats at each of the 24 MIRU loci are also used to generate a pattern
- The first 12 loci are the MIRU1 pattern and the second 12 loci are the MIRU2 pattern
- Isolates that have the exact same spoligotype and 24 locus MIRU pattern are assigned the same GENType, which are named as the letter G followed by 5 numbers
- You can see in this example that these isolates match exactly and they are both designated as G00010

National TB Genotyping Service

- **GENTypes that differ at one locus**
  - SLV = Single Locus Variant
  - MML = Mixed or Missing Locus

	Spoligotype	MIRU1 (12 loci)	MIRU2 (12 loci)	GENType
Isolate 1	000000000003771	223325173533	444534423428	G00010
Isolate 2	000000000003771	223325173533	444534423428	G00010
Isolate 3	000000000003771	223325163533	444534423428	**G00818**

→ G00818 is a single locus variant of G00010

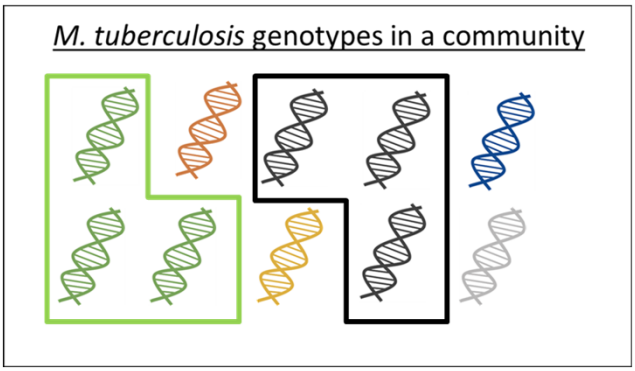
- Some GENTypes differ from each other at only one locus. These are referred to as single locus variants or SLVs
- Here in this example, Isolate 3 has 6 repeats at one of the loci in MIRU1 where Isolate 1 and Isolate 2 have 7 repeats
- Since Isolate 3 does not have a pattern that is an exact match, it is assigned a different GENType, G00818
- And we would say that G00818 is a single locus variant of G00010
- Similarly, if the difference is due to a mixed locus or missing locus it is referred to as an MML for mixed or missing locus
- In practice, TB cases with isolates that are SLVs or MMLs can sometimes be linked through recent transmission

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 might be due to recent transmission

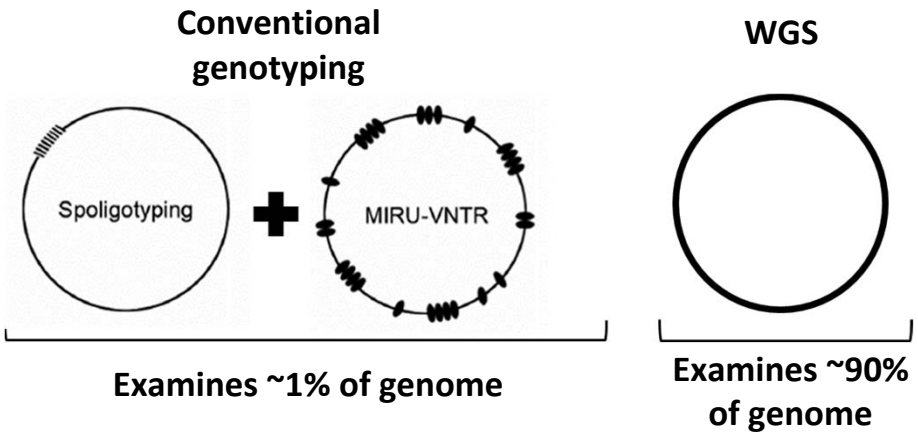
<u>CDC cluster detection methods</u>

- 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



- 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

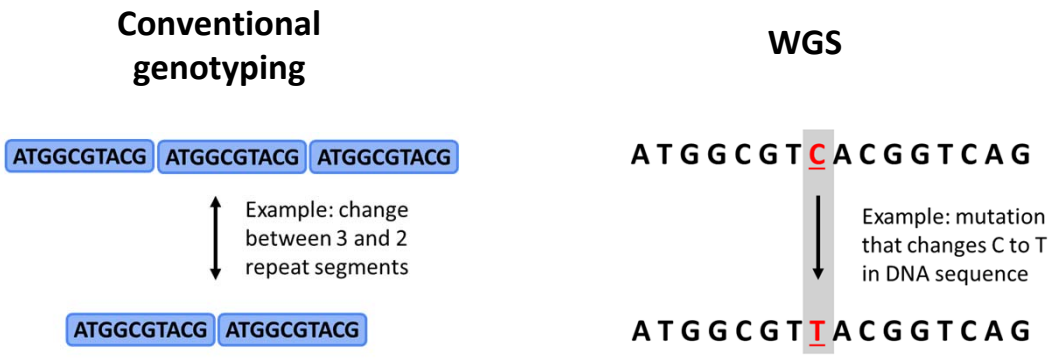
Whole-genome sequencing (WGS) can provide added resolution for examining genetic relatedness of isolates



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

- However, the current conventional genotyping methods provide relatively low resolution for examining the genetic relatedness of isolates because they only examine a small portion of the genome, about 1%
- Whole-genome sequencing (or WGS) can provide added resolution by expanding the coverage of the genome to about 90%, capturing much more of the genomic changes that occur

Whole-genome sequencing (WGS) can provide added resolution for examining genetic relatedness of isolates

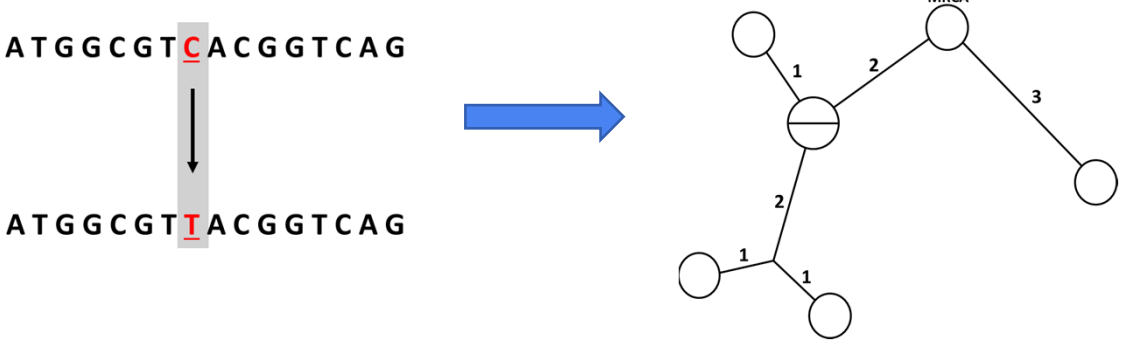


Gain or loss of segments of DNA sequence within certain repetitive regions

Single nucleotide polymorphisms (SNPs) throughout the genome

- WGS can also provide added resolution by examining changes at the nucleotide level
- Conventional genotyping examines the gain or loss of segments of DNA sequence within certain repetitive regions as depicted on the left by an example of a change between 3 and 2 repeat segments at a particular location in the genome
- There is a bidirectional arrow to indicate that these changes are reversible
- On the other hand, WGS examines millions of nucleotides throughout the genome to identify single nucleotide polymorphisms (or SNPs)
- A SNP is a mutation at a single position in the DNA sequence
- Here on the right, I'm showing an example of a mutation that changes a C to a T in the DNA sequence
- There is a unidirectional arrow because these mutations are unlikely to reverse

Whole-genome single nucleotide polymorphism comparison (wgSNP)



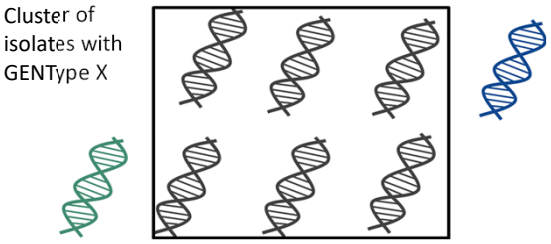
SNPs that differ between isolates in a cluster are identified

SNPs are mapped on to a phylogenetic tree to diagram the genetic relationship among isolates

- One type of analysis that can be performed using WGS data is called whole-genome single nucleotide comparison (or wgSNP)
- We first identify SNPs that differ between isolates in a genotype cluster
- And then we can map the SNPs on to a phylogenetic tree to diagram the genetic relationship among the isolates

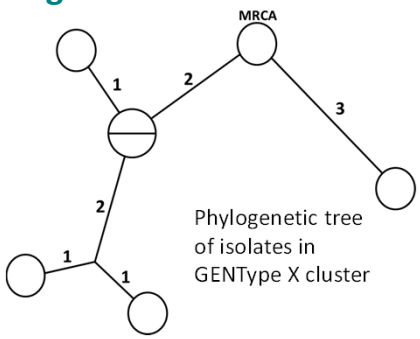
wgSNP comparison can further distinguish clustered isolates

Step 1. Detect a cluster



Conventional genotyping
Isolates are clustered based on matching GENType

Step 2. Examine genetic relationship among isolates in the cluster



Whole-genome SNP comparison
Isolates in a cluster may be further distinguished

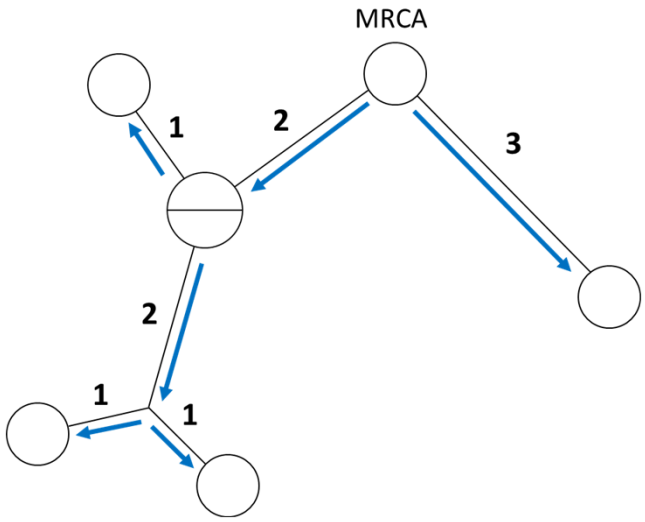
- Since 2012, we have increasingly been using WGS data to perform whole-genome SNP comparison and phylogenetic analysis to further distinguish isolates in a genotype-matched cluster and understand transmission
- Clusters are first detected based on isolates having a matching GENType
- Whole-genome SNP comparison can then be performed to examine the genetic relationship among isolates in the cluster

Results of whole-genome SNP comparison: the phylogenetic tree

- Nodes (circles) represent isolates
- Branches (lines) are proportional in length to the number of SNPs that differ between the isolates

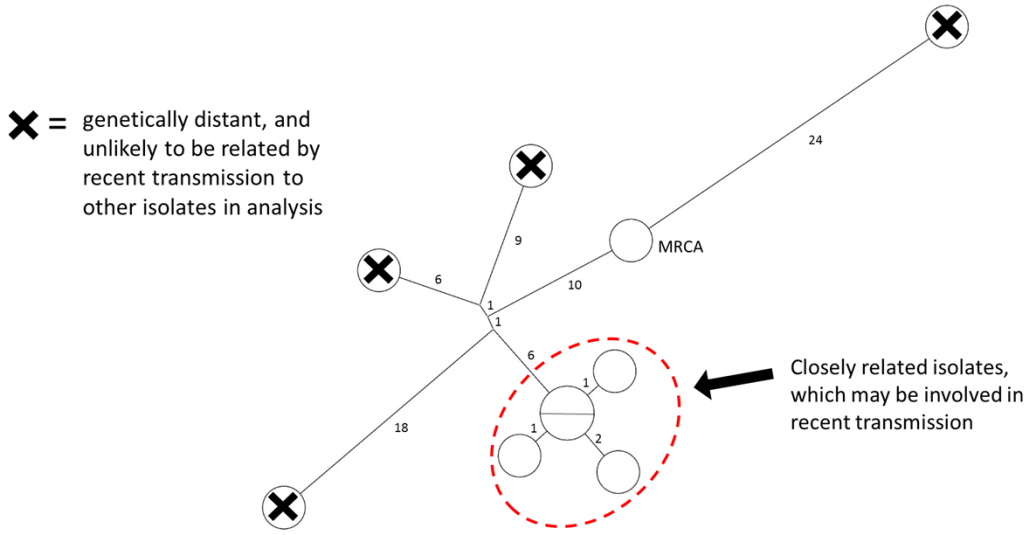
MRCA = Most Recent Common Ancestor

- Hypothetical genome type from which all isolates on the tree are descended
- Serves as a reference point for examining the direction of genetic change (→)



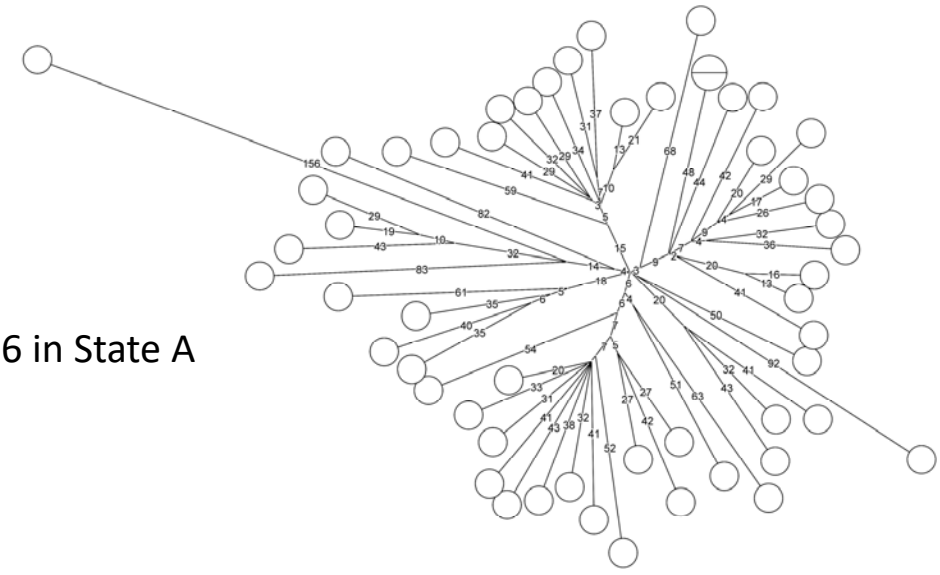
- The results of the whole-genome SNP comparison are delivered to programs in the form of a phylogenetic tree
- The nodes (or circles) represent the isolates and the branches (or lines) that connect the nodes are proportional in length to the number of SNPs that differ between the isolates
- The tree also has a node labeled MRCA, which stands for most recent common ancestor
- It represents a hypothetical genome type from which all isolates on the tree are descended and serves as a reference point for examining the direction of genetic change, which starts at the MRCA and moves out from there as shown by these blue arrows

Phylogenetic trees can be used to inform epidemiologic investigations



- The phylogenetic trees can be used to inform epidemiologic investigations by identifying groups of closely related isolates that may be involved in recent transmission and ruling out genetically distant isolates that are unlikely to be involved in recent transmission
- Identifying TB cases that might be due to recent transmission is important because these represent a chain of transmission for which public health intervention to interrupt further transmission might still be productive

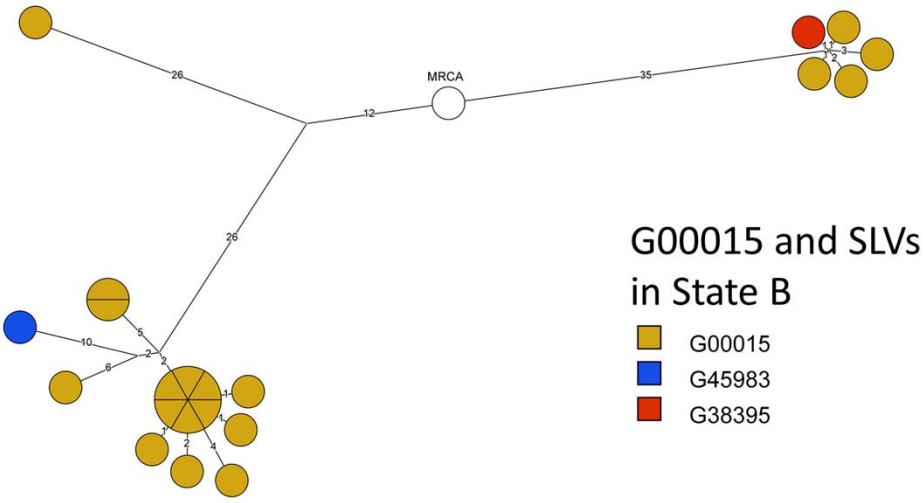
Example 1: Isolates with matching GENType are genetically distant



G00016 in State A

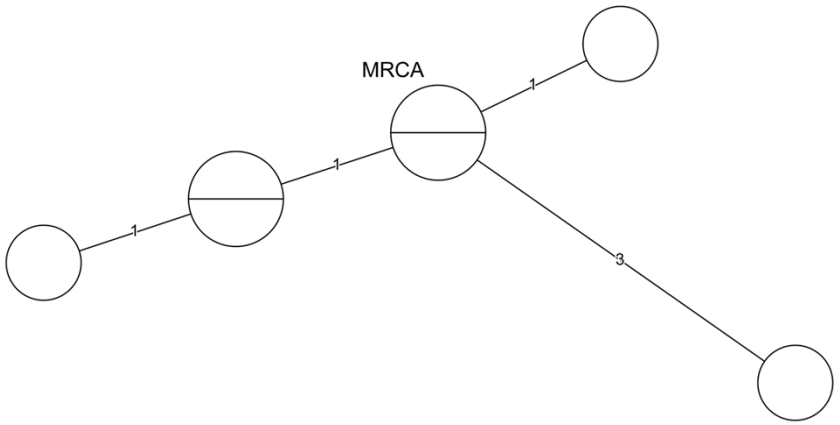
- When a GENType cluster is analyzed by whole-genome SNP comparison, sometimes the isolates are genetically distant from each other and the cluster gets broken up
- We can see that here in this example showing the tree for isolates with the common GENType G00016

Example 2: Isolates with matching GENType separate into subclusters, SLVs can be closely related



- Whole-genome SNP comparison also sometimes separates isolates into subclusters of closely related isolates and can be used to determine if SLVs or MMLs are closely related as is seen in this example of the tree for G00015 isolates, shown in yellow, and SLVs G45983 and G38395, shown in blue and red

Example 3: Isolates with matching GENType are all closely related, phylogenetic tree can provide additional information about genetic relationships

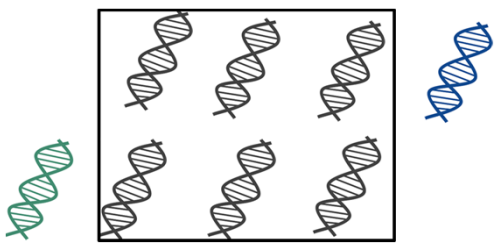


G08090 in State C

- And of course, sometimes all the isolates in the cluster are also closely related based on whole-genome SNP comparison
- But the phylogenetic tree can still provide additional information about the genetic relationships among the isolates
- This information combined with available epidemiologic and clinical data can be used to make inferences about transmission among cases in a cluster

Transition to WGS data for cluster detection

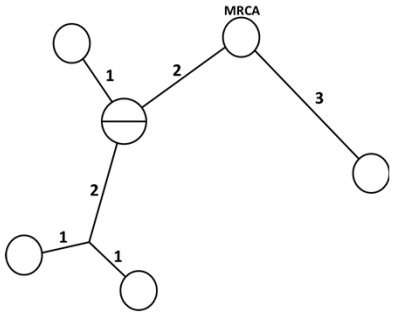
Step 1. Detect a cluster



GENType
(conventional genotyping)

wgMLSType (WGS data)

Step 2. Examine genetic relationship among isolates in the cluster



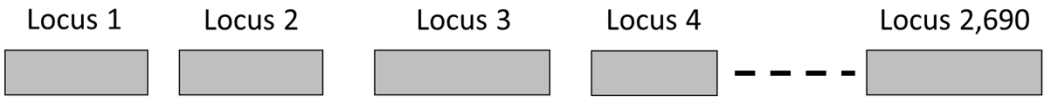
Whole-genome SNP comparison

- In 2018, the National Tuberculosis Molecular Surveillance Center began sequencing all isolates as part of a planned transition to replace conventional genotyping methods with whole-genome sequencing for cluster detection as well
- Instead of using spoligotyping and MIRU to assign isolates a GENType, we have begun using the WGS data to perform whole-genome multilocus sequence typing (wgMLST) to assign isolates a wgMLSType
- And further distinguishing isolates in a cluster will continue to be performed by whole-genome SNP comparison

Whole-genome multilocus sequence typing (wgMLST)

- In this next section, I'll explain what whole-genome multilocus sequence typing is and how it will be used for cluster detection

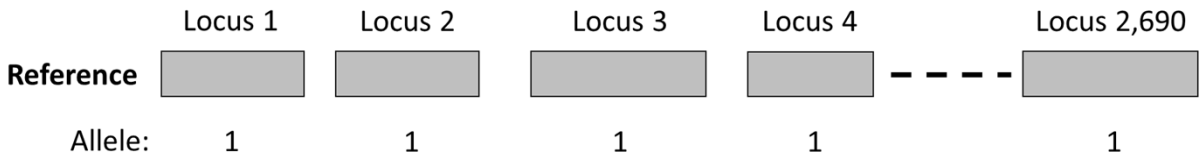
Whole-genome multilocus sequence typing (wgMLST)



- **Compares sequence at 2,690 loci throughout the genome**
  - Covers ~70% of the genome
- **Locus: location in the genome**
  - In this case, each locus is an individual gene

- wgMLST is a genotyping scheme that uses WGS data to examine and compare sequence at thousands of loci throughout the genome
- The wgMLST scheme developed by CDC includes 2,690 loci, which covers about 70% of the genome
- A locus is the location in the genome and in this case, each locus is an individual gene in the genome

Whole-genome multilocus sequence typing (wgMLST)

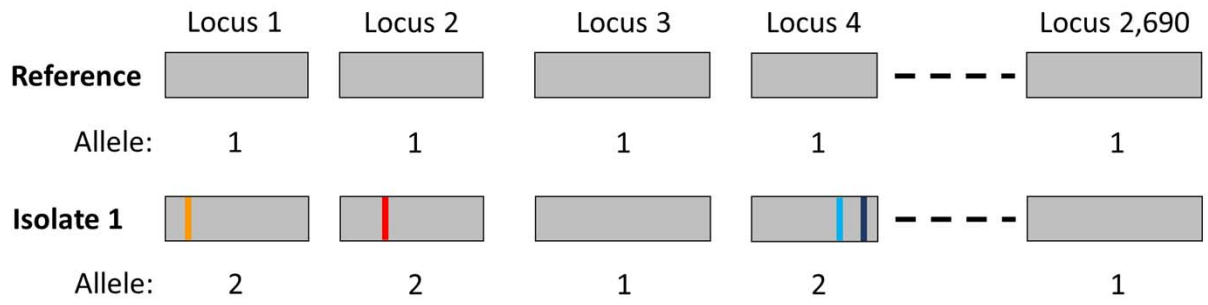


Locus: location in the genome; in this case, each locus is an individual gene

Allele: variant form of a gene

- The sequence at each of the 2,690 loci is examined and assigned an allele number
- An allele is a variant form of a gene
- A reference sequence for each locus is stored in a curator database and is identified as allele 1

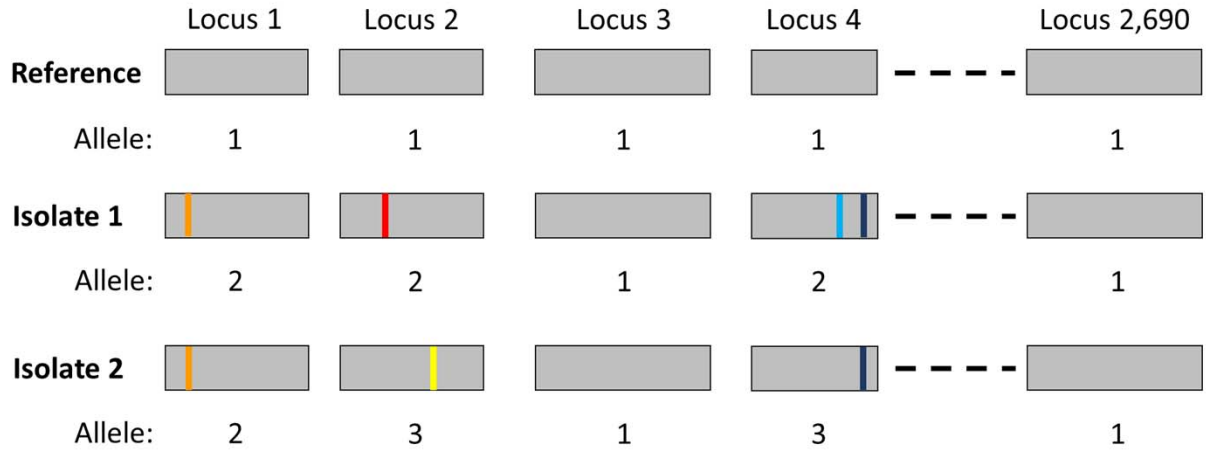
Whole-genome multilocus sequence typing (wgMLST)



Colored lines indicate a single nucleotide difference compared to the reference

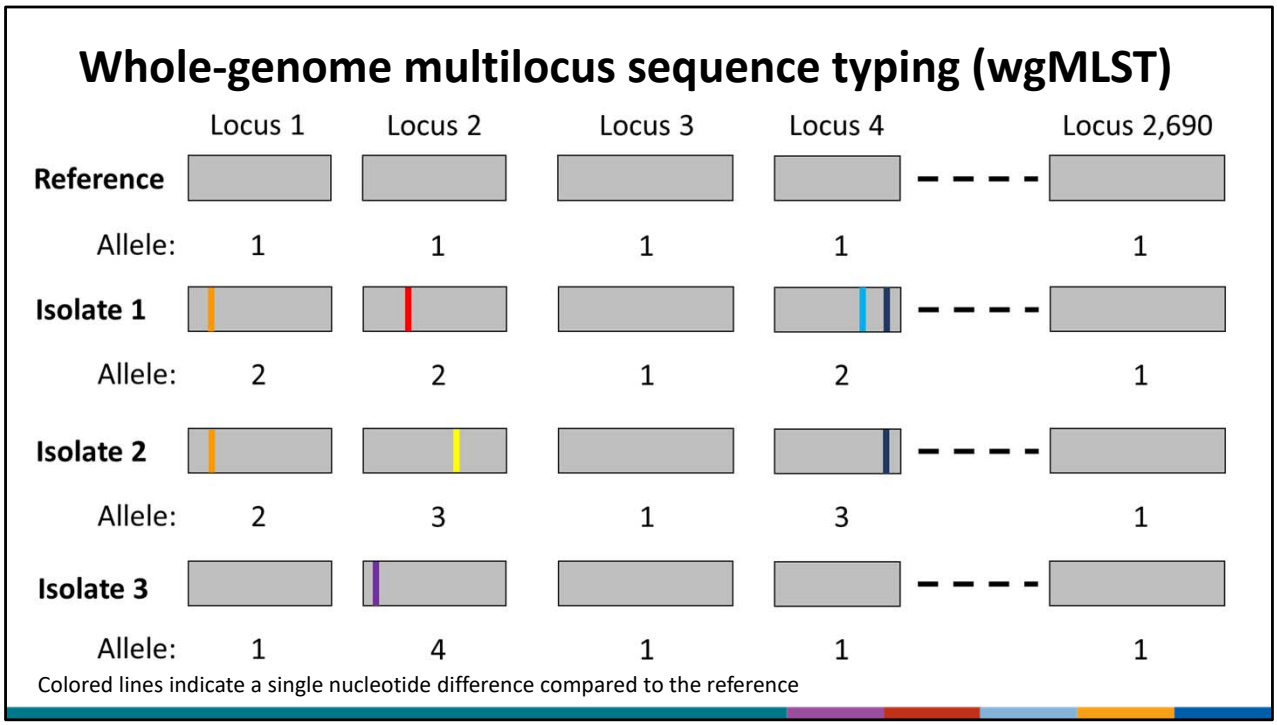
- Isolates that have the same sequence at a given locus are assigned the same allele number for that locus
- But if the sequence is different, then it is assigned a new allele number
- Here we have Isolate 1
- The colored lines are indicating a single nucleotide difference in comparison to the reference sequence
- At Locus 1, Isolate 1 has a different sequence than the reference because there is one nucleotide that is different
- So Isolate 1 is assigned as having Allele 2 at Locus 1
- Isolate 1 also has a different sequence at Locus 2 so it is assigned as having Allele 2 at Locus 2
- At Locus 3, it has sequence that matches to the Reference so it is assigned as Allele 1 at Locus 3
- At Locus 4, it differs from the Reference at two nucleotide positions, but wgMLST does not quantify how many differences there are at a particular locus, it only considers whether they are different or not
- So Isolate 1 gets assigned as having Allele 2 at Locus 4
- This process then gets repeated at each of the 2,690 loci

Whole-genome multilocus sequence typing (wgMLST)

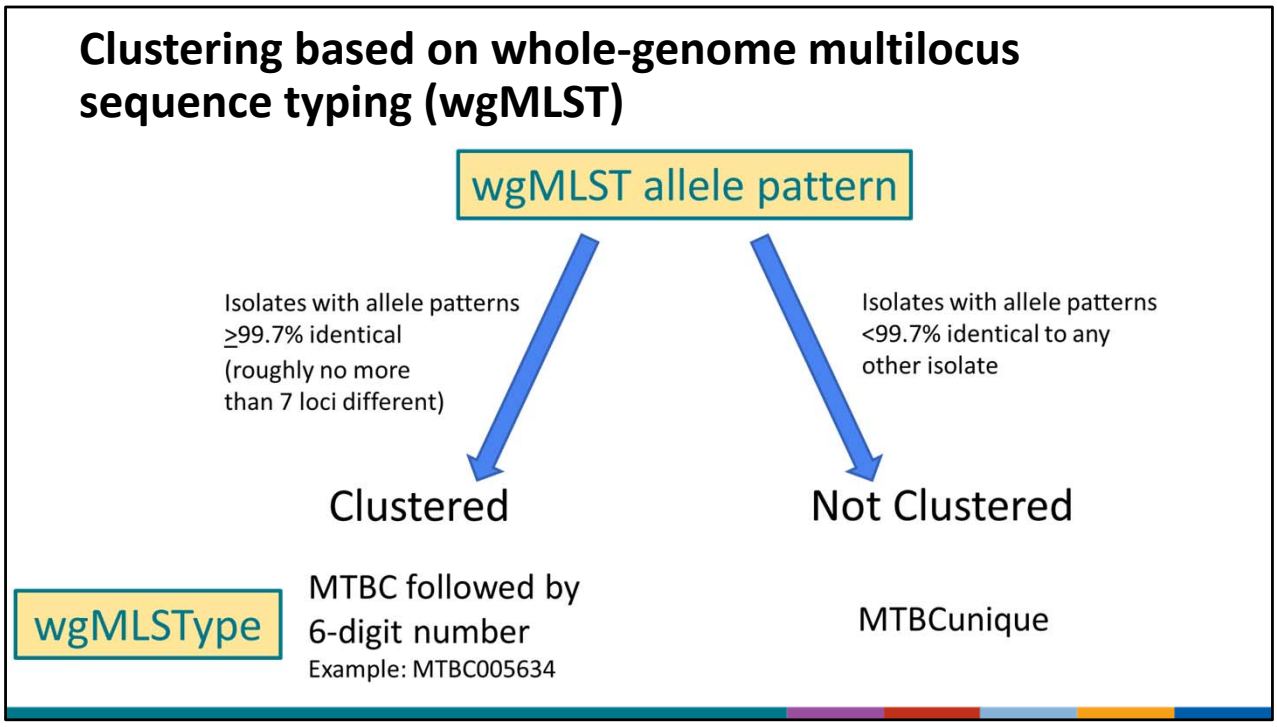


Colored lines indicate a single nucleotide difference compared to the reference

- Now we are going to add sequence from another isolate, Isolate 2, to the database
- Isolate 2 has the same sequence at Locus 1 as Isolate 1 so it gets assigned the same allele number, Allele 2 at Locus 1
- At Locus 2, Isolate 2's sequence does not match the Reference or Isolate 1 so it gets assigned a new allele number, Allele 3 at Locus 2
- At Locus 3, it has the same sequence as the Reference and Isolate 1 so it gets assigned as having Allele 1 at Locus 3
- At Locus 4, it does not exactly match with either the Reference or Isolate 1
- Although Isolate 2 has the same nucleotide difference shown in dark blue as Isolate 1, it does not have the nucleotide difference shown in light blue
- Since it does not match exactly, it gets assigned a new allele number, Allele 3 at Locus 4



- And now we will add another isolate, Isolate 3, to the database
- At Locus 1, this isolate's sequence matches to the Reference so it is assigned Allele 1 at Locus 1
- At Locus 2, it does not match to any of the sequences for this locus that are already stored in the database so it gets assigned a new allele number, Allele 4 at Locus 2
- And at Locus 3 and 4, it matches to the Reference so it is assigned as having Allele 1 at Locus 3 and Allele 1 at Locus 4
- So each isolate then ends up with a wgMLST allele pattern, which is a code made up of the assigned allele numbers for all 2,690 loci

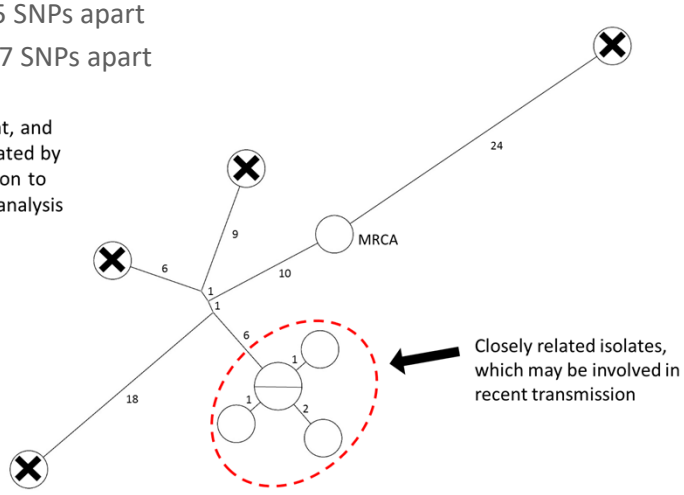


- The wgMLST allele pattern is then used for defining clusters
- Unlike with conventional genotyping, isolates do not need to match at all 2,690 loci to be clustered
- Isolates that match for over 99.7% of the loci get grouped together as a cluster
- This is roughly no more than 7 loci different
- These clusters are assigned a name, called a wgMLSType
- The wgMLSType is formatted as MTBC followed by a 6-digit number
- Unlike with GENType, with wgMLSType, we only assign a number to it if it's a cluster (2 or more isolates nationally that are more than 99.7% identical). Otherwise, it is designated as MTBCunique
- However, it should be noted that if duplicate isolates are submitted for a patient, those two isolates will cluster with each other and be assigned a cluster number

wgMLST clustering threshold informed by experience with wgSNP comparison

- **Isolates from epidemiologically linked cases**
  - Generally within 5 SNPs apart
  - Occasionally 6 or 7 SNPs apart

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



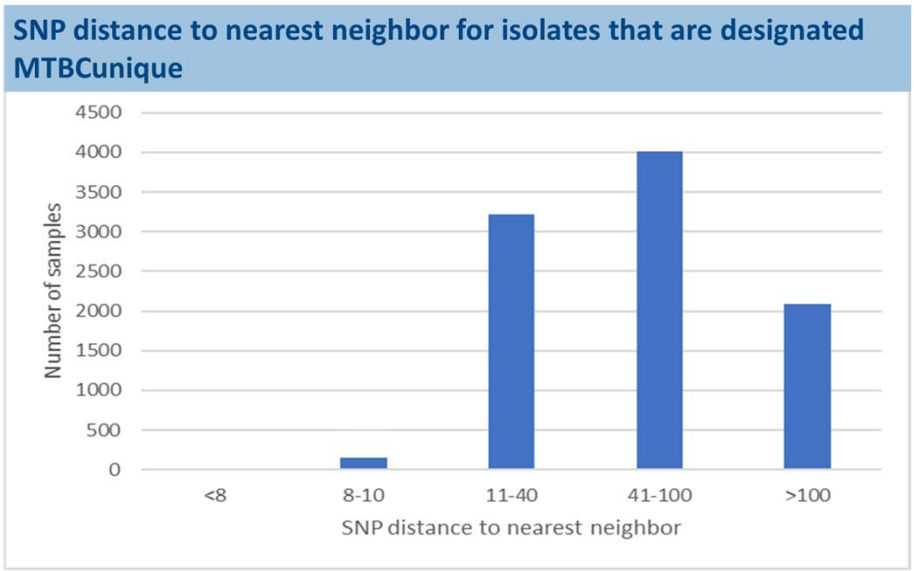
- This threshold for clustering was based on our experience with wgSNP comparison and that isolates from cases that are epi linked are generally within 5 SNPs apart from each other, but we do occasionally see isolates from epi linked cases that are 6 or 7 SNPs apart
- Because wgMLST is being used as a first pass to detect clusters that may represent recent transmission, we wanted to cast a wide net to make sure we were being inclusive of all cases that could be due to recent transmission

wgMLST clustering threshold informed by experience with wgSNP comparison

- **Want to make sure isolates within 7 SNPs apart are clustered**
- **Set threshold for clustering to be up to 7 loci different**
  - 2,690 loci so threshold would be 99.7%

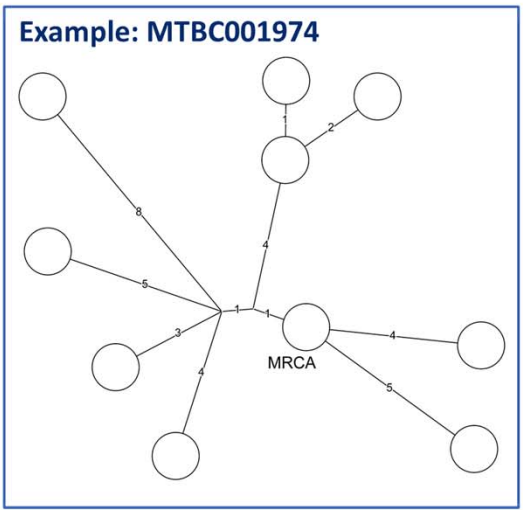
- To make sure that we were capturing all isolates that were within 7 SNPs of another isolate, we set the threshold for clustering to be up to 7 wgMLST loci different
- Since there are 2,690 wgMLST loci, this works out to be a threshold of 99.7% identity between wgMLST allele patterns

wgMLSType cluster threshold captures isolates within 7 SNPs

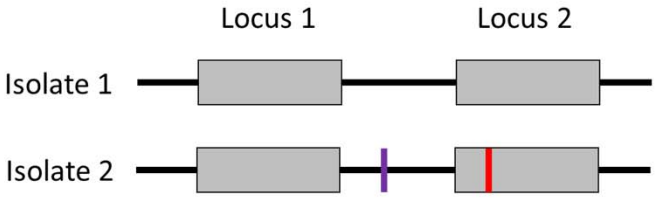


- To ensure that clustering using a threshold of 99.7% identity between wgMLST allele patterns was not missing any isolates that are within 7 SNPs, we analyzed all the isolates designated as MTBCunique to identify their nearest neighbor (the isolate that is closest to it) and measured the SNP distance between them
- You can see on this graph that most of the isolates that were designated as MTBCunique were more than 10 SNPs away from their next nearest neighbor and no MTBCunique isolates were within 7 SNPs of their nearest neighbor
- Because a wgMLSType cluster includes all isolates that are considered closely related enough to be consistent with recent transmission, it will not be necessary to also search for isolates that are single locus variants (SLVs) or have a mixed or missing locus (MMLs) like we do with GENType

wgMLSType cluster threshold may also capture isolates greater than 7 SNPs apart



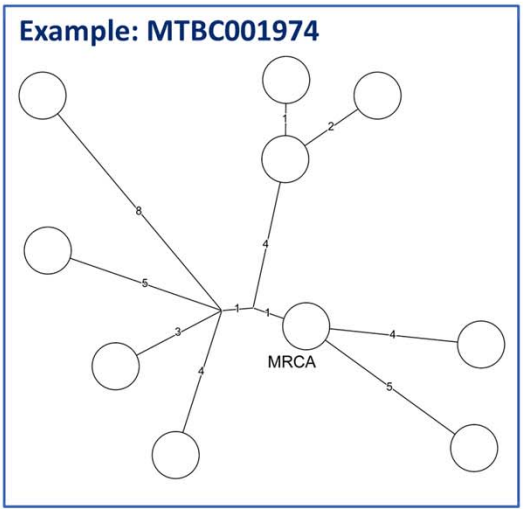
wgMLST: covers ~70% of genome
wgSNP: covers ~90% of genome



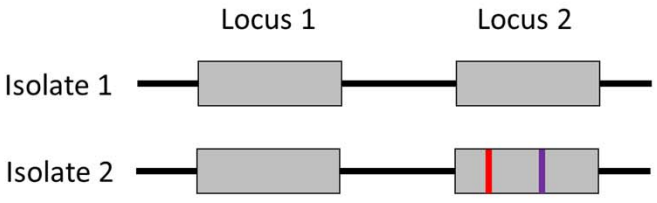
wgMLST: 1 allele difference
wgSNP: 2 SNP differences

- However, wgMLSType clusters can also include isolates that are more than 7 SNPs from their nearest neighbor
- This is because wgMLST covers about 70% of the Mtb genome and wgSNP covers about 90% so there could be SNPs identified by wgSNP comparison that are outside of the regions of the genome included in the wgMLST scheme

wgMLSType cluster threshold may also capture isolates greater than 7 SNPs apart



wgMLST: counts allele differences
wgSNP: counts SNPs individually



wgMLST: 1 allele difference
wgSNP: 2 SNP differences

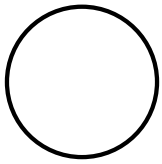
- In addition, if there were more than one SNP in a single gene, each SNP would be counted individually by wgSNP comparison, but would just be counted as one allele difference by wgMLST

Isolates designated as MTBCunique can later become clustered

Example:

Isolate X

Sequenced September 2021



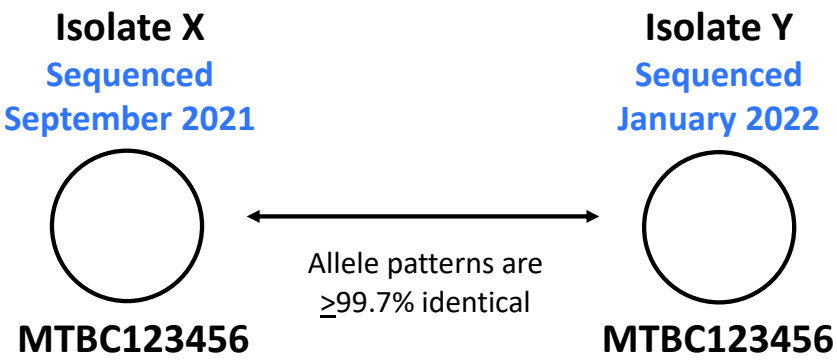
MTBCunique

(<99.7% identical to any other isolate in the database *at the time*)

- Because wgMLSType is designated based on whether the isolate clusters with another isolate in the database at the time, this means that the wgMLSType for an isolate can later change as more isolates are added to the database
- The most common scenario is an isolate that is designated as MTBCunique later becoming clustered
- For example, we have this isolate, Isolate X, that is sequenced in September 2021
- Its wgMLST allele pattern is less than 99.7% identical to any other isolate in the database at that time so it is designated as MTBCunique

Isolates designated as MTBCunique can later become clustered

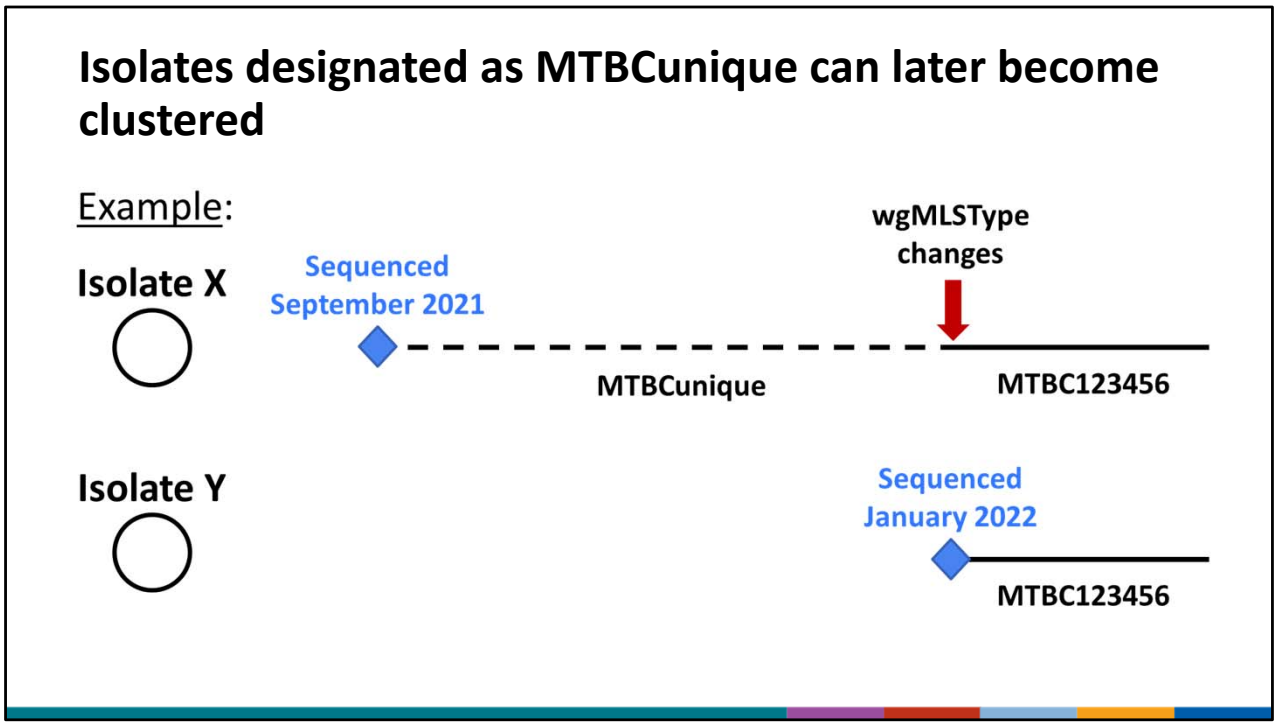
Example:



Isolate X's wgMLSType changes from MTBCunique to MTBC123456 in January 2022

- But then later, in January 2022, Isolate Y gets submitted for sequencing and its wgMLST allele pattern is over 99.7% identical with Isolate X
- Since these two isolates are clustered, they both get assigned a numbered wgMLSType, in this case MTBC123456
- So Isolate X's wgMLSType would change from MTBCunique to MTBC123456 in January 2022

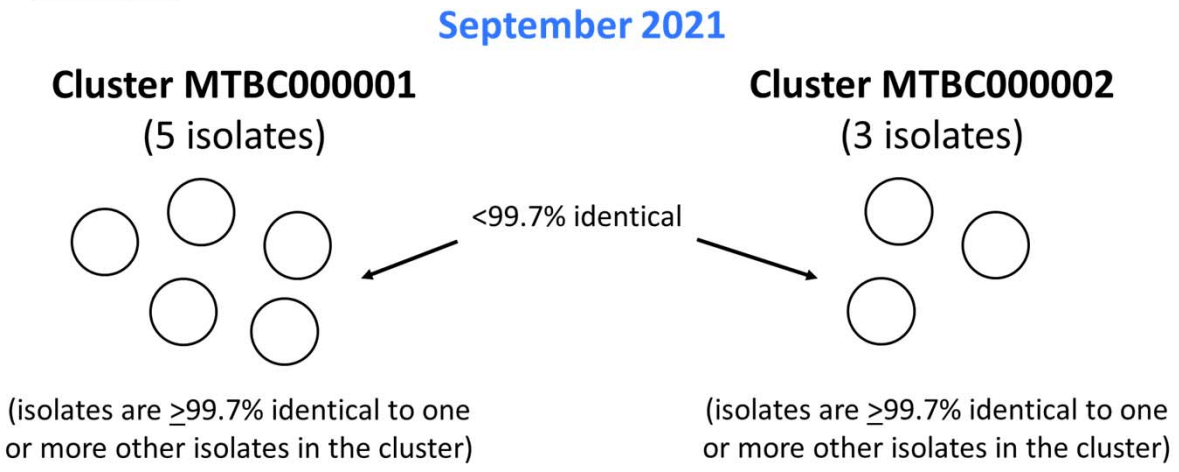
Isolates designated as MTBCunique can later become clustered



- So Isolate X's wgMLSType would have been listed as MTBCunique starting in September 2021 until January 2022 when it was updated to MTBC123456 because it clustered with a newly added isolate

Two wgMLSType clusters can combine into one cluster

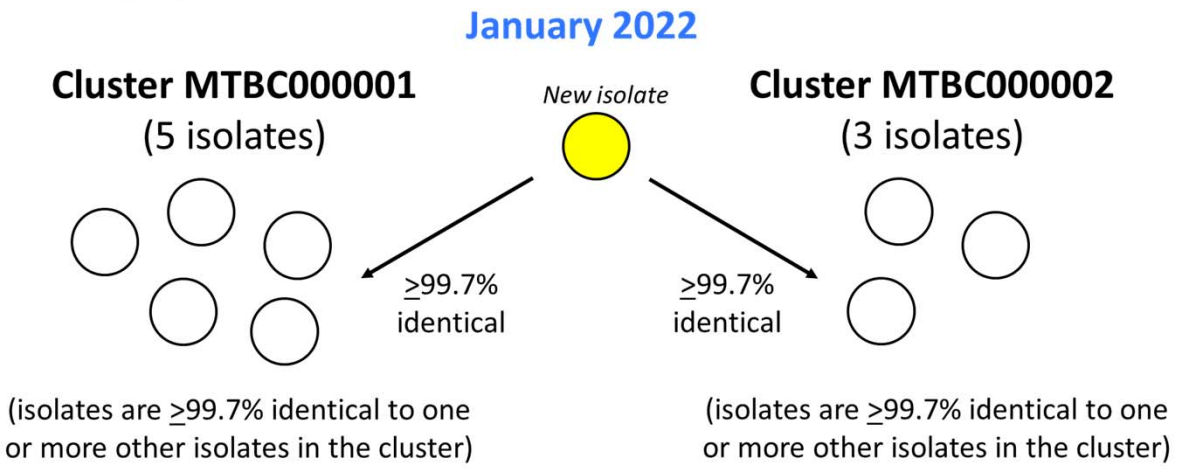
Example:



- A less common scenario is that two wgMLSType clusters can combine into one cluster
- In this example, we have two separate clusters in September 2021, MTBC000001 and MTBC000002
- All five isolates in cluster MTBC000001 are less than 99.7% identical to the three isolates in cluster MTBC000002 and vice versa

Two wgMLSType clusters can combine into one cluster

Example:



- But the addition of a new isolate could bridge the gap between these two clusters
- In this example, a new isolate shown in yellow is submitted and sequenced in January 2022 and the wgMLST allele pattern of this new isolate is over 99.7% identical with at least one of the isolates in cluster MTBC000001 and at least one of the isolates in cluster MTBC000002

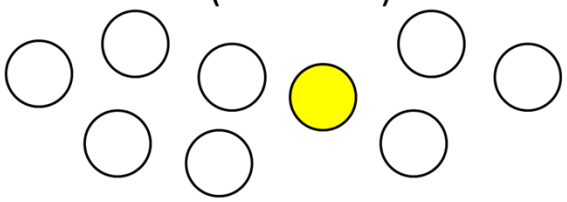
Two wgMLSType clusters can combine into one cluster

Example:

January 2022

Cluster MTBC000001

(9 isolates)



(isolates are ≥99.7% identical to one or more other isolates in the cluster)

Isolates previously designated as MTBC000002 change to MTBC000001 in January 2022

- So addition of this new isolate causes the two clusters MTBC000001 and MTBC000002 to combine into one cluster
- The combined cluster will be named MTBC000001 because that was the cluster with the greatest number of isolates at the time
- This means that the three isolates that were previously designated as MTBC000002 will change to MTBC000001 in January 2022
- These changes in cluster name will be rare

Summary

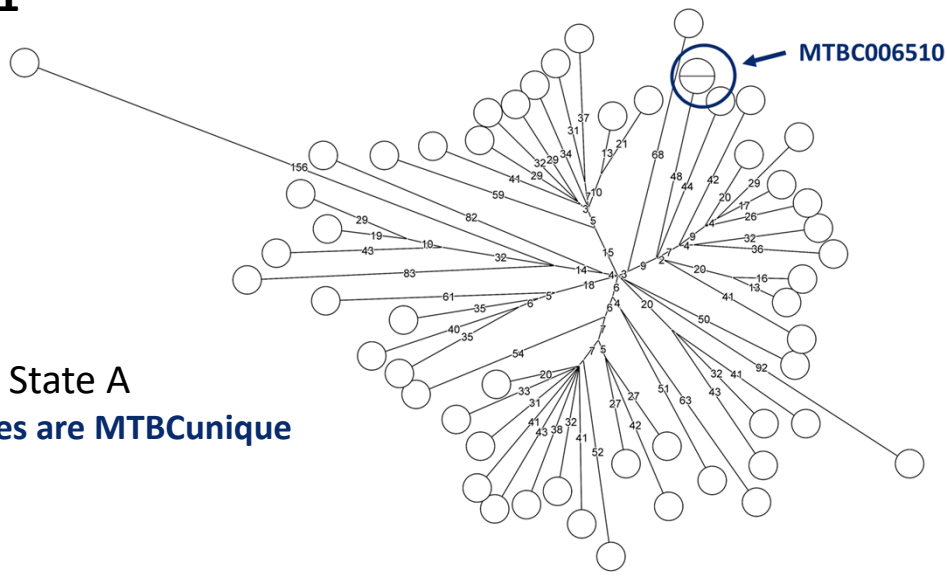
- **wgMLST is a genotyping scheme that uses WGS data to compare sequence at thousands of loci throughout the genome**
  - Isolates that match ≥ 99.7% of loci with another isolate will be clustered
- **Isolates are assigned a wgMLSType**
  - For clustered isolates, this is MTBC followed by a 6-digit cluster number
  - For nonclustered isolates, this is MTBCunique
- **Isolates in a wgMLSType cluster are not necessarily closely related enough to be consistent with recent transmission**
- **The wgMLSType of an isolate can change as new isolates are added to the database**
  - This will usually be going from MTBCunique to a cluster name

- To summarize, wgMLST is a genotyping scheme that uses WGS data to compare sequence at thousands of loci throughout the genome
- Isolates that match over 99.7% of the loci with another isolate will be clustered
- Isolates are assigned a wgMLSType
- For clustered isolates, this is MTBC followed by a 6-digit cluster number
- For non-clustered isolates, this is MTBCunique
- Isolates in a wgMLSType cluster are not necessarily closely related enough to be consistent with recent transmission
- And the wgMLSType of an isolate can change as new isolates are added to the database, but this will usually be going from MTBCunique to a cluster name

How clustering with wgMLSType compares to clustering with GENType

- Next I will present how clustering with wgMLSType compares to clustering with GENType based on some of our preliminary data

Example 1

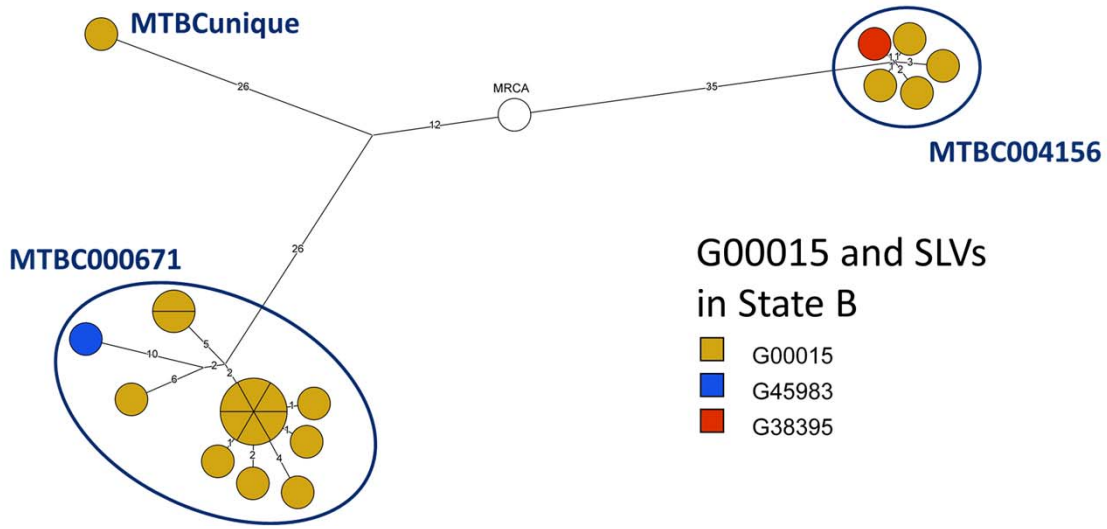


G00016 in State A

Most isolates are MTBCunique

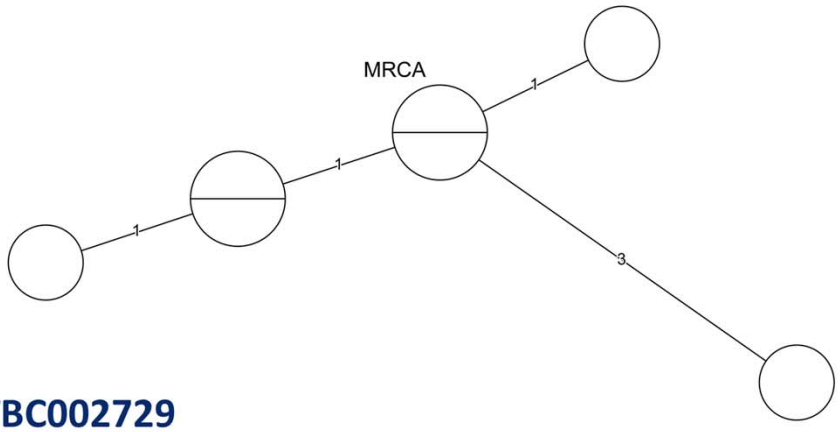
- Now we can revisit our examples from before and look at how isolates in these GENTypes would be clustered by wgMLSType
- We saw that isolates in G00016 were quite distant from each other based on wgSNP comparison so most of these get designated as MTBCunique
- Except for these two isolates up at the top that are 0 SNPs apart, these two form a cluster that was assigned the wgMLSType MTBC006510

Example 2



- When we look at the wgSNP comparison for G00015, we see that isolates in this GENType were broken up into two subclusters of closely related isolates
- Isolates in these two subclusters formed two different wgMLSType clusters: MTBC000671 and MTBC004156
- Notice here that SLVs of G00015, G45983 and G38395, get grouped in with G00015 isolates in the wgMLSType clusters
- Also notice that here we see in wgMLSType cluster MTBC000671, over on the left, an example of isolates that are further than 7 SNPs from their next nearest neighbor but are still clustered by wgMLST

Example 3



G08090 in State C

All isolates are MTBC002729

- And then when we look at G08090 where all the isolates were closely related to each other based on wgSNP comparison, we see that these isolates are all in the same wgMLSType MTBC002729 as well

Isolate diversity within GENType and wgMLSType clusters

Cluster Name	Average SNP distance between all pairs
G00010 (n=114)	21
G00012 (n=115)	150*
G00016 (n=121)	117
G00017 (n=167)	38
G05056 (n=85)	133*
MTBC000151 (n=146)	13
MTBC000025 (n=110)	14
MTBC000164 (n=112)	7
MTBC000145 (n=120)	11
MTBC000013 (n=98)	10

*These averages are underestimates

- To look at how wgMLST affects isolate diversity within a cluster, we compared the average SNP distance between all pairs of isolates in the top five most common GENType clusters and the top five most common wgMLSType clusters
- And we see that the average SNP distance is greatly reduced for the wgMLSType clusters
- Because wgMLST is able to capture much more of the genetic differences throughout the Mtb genome, it provides increased molecular resolution for defining clusters and we see less diversity among isolates in a cluster

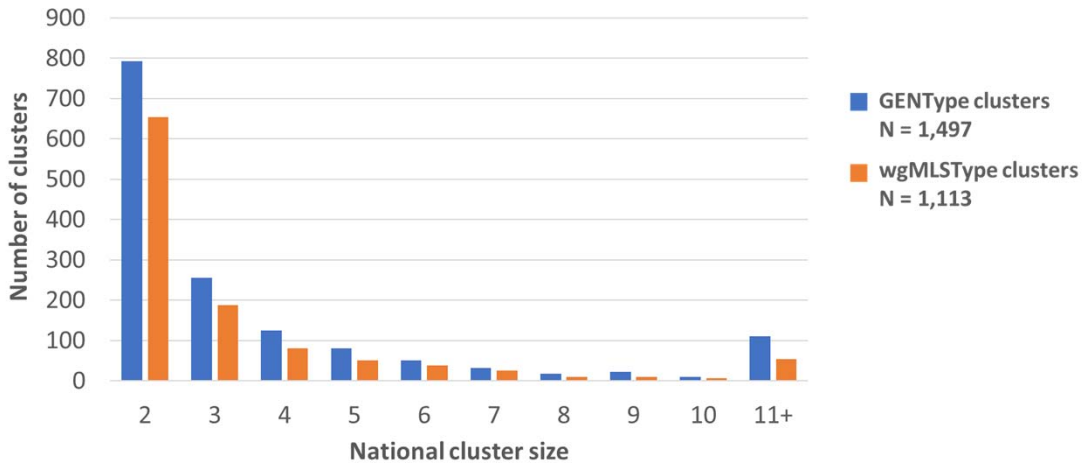
Percent of cases that are clustered

Percent of cases that cluster with another case within geographic boundary

Geographic Boundary	GENType	wgMLSType
National	46.5%	30.7%
County	18.5%	16.8%

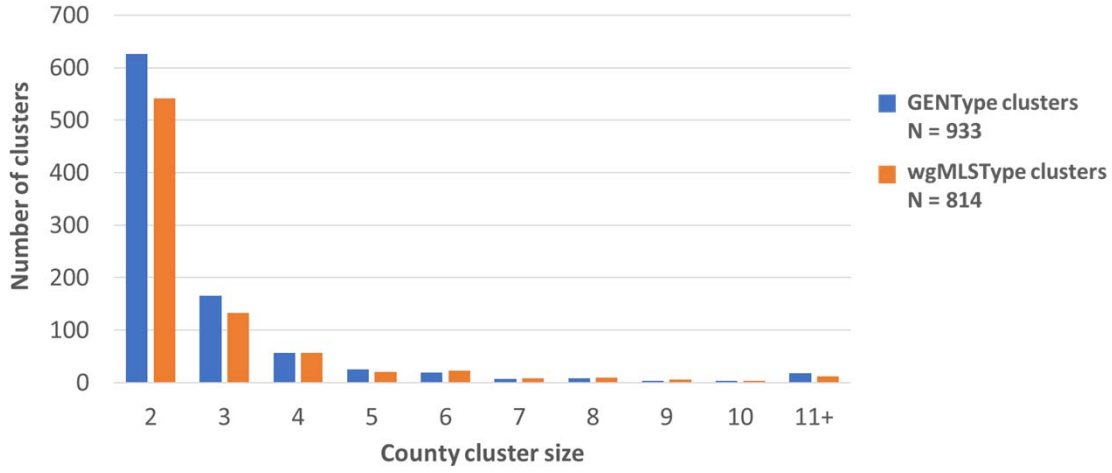
- We also see an overall decrease in the percent of cases that are clustered
- This decrease is more pronounced at the national level, where it goes from 46.5% down to 30.7%, than at the county level, where it goes from 18.5% to 16.8%
- Compared to cases that have the same GENType nationally, cases with the same GENType that are also from the same county are more likely to be due to recent transmission and have closely related Mtb isolates, so those cases are clustered by wgMLST as well

National cluster size distribution for GENType clusters and wgMLSType clusters



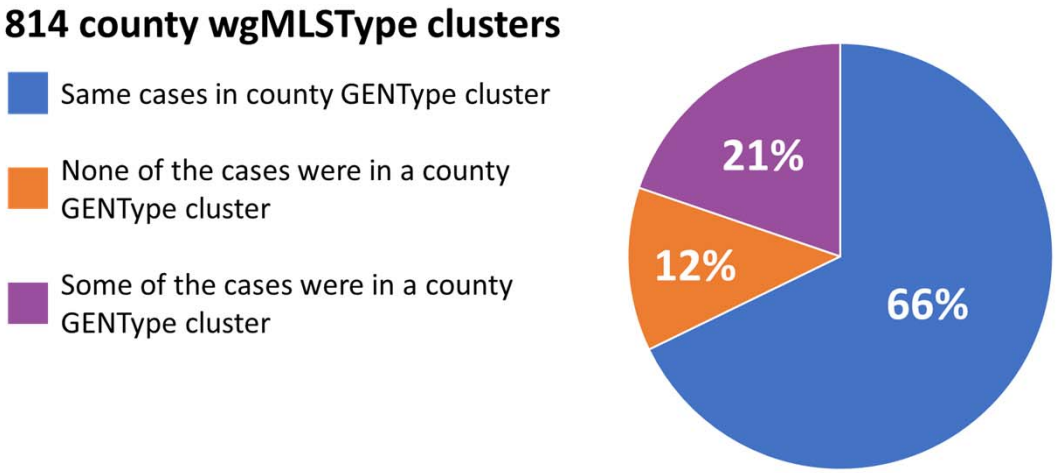
- If we look at the national cluster size distribution for GENType clusters and wgMLSType clusters, we see that there are fewer wgMLSType clusters overall
- This decrease in the number of clusters is seen among all cluster sizes, and particularly for the larger sized clusters with 11 or more cases

County cluster size distribution for GENType clusters and wgMLSType clusters



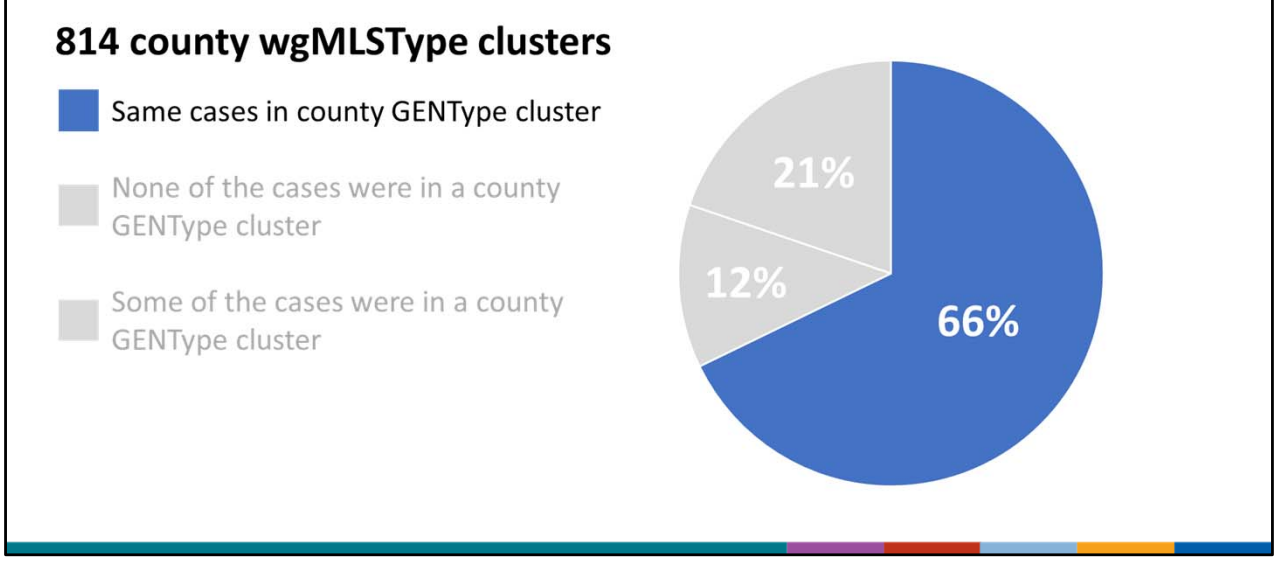
- Looking at the cluster size distribution for county clusters, we also see fewer clusters overall
- There were 933 county GENType clusters and 814 county wgMLSType clusters
- However, with the county clusters, it was mainly the smaller cluster sizes of 2 or 3 that had fewer wgMLSType clusters

Comparison of cases in county wgMLSType clusters and county GENType clusters



- If we then dive deeper into these 814 county wgMLSType clusters and compare which cases are in these clusters compared to which cases are in county GENType clusters, we see that most (66%) of the county wgMLSType clusters are comprised of the same cases as the county GENType cluster
- 12% of the wgMLSType clusters were comprised of cases that were not clustered by GENType
- And 21% of the wgMLSType clusters had some cases that were also clustered by GENType but the clusters were not comprised of the exact same set of cases
- I will give some examples of these different scenarios in the following slides

County wgMLSType clusters comprise same cases as county GENType cluster



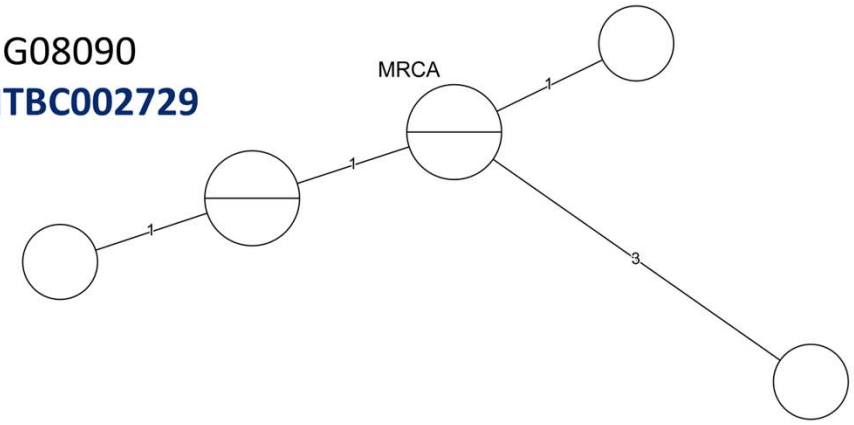
- First, let's look more closely at the county wgMLSType clusters that comprise the same cases as the county GENType cluster

County wgMLSType clusters comprise same cases as county GENType cluster

Example:

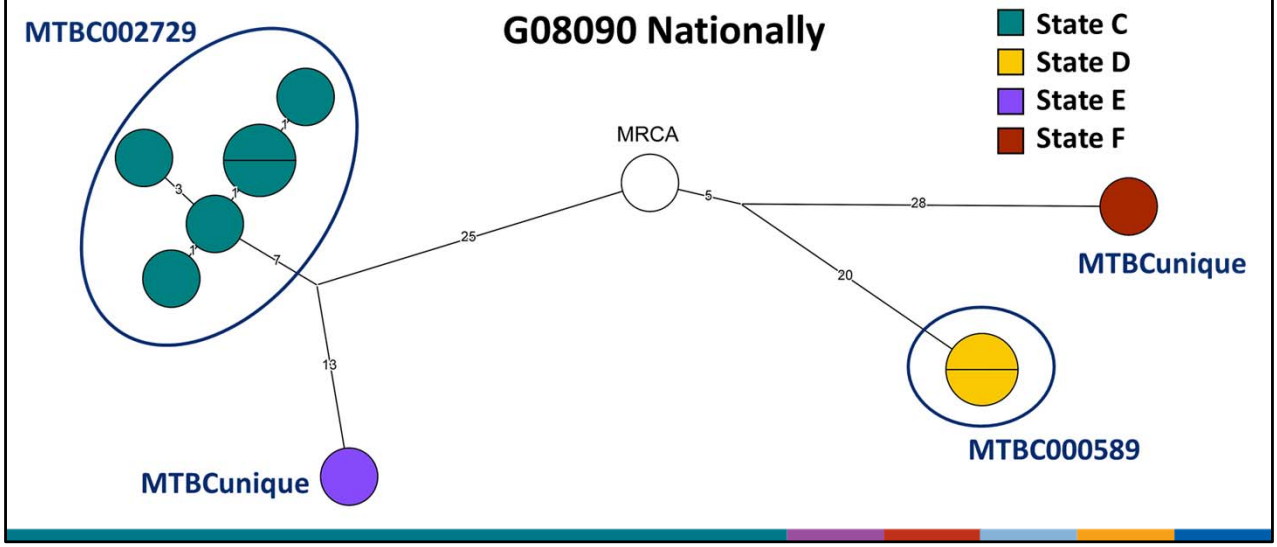
County cluster of G08090

All isolates are MTBC002729



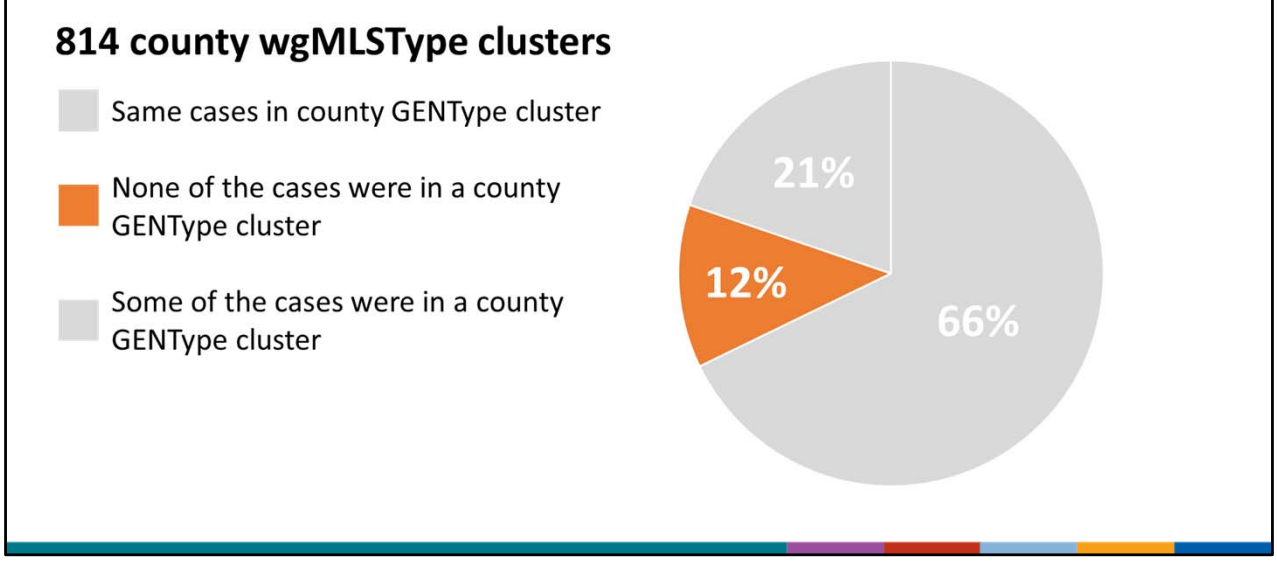
- An example of this would be the cluster of G08090 isolates that I presented earlier
- There were 6 cases in a county that all were GENType G08090
- When wgMLST was applied as the genotyping method for cluster detection, the wgMLSType cluster MTBC002729 comprised the same 6 cases

County wgMLSType clusters comprise same cases as county GENType cluster



- However, there could be changes to which cases share the same genotype nationally with cases in this county cluster
- G08090 is a somewhat common GENType and there are cases outside the county in other states with GENType G08090
- But when wgMLST is applied as the genotyping method, no other cases outside the county have a matching wgMLSType to the cases in the county
- On this tree of all G08090 isolates nationally, you can see that the county wgMLSType cluster MTBC002729 in State C is over on the left in green
- The other G08090 isolates from cases in different states, shown in purple, yellow and red, are all genetically distant so they are not included in wgMLSType MTBC002729

County wgMLSType clusters comprise cases not clustered by GENType



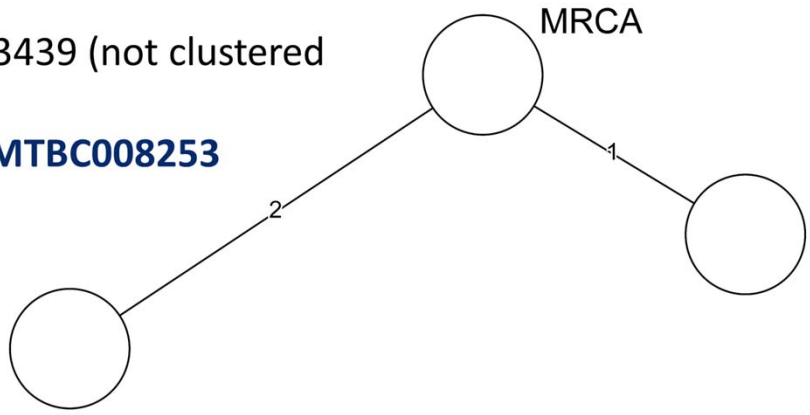
- For the wgMLSType clusters that had none of the cases in a county GENType cluster, these were clusters for which wgMLST clustered isolates that were different GENTypes like SLVs or MMLs
- These were almost entirely 2 case clusters

County wgMLSType clusters comprise cases not clustered by GENType

Example:

G43440 and G43439 (not clustered by GENType)

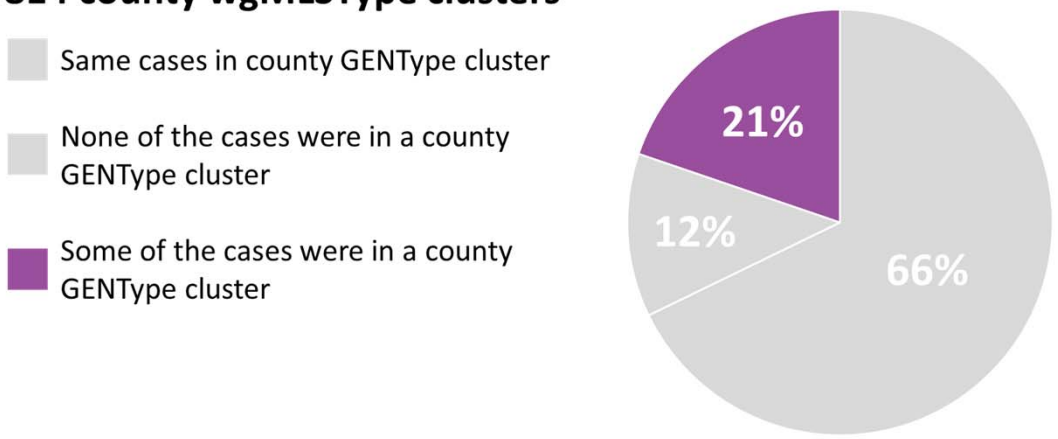
County cluster MTBC008253



- For example, this county wgMLSType cluster is made up of two cases
- One had GENType G43440 and the other had GENType G43439 and no other cases in this county had either of those two GENTypes
- These two GENTypes are SLVs and, in this case, the isolates are closely related
- The isolates are only 3 SNPs apart and so they form a wgMLSType cluster

County wgMLSType clusters comprise some of the same cases as county GENType cluster

814 county wgMLSType clusters



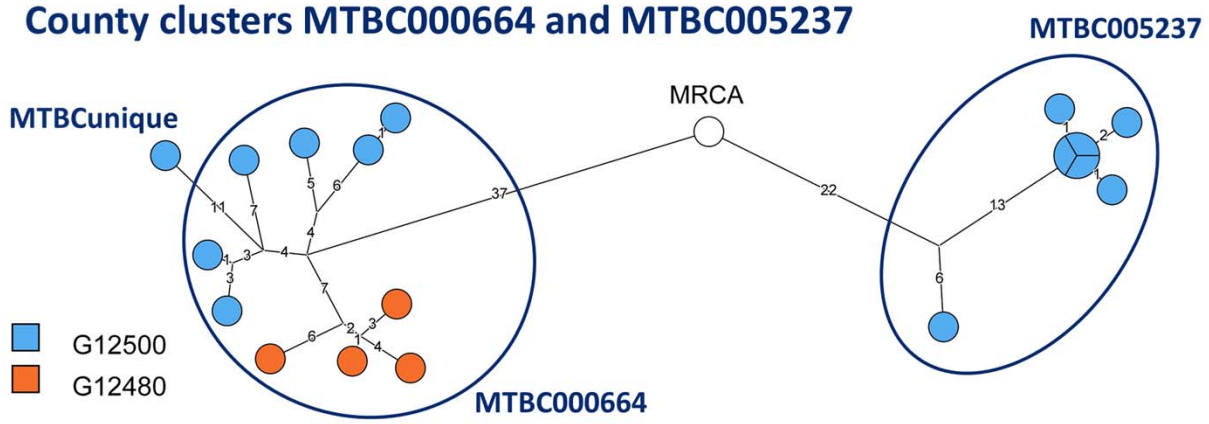
- And the remainder of the county wgMLSType clusters had at least some of the same cases that were in a county GENType cluster, but the cluster did not comprise the exact same set of cases
- This can be due to distant isolates in a GENType cluster being excluded from the wgMLSType cluster, or due to SLVs or MMLs being included in the wgMLSType cluster, or a combination of the two

County wgMLSType clusters comprise some of the same cases as county GENType cluster

Example:

G12500 and G12480 (SLVs)

County clusters MTBC000664 and MTBC005237



- For example, if we look at G12500, the isolates shown in blue in the tree below would all form one county GENType cluster
- And the SLV of G12500, G12480, shown in orange would form a separate county GENType cluster
- But when wgMLST is applied, some of the G12500 isolates are in county wgMLSType cluster MTBC005237, circled on the right side of the slide
- And some of the other G12500 isolates are in a different county wgMLSType cluster MTBC000664 along with the G12480 isolates, circled over here on the left side of the slide

Summary

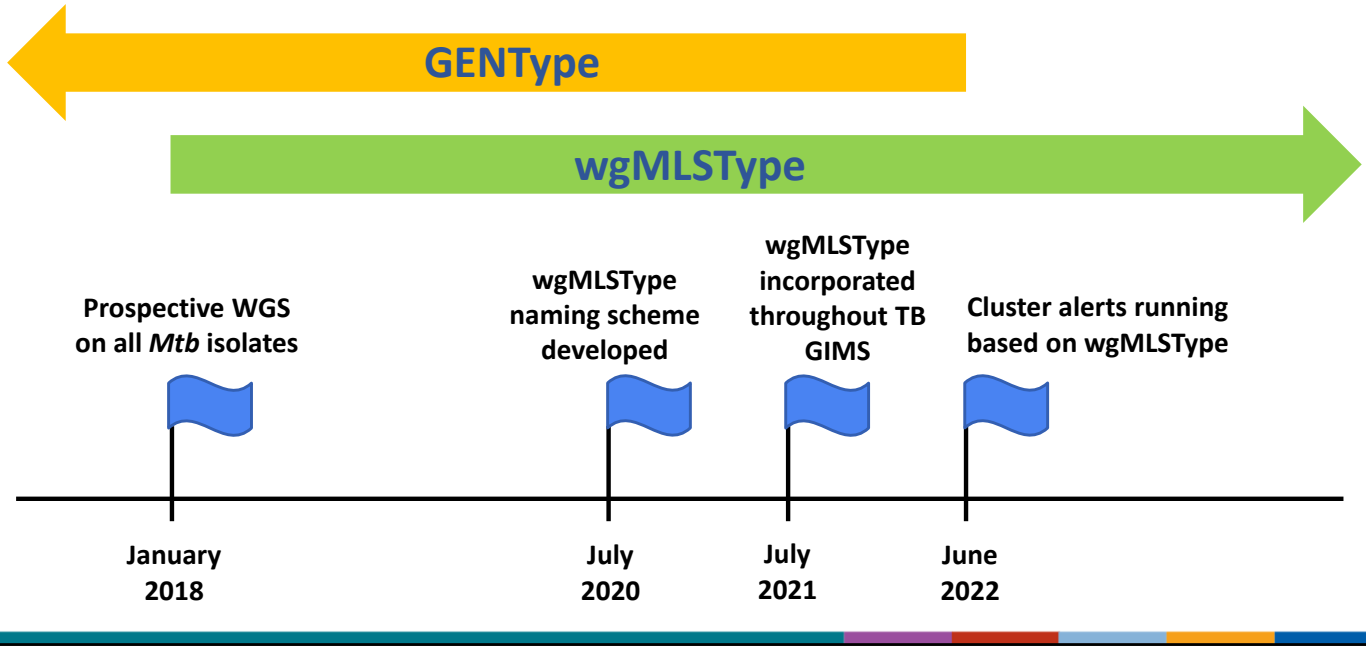
- **Clustering based on wgMLST**
  - Includes closely related SLVs and MMLs and excludes distantly related isolates
  - Decreases the percent of cases that are clustered and number of clusters
  - Greatly decreases isolate diversity within common genotypes
- **The majority of county wgMLSType clusters comprise the same cases as the corresponding county GENType cluster**

- In summary, clustering based on wgMLST includes closely related SLVs and MMLs and excludes distantly related isolates, decreases the percent of cases that are clustered and number of clusters, and greatly decreases isolate diversity within common genotypes
- However, the majority of county wgMLSType clusters comprise the same cases as the corresponding county GENType clusters

Where we are in the transition

- And lastly, I'll give a brief orientation to where we are in this transition to wgMLSType

Timeline for transition to wgMLSType



- Prospective whole-genome sequencing of all Mtb isolates began in 2018, but the conventional genotyping methods (spoligotyping and MIRU-VNTR) that are used to assign an isolate a GENType have continued to be performed
- And we're currently still in this overlap period where an isolate has both conventional genotyping and whole-genome sequencing performed
- In summer of 2020, once there was enough years of whole-genome sequence data available, we were able to finalize the wgMLST methods and the wgMLSType naming scheme
- The next step was to begin reporting out the wgMLSType to state and local TB programs and labs through the TB Genotyping Information Management System (TB GIMS)
- This is a web-based system that combines TB case surveillance data with the corresponding isolate genotyping data
- Aside from housing the data, the system also has a lot of other functionality such as generating various reports and requesting wgSNP analysis for cluster resolution
- In July 2021, wgMSLType was incorporated throughout the system for all users
- The wgMLSType designations going back to isolates that were submitted in 2018 were back populated and wgMLSType designations for newly submitted isolates are being reported prospectively through TB GIMS
- However, the transition is not yet complete because the cluster alerts that we run each week are still based on GENType

- We have been performing analyses to determine how to best adjust our cluster alerting algorithms to accommodate wgMLSType and plan to begin running those based on wgMLSType within TB GIMS by next summer, at which point conventional genotyping to assign GENType will no longer be performed

Acknowledgments

- **Division of TB Elimination**
  - Jamie Posey
  - Lauren Cowan
  - Steve Kammerer
- **Michigan State Public Health Laboratory**
- **Wadsworth Center**

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.