National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention
Division of Tuberculosis Elimination

A look at the methods behind whole-genome single nucleotide polymorphism (SNP) comparison and phylogenetic analysis for TB

Sarah Talarico, PhD, MPH

Laboratory Branch and Surveillance, Epidemiology, and Outbreak Investigations Branch

- Hello, I'm Sarah Talarico. I'm an epidemiologist with the Laboratory Branch and the Surveillance, Epidemiology, and Outbreak Investigations Branch in the Division of Tuberculosis Elimination at CDC
- I will be presenting the following set of training slides that provide a look at the methods CDC uses to perform whole-genome single nucleotide polymorphism (or SNP) comparison and phylogenetic analysis for TB

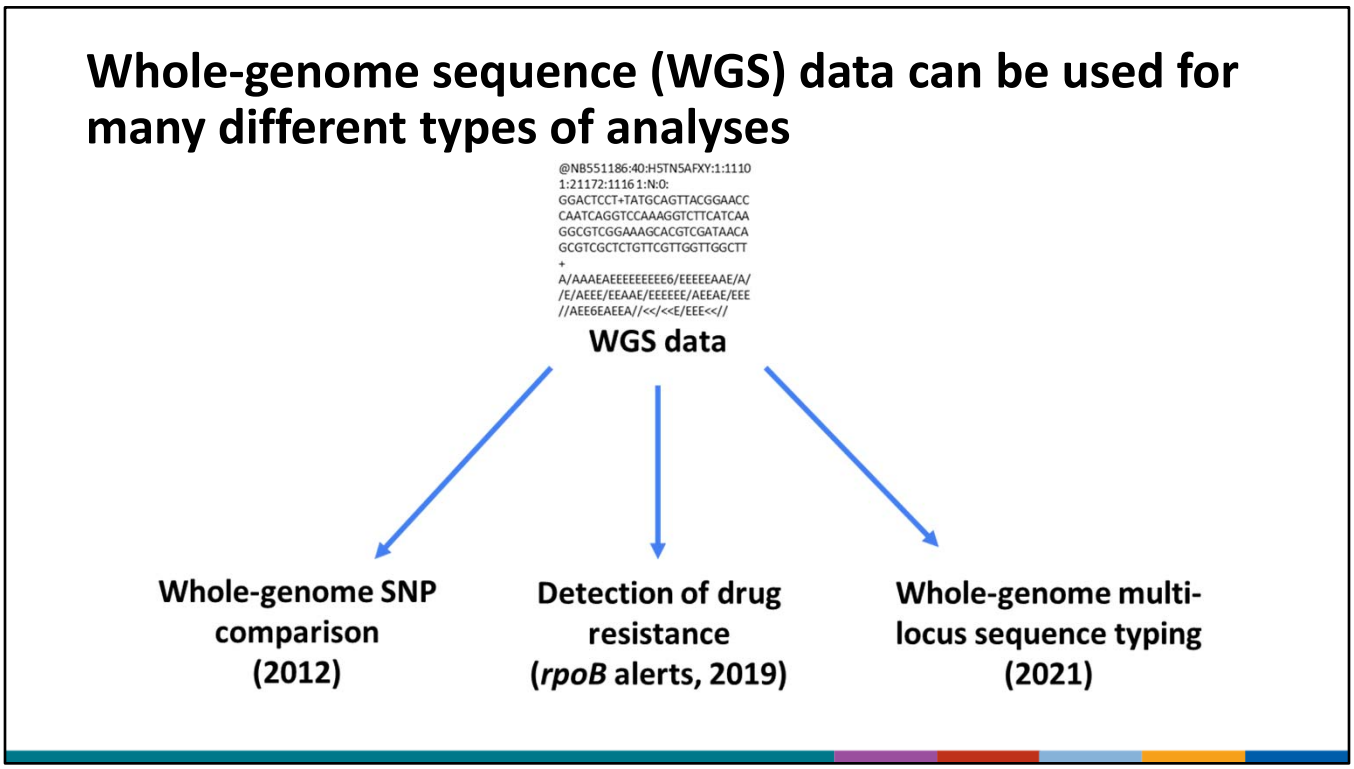
Learning objectives

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

- The analytic steps involved in whole-genome SNP comparison
- How a phylogenetic tree is built using the results of whole-genome SNP comparison
- How the placement of the most recent common ancestor (MRCA) is determined
- How adding or removing isolates from the comparison can affect results

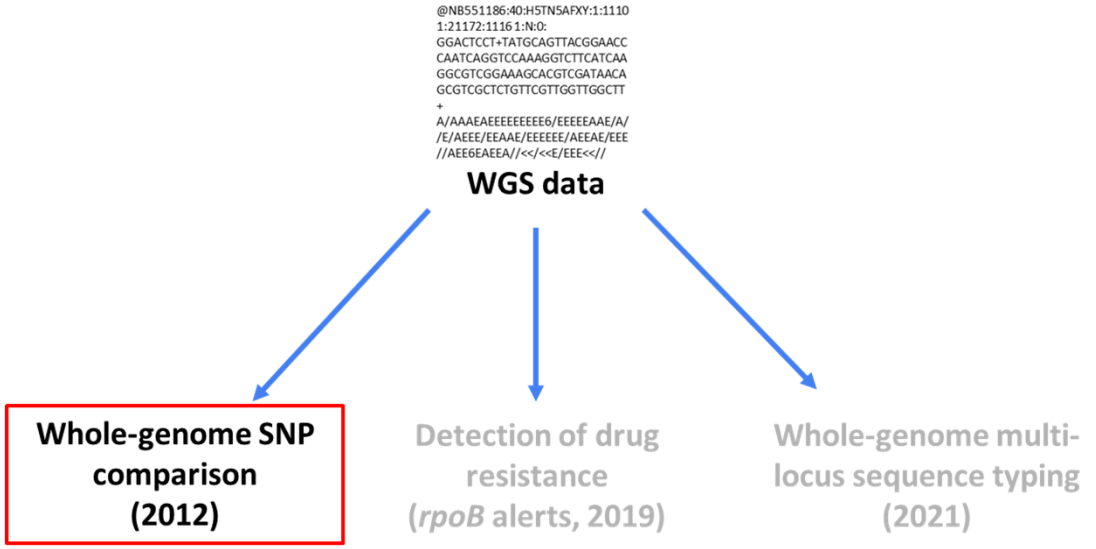
At the end of this presentation, participants will be able to describe:

- The analytic steps involved in whole-genome SNP comparison
- How a phylogenetic tree is built using the results of whole-genome SNP comparison
- How the placement of the most recent common ancestor (MRCA) is determined
- How adding or removing isolates from the comparison can affect results



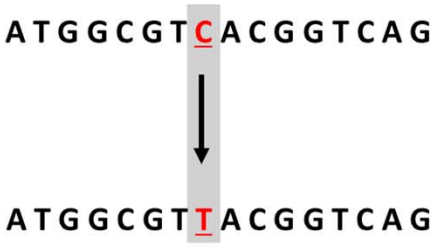
- Whole-genome sequence (or WGS) data can be used for many different types of analyses that serve different purposes
- CDC began using WGS data for whole-genome SNP comparison of isolates in genotype-matched clusters in 2012
- Use of WGS data to detect mutations in the rpoB gene that confer resistance to rifampicin began in 2019
- And starting in 2021, we will begin using the WGS data for whole-genome multi-locus sequence typing, which is a genotyping scheme that will replace GENType for cluster detection and alerting

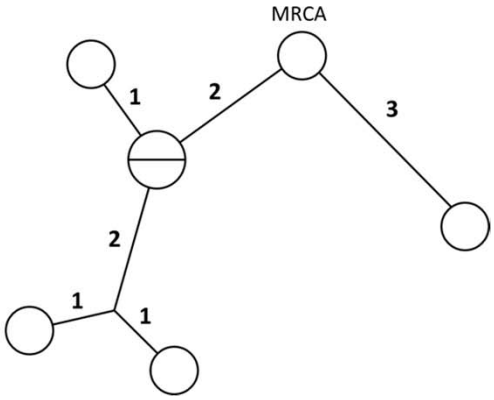
WGS data can be used for many different types of analyses



- This training module will focus on how WGS data is used for whole-genome SNP comparison

Whole-genome SNP comparison





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

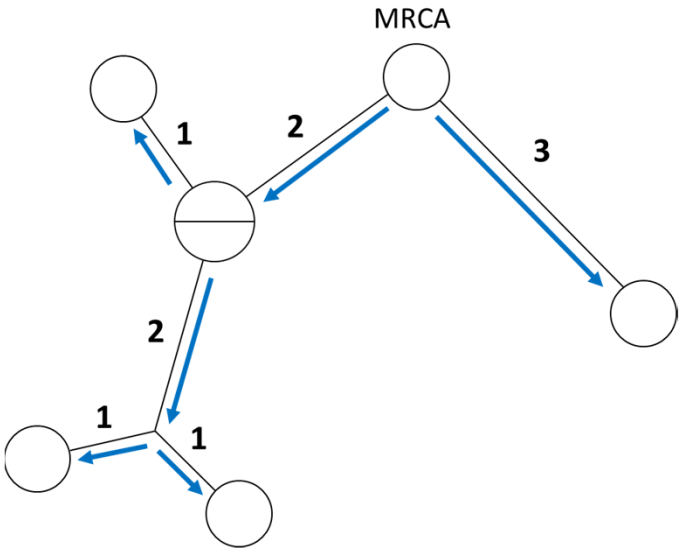
- Whole-genome SNP comparison is performed to identify SNPs that differ between isolates in a genotype-matched cluster
- SNP stands for single nucleotide polymorphism, which is a mutation at a single position in the DNA sequence
- The identified SNPs can then be mapped on to a phylogenetic tree to diagram the genetic relationship among the isolates

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

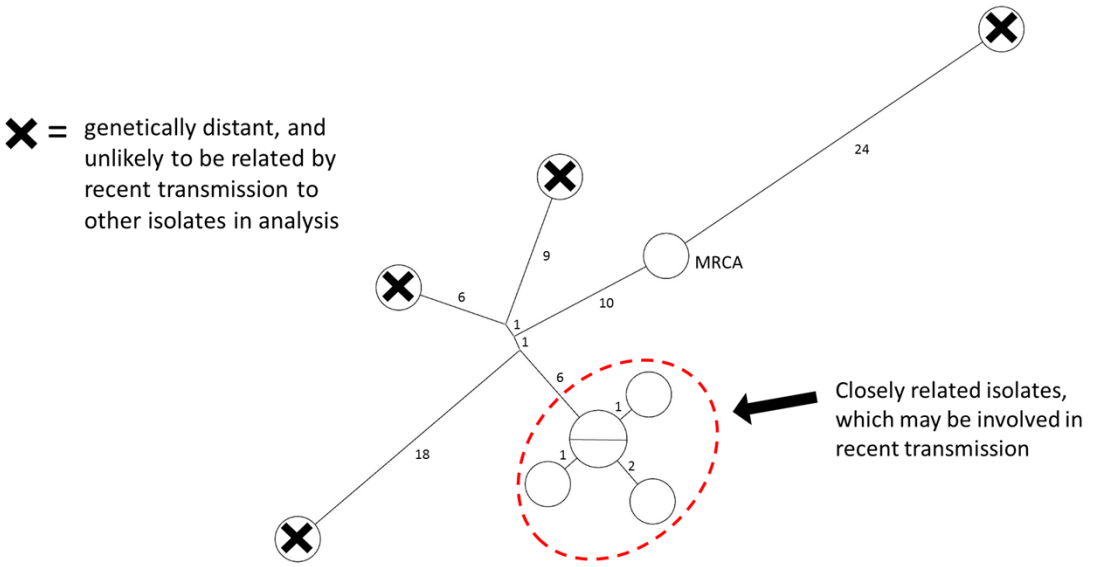
MRCA = Most Recent Common Ancestor

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



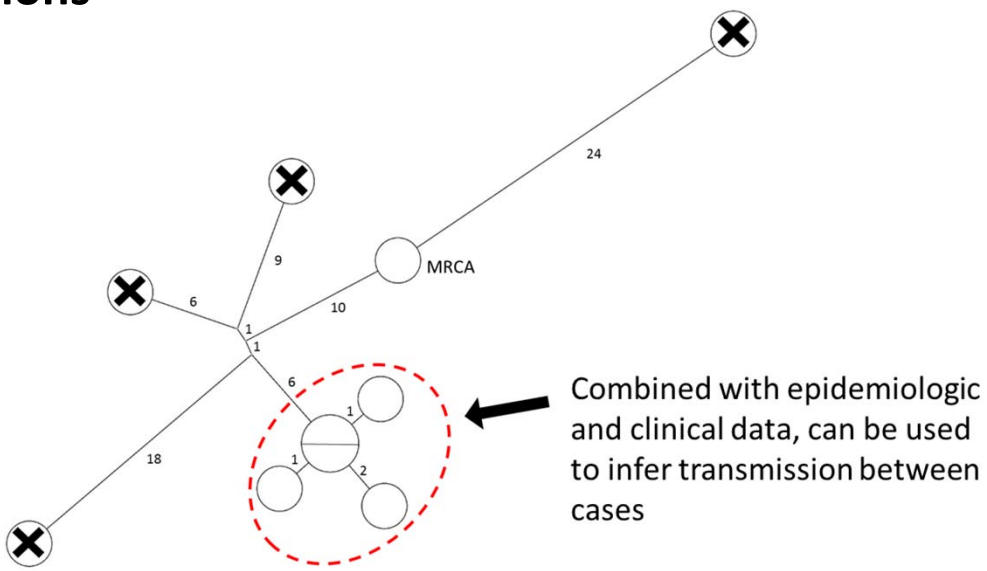
- The results of the whole-genome SNP comparison are delivered to programs in the form of a phylogenetic tree
- The nodes (or circles) represent the isolates and the branches (or lines) that connect the nodes are proportional in length to the number of SNPs that differ between the isolates
- The tree 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 in two ways
- First, they can be used for 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

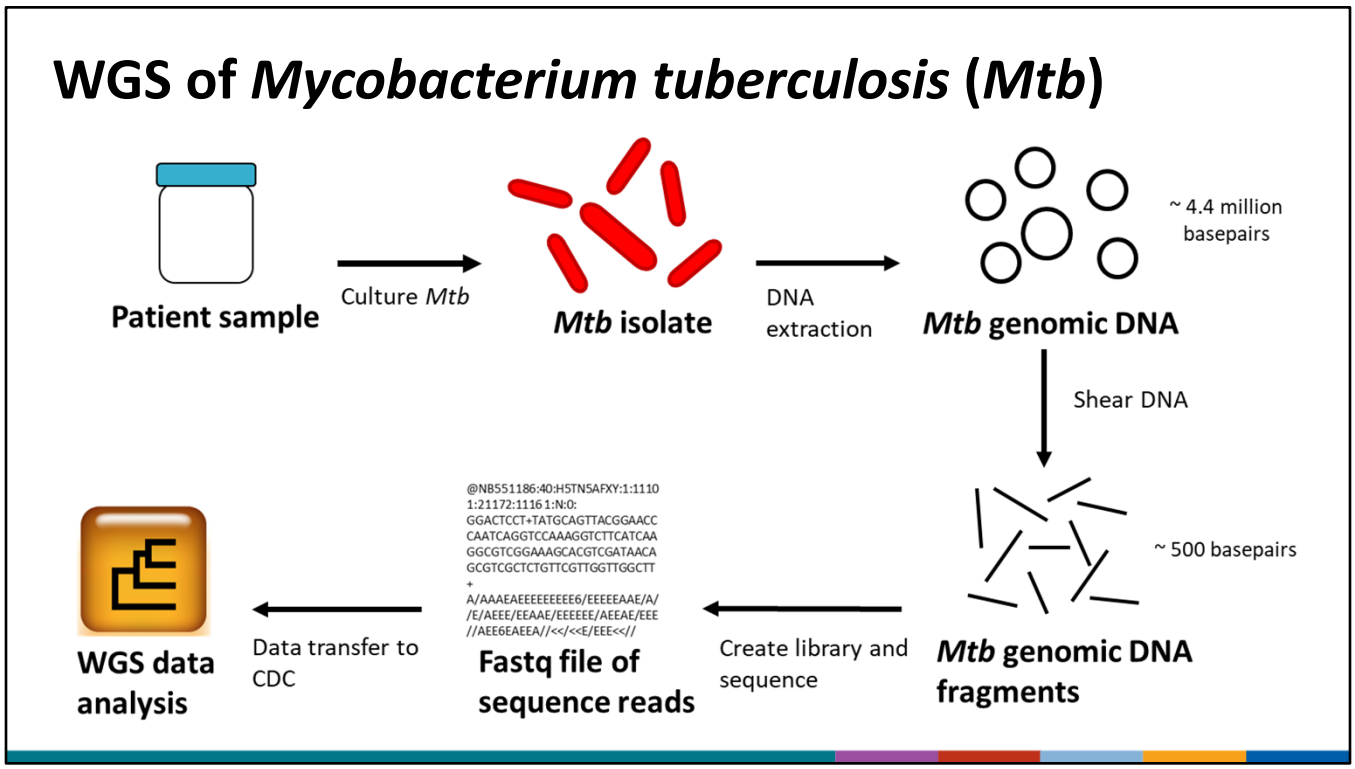
Phylogenetic trees can be used to inform epidemiologic investigations



- Secondly, the genetic relationship among isolates that are closely related can be examined further by looking at SNP distance between isolates, the direction of SNP accumulation based on the MRCA, and the structure of the tree
- This information combined with available epidemiologic and clinical data, such as the timing and infectiousness of the cases and any known epidemiologic links, can be used to make inferences about transmission among cases in a cluster

Whole-genome sequencing and SNP comparison

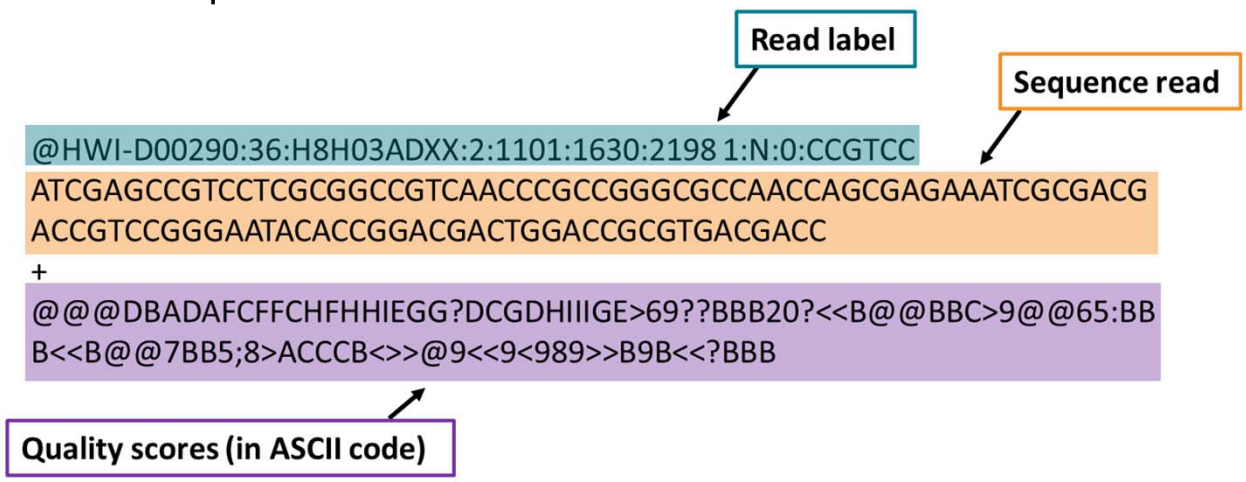
- In the next section, I will present the basics of whole-genome sequencing and details of how whole-genome SNP comparison is performed



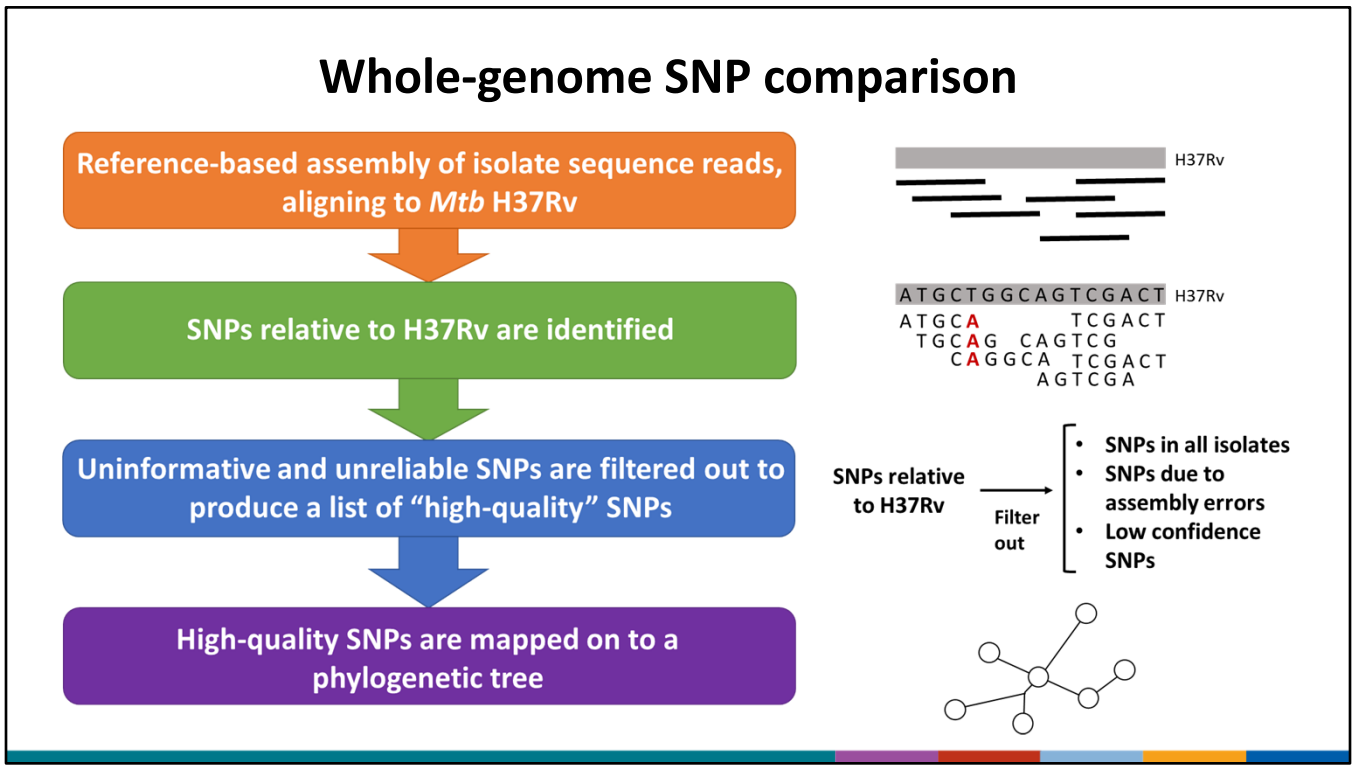
- First I'll start with a big picture overview of how we go from the patient sample to having data that we can analyze
- The process starts with a patient sample, which would usually be a sputum sample, and the sample is cultured for Mycobacterium tuberculosis (or Mtb)
- That yields an Mtb isolate
- The isolate is the population of Mtb that grew from the patient sample so it should approximately reflect any genetic diversity that is present within the patient sample
- Then the genomic DNA is extracted from the Mtb isolate
- The size of the Mtb genome is about 4.4 million basepairs
- The genomic DNA gets sheared to break up the genome into smaller fragments that are around 500 bp long
- A library is created from these DNA fragments, which involves adding special adapters to the ends of the fragments, and the library is sequenced
- The sequence data from the DNA fragments are called sequence reads and they are stored in the form of a fastq file
- I'll explain the data format for a fastq file in the next slide
- These fastq files are what gets transferred over to CDC from the public health laboratory that is doing the sequencing
- The WGS data is then analyzed at CDC using a software called BioNumerics
- The rest of this training module will focus on the details of this last step: the WGS data analysis and specifically the whole-genome SNP comparison

Fastq file: what's that?

A text file with sequence reads and quality scores for each base call in the sequence read



- The fastq file contains the WGS data that gets transferred to CDC for analysis
- A fastq file is a text file with sequence reads and quality scores for each base call in the read
- It has a label for an individual read, followed by the actual sequence read itself, and then the quality scores that are in a special code which can be translated to a number
- The number reflects what percent of the time that base call is expected to be an error



- This is the overall workflow for the whole-genome SNP comparison going from the fastq file of sequence reads to a phylogenetic tree
- I will present the overview of the workflow here and then go into each of these steps in much more detail
- The first step is aligning the isolate sequence reads to a reference genome, we use M. tuberculosis H37Rv
- Then, SNPs relative to the reference genome H37Rv are identified
- The next step is that uninformative and unreliable SNPs are filtered out to produce a list of high-quality SNPs
- Lastly the high-quality SNPs are mapped on to a phylogenetic tree to diagram the genetic relationships between the isolates

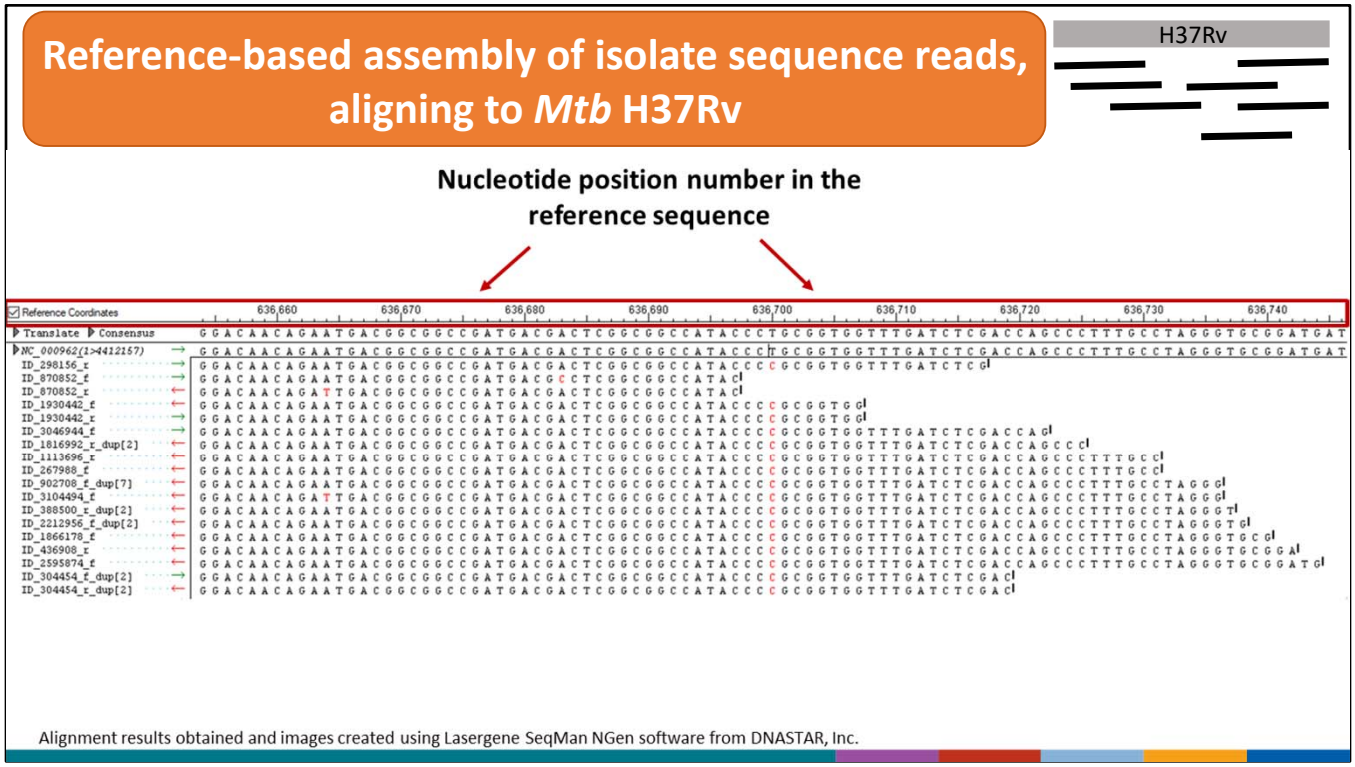
Reference-based assembly of isolate sequence reads, aligning to *Mtb* H37Rv

H37Rv

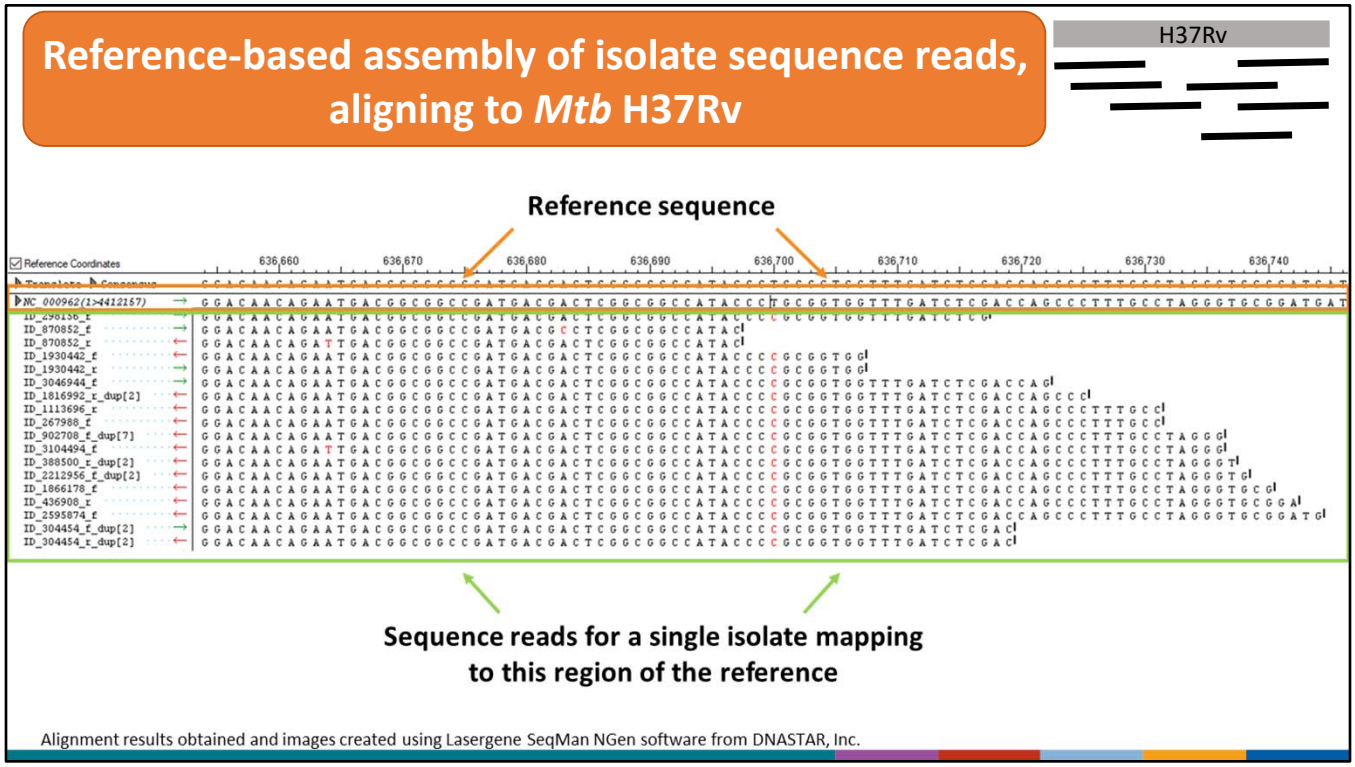
Alignment results obtained and images created using Lasergene SeqMan NGen software from DNASTAR, Inc.

- The first step is the reference-based assembly
- This is showing what it looks like when you align the sequence reads to the Mtb reference genome H37Rv

Reference-based assembly of isolate sequence reads, aligning to *Mtb* H37Rv

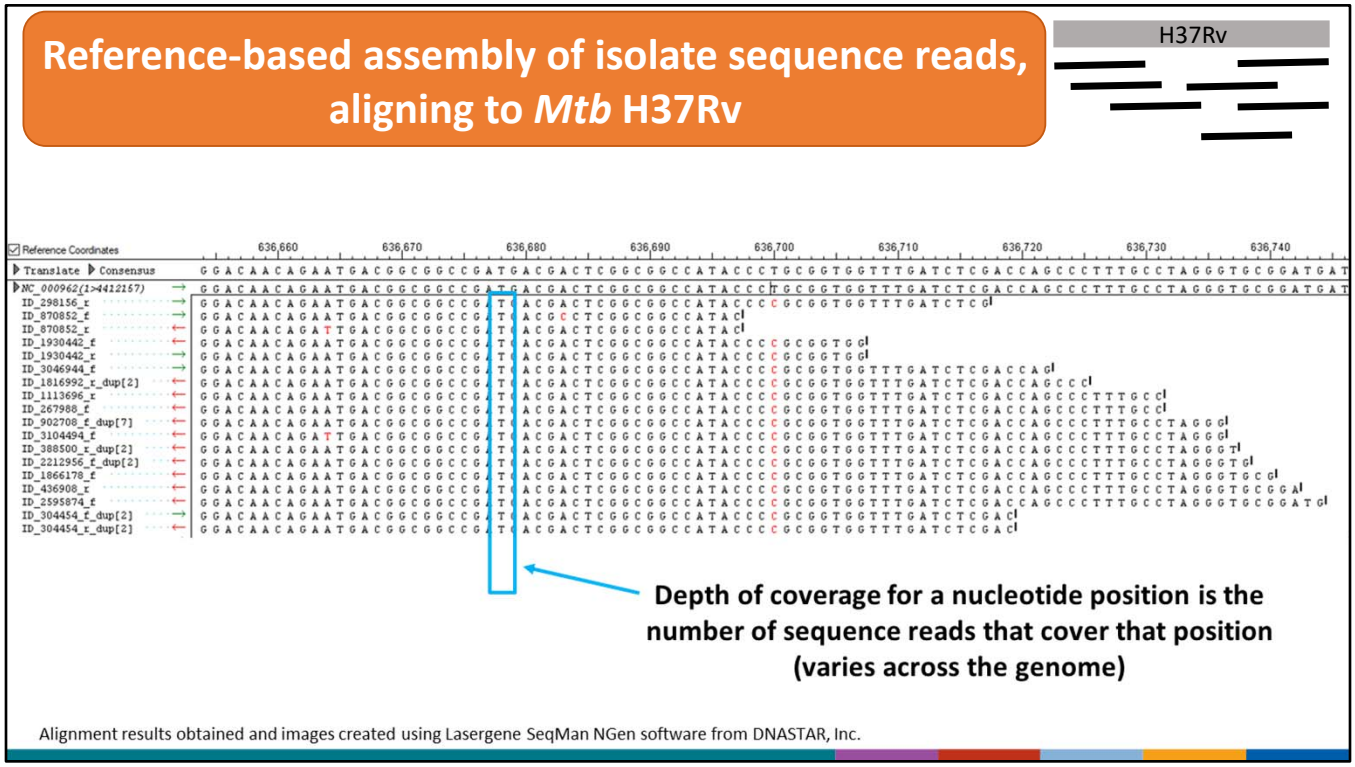


- Each of the 4.4 million nucleotides in the reference genome has a position number starting at 1
- If we zoom in, we can see these numbers boxed in red at the top that show the nucleotide position number in the reference sequence
- And we can see we are looking at part of the genome and we are looking at position 636,654 to position 636,746



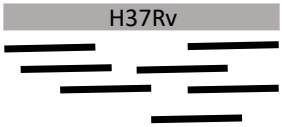
- We can also see the reference sequence in the orange box, and below that in the green box are all the sequence reads for a single isolate that are mapping to this region of the reference

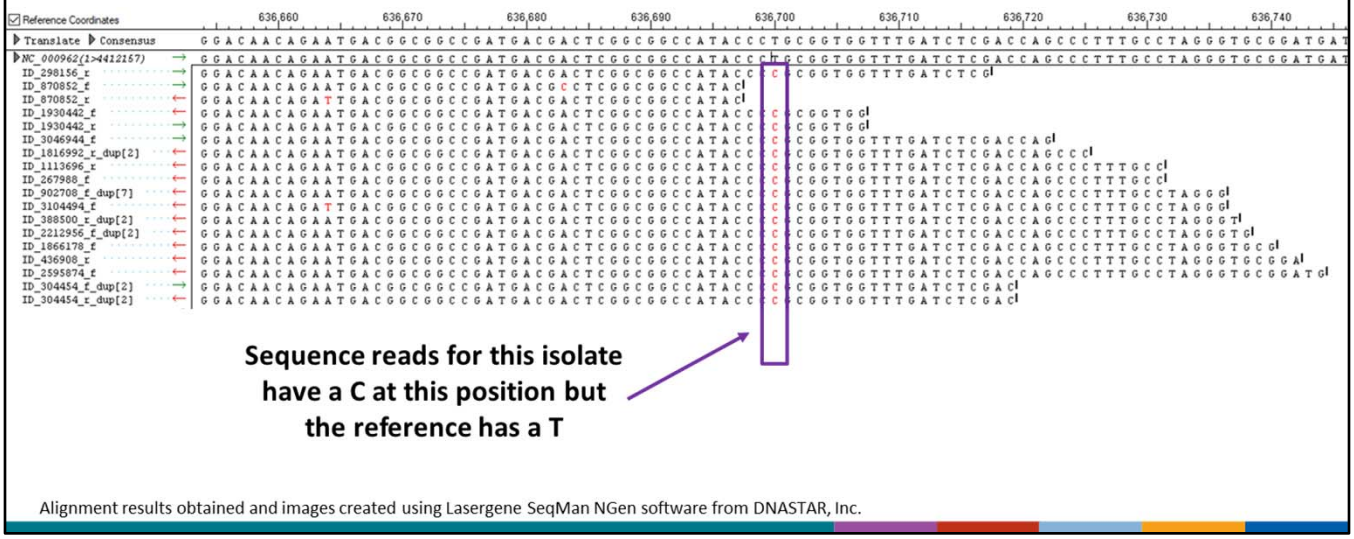
Reference-based assembly of isolate sequence reads, aligning to *Mtb* H37Rv



- If we were to look at one particular nucleotide position, for example this position boxed in blue, and look down at how many sequence reads cover that position, that would be the depth of coverage for that particular position
- The depth of coverage varies across the genome so there may be a lot of sequence reads for some regions and very few or none for other regions

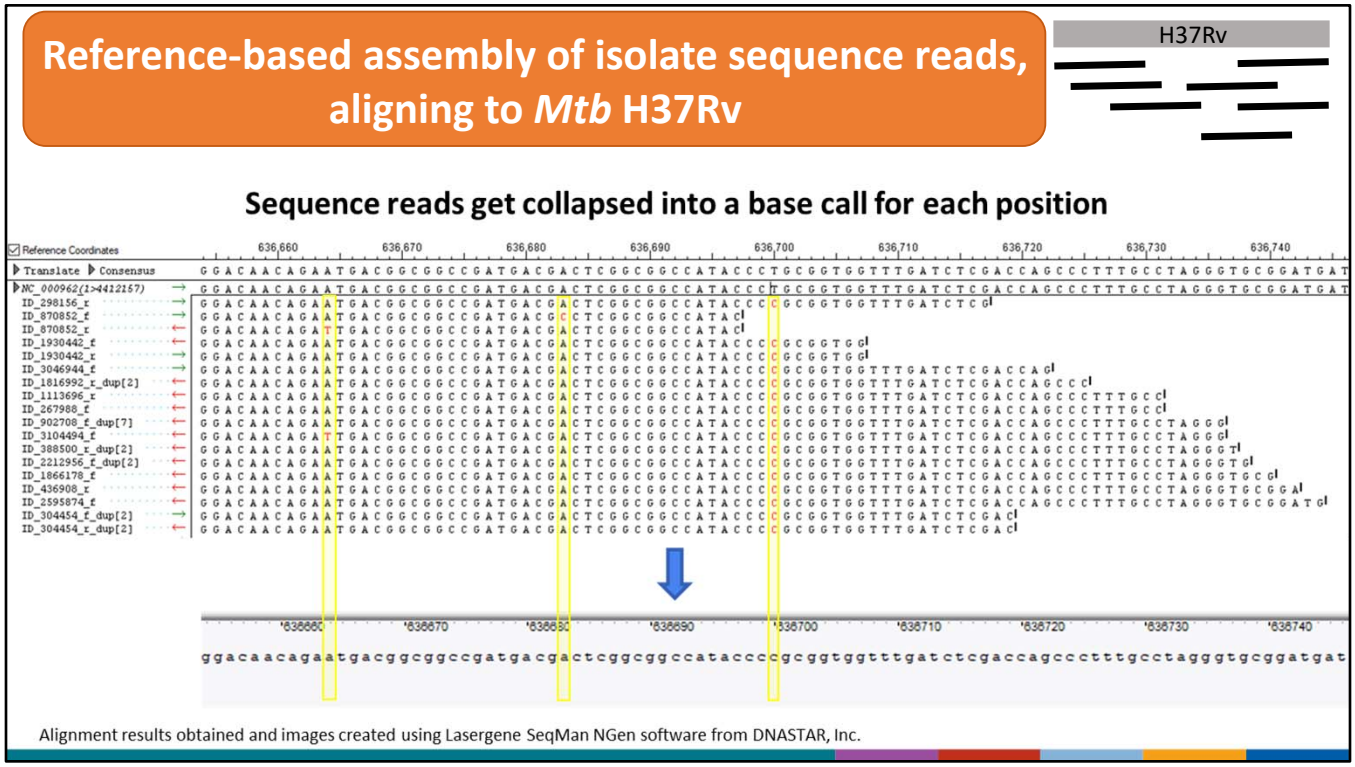
Reference-based assembly of isolate sequence reads, aligning to *Mtb* H37Rv





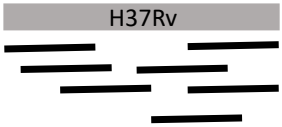
- Any bases in the sequence reads that do not match to the reference sequence are shown in red
- If we look at the nucleotide position 636,700 boxed in purple, we can see that the sequence reads for this isolate are all showing a C at this position but the reference has a T
- The other nucleotide positions in this region of the genome pretty much all match up with the reference, but some reads will occasionally have a different base call which may be a sequencing error

Reference-based assembly of isolate sequence reads, aligning to *Mtb* H37Rv



- The sequence reads then get collapsed into a base call for each position. A few are highlighted here in yellow
- We can see that the nucleotide position 636,700 where the reads all had a C but the reference has a T gets called as a C for this isolate
- Other positions might have a few reads that don't match the reference but they still get called as reference

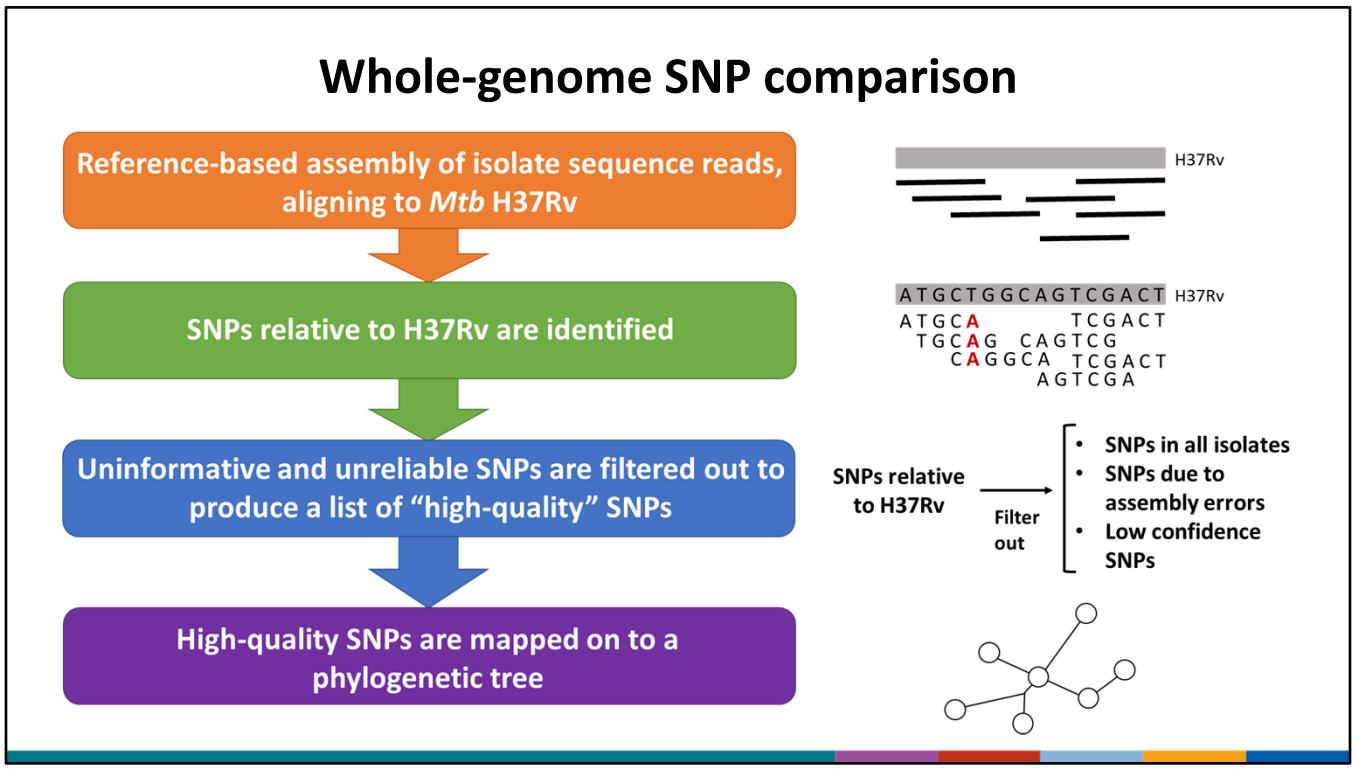
Reference-based assembly of isolate sequence reads, aligning to *Mtb* H37Rv



A,C, G, T…and beyond!

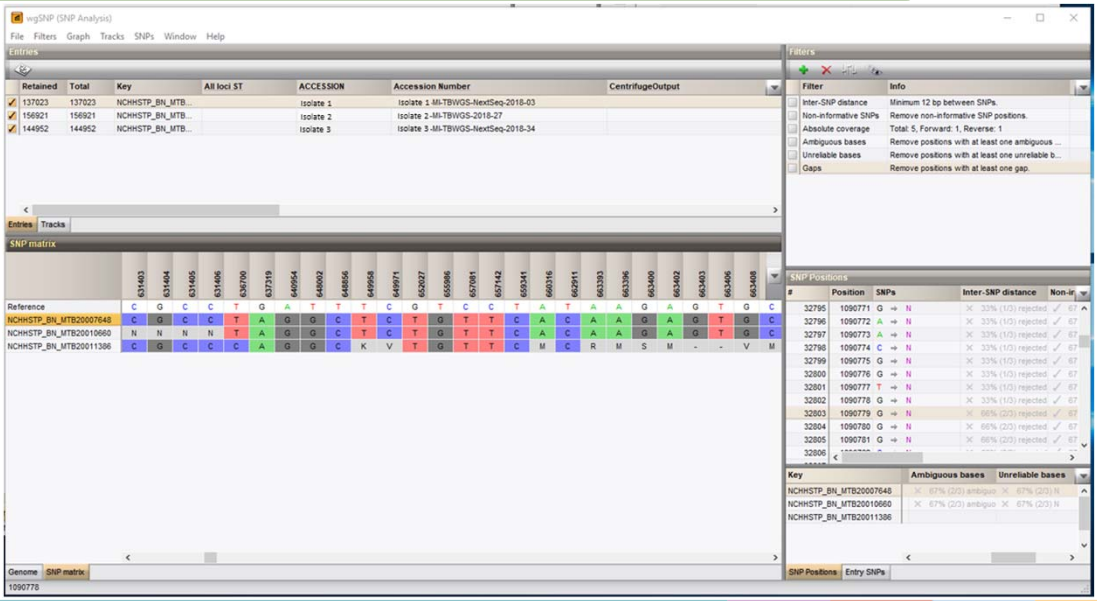
Symbol	Description	Bases			
A	Adenine	A			
C	Cytosine		C		
G	Guanine			G	
T	Thymine				T
W	Weak interaction	A			T
S	Strong interaction		C	G	
M	aMino	A	C		
K	Keto			G	T
R	puRine	A		G	
Y	pYrimidine		C		T
B	not A (B comes after A)		C	G	T
D	not C (D comes after C)	A		G	T
H	not G (H comes after G)	A	C		T
V	not T (V comes after T)	A	C	G	
N	any nucleotide but not a gap	A	C	G	T
-	Gap				

- But there are more base calls that can be made besides just A,C,G, and T
- There are other letters that stand for different mixtures of bases
- A mix of A and T gets called as a W, a mix of C and G is an S and so on
- If the sequence data at a position is unreliable, it gets assigned as N
- A dash indicates there is a gap in the alignment
- This could be due to an actual insertion/deletion event in the sequence or just no sequence coverage for a region



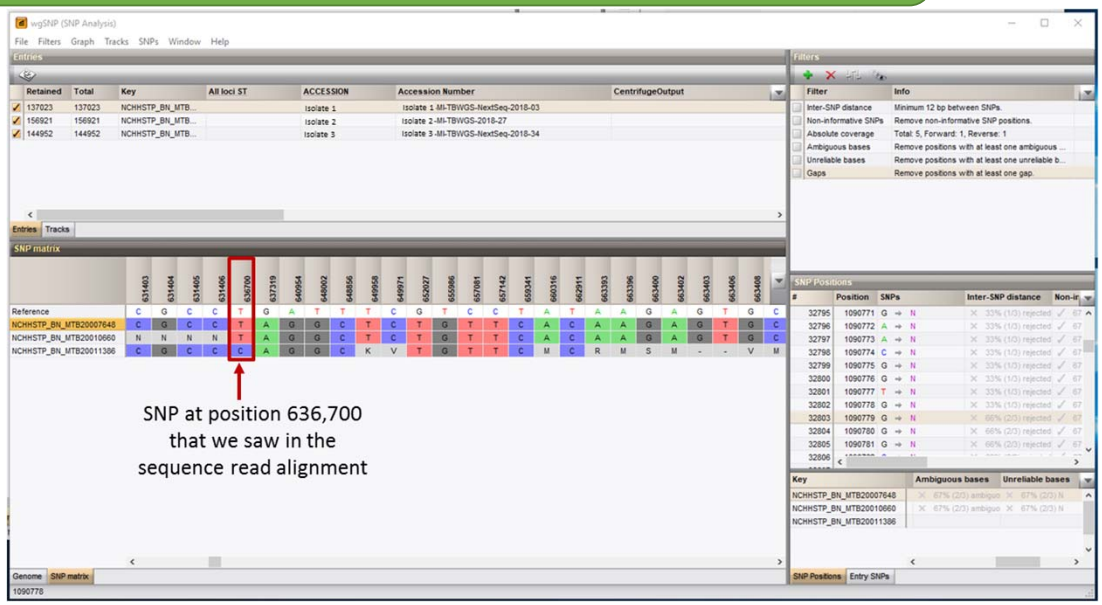
- Once we have the sequence reads assembled by aligning to the reference genome we can identify SNPs relative to the reference sequence

SNPs relative to H37Rv are identified

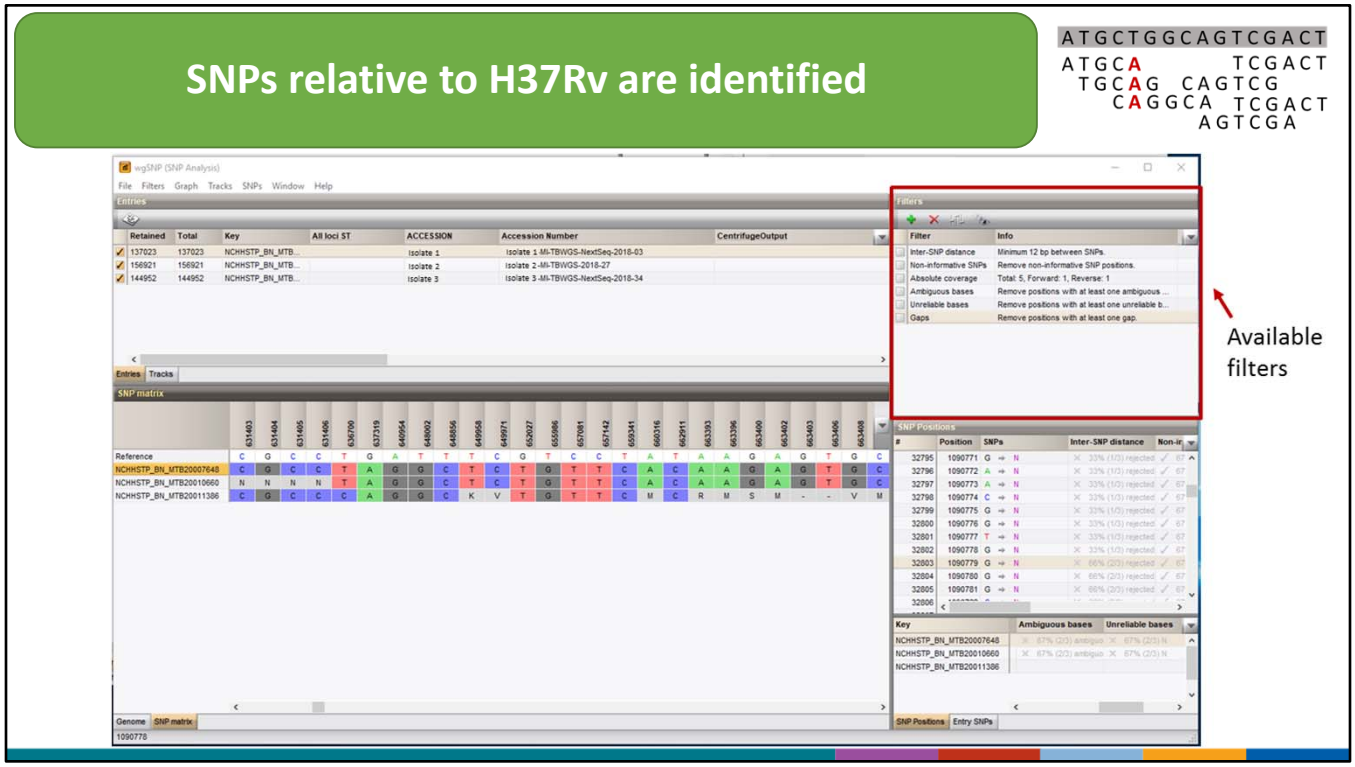


- This is showing the SNP analysis window of the BioNumerics software that is used for the whole-genome SNP comparison
- In this example, we will be analyzing three isolates that are all the same genotype
- In the top portion of the window there is some information about those three isolates
- Below that are all the nucleotide positions that are different from the reference sequence in at least one of these three isolates
- So now we are no longer looking at every single nucleotide position in the genome. We are now looking at only the positions where there is a difference

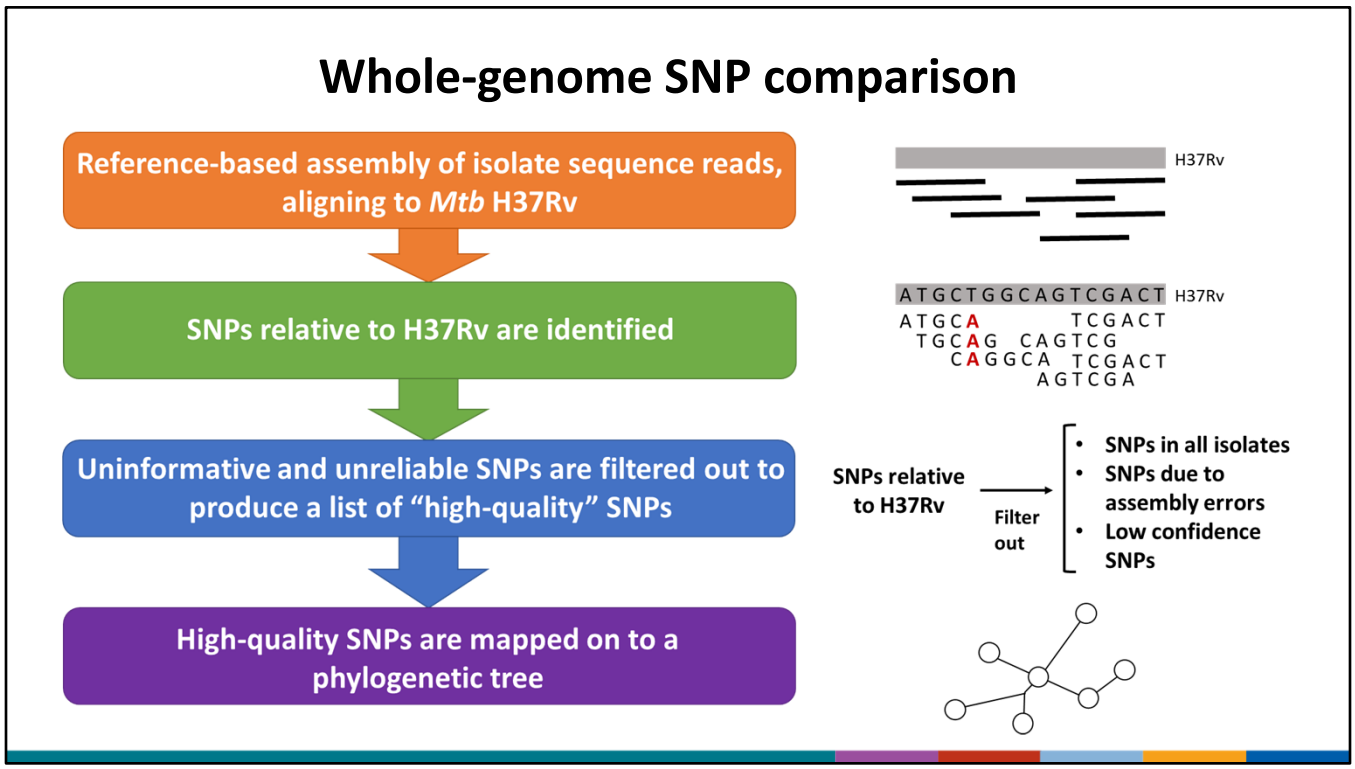
SNPs relative to H37Rv are identified



- And we can find the SNP at nucleotide position 636,700 that we saw in the alignment of sequence reads where one isolate had a C where the reference has a T
- We can see that the other two isolates in the analysis have a T at that position

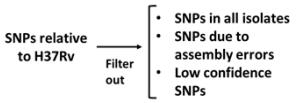


- Most of these nucleotide positions will ultimately get filtered out in the next step of the analysis, but right now I'm showing how it looks before the filtering
- The available filters are shown over here on the right in the red box, but they are all unchecked right now meaning they are turned off



- So the next step is to filter out uninformative and unreliable SNPs to produce a list of high-quality SNPs
- Let's take a look at what these different filters are

Uninformative and unreliable SNPs are filtered out to produce a list of "high-quality" SNPs



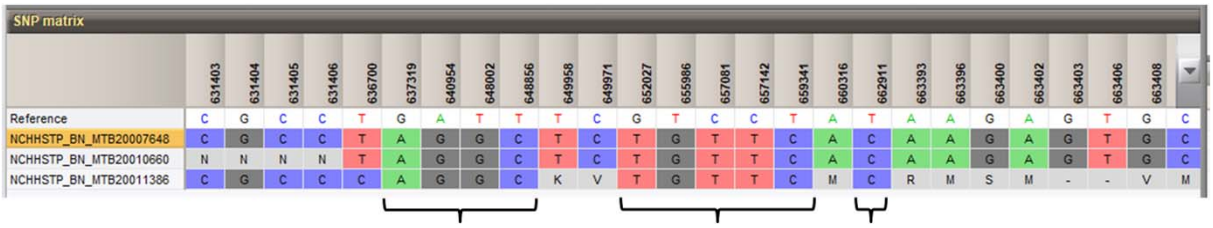
- **Six different filters**
  - Non-informative filter
  - Absolute coverage filter
  - Gaps filter
  - Unreliable bases filter
  - Ambiguous bases filter
  - Inter-SNP distance filter
- **Filters are not mutually exclusive**
  - Most SNPs get filtered for multiple reasons
- **Nucleotide position with a SNP needs to pass all six filters for all isolates in the comparison**

- There are six different filters that are applied. They are:
- The non-informative filter
- The absolute coverage filter
- The gaps filter
- The unreliable bases filter
- The ambiguous bases filter
- And the inter-SNP distance filter
- The filters are not mutually exclusive so most SNPs get filtered for multiple reasons
- And a nucleotide position with a SNP needs to pass all six filters for all isolates in the comparison to make it on to the list of high-quality SNPs
- I will explain all six of these filters in the following slides

Uninformative and unreliable SNPs are filtered out to produce a list of "high-quality" SNPs

SNPs relative to H37Rv → Filter out →
- SNPs in all isolates
- SNPs due to assembly errors
- Low confidence SNPs

Non-informative Filter



Non-informative filter: Removes positions with SNPs that are present in all isolates in the analysis

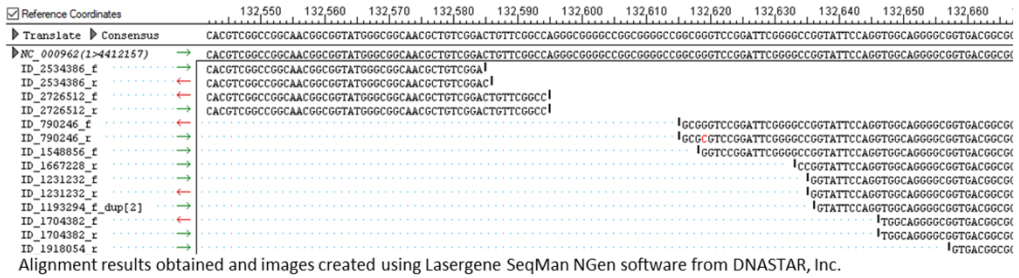
Rationale: Since these SNPs are present in all the isolates, they don't tell us anything about the genetic relationship among the isolates

- The non-informative filter removes positions with SNPs that are present in all isolates in the analysis
- Some are indicated here with the brackets
- Since these SNPs are present in all the isolates, they don't tell us anything about the genetic relationship among the isolates so they are removed

Uninformative and unreliable SNPs are filtered out to produce a list of "high-quality" SNPs

SNPs relative to H37Rv → Filter out →
- SNPs in all isolates
- SNPs due to assembly errors
- Low confidence SNPs

Absolute Coverage Filter



Alignment results obtained and images created using Lasergene SeqMan NGen software from DNASTAR, Inc.

Absolute coverage filter: Removes positions that don't have a total depth of coverage of at least 5 sequence reads, with at least one forward and one reverse read

Rationale: Since there are very little sequence data available for positions with low coverage, we have low confidence in these SNPs. They could be due to sequencing errors or assembly errors

- The absolute coverage filter removes positions that don't have a total depth of coverage of at least 5 reads, with at least one forward read and one reverse read
- You can see in this alignment that there is a portion of the reference genome that is not covered by the sequence reads as well as adjacent portions that are covered by less than five reads
- Since there are very little sequence data available for positions with low coverage, we have low confidence in these SNPs
- They could just be due to sequencing errors or assembly errors
- Or sometimes there's a SNP in one of the isolates and the coverage is good for that isolate but if one of the other isolates in the analysis has low coverage or no coverage at that position we can't know whether it has the reference base or the SNP so we wouldn't know how to place it on the tree

Uninformative and unreliable SNPs are filtered out to produce a list of "high-quality" SNPs

SNPs relative to H37Rv → Filter out
- SNPs in all isolates
- SNPs due to assembly errors
- Low confidence SNPs

Gaps Filter

SNP matrix

	631403	631404	631405	631406	636700	637319	640954	648002	648856	649958	649971	652027	655986	657081	657142	659341	660316	662911	663393	663396	663400	663402	663403	663406	663408
Reference	C	G	C	C	T	G	A	T	T	T	C	G	T	C	C	T	A	T	A	A	G	A	G	T	G	C
NCHHSTP_BN_MTB20007648	C	G	C	C	T	A	G	G	C	T	C	T	G	T	T	C	A	C	A	A	G	A	G	T	G	C
NCHHSTP_BN_MTB20010660	N	N	N	N	T	A	G	G	C	T	C	T	G	T	T	C	A	C	A	A	G	A	G	T	G	C
NCHHSTP_BN_MTB20011386	C	G	C	C	C	A	G	G	C	K	V	T	G	T	T	C	M	C	R	M	S	M	-	-	V	M

Gaps filter: Removes positions where at least one isolate in the analysis has a gap

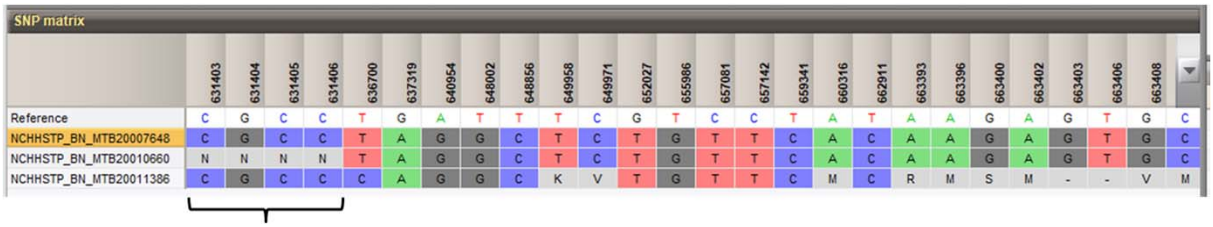
Rationale: Whole-genome SNP analysis only considers SNPs, not insertions or deletions. Since there is no base call for that position in the isolate with the gap, would not be able to assign it as either matching to the reference or having the SNP

- Somewhat similar to that is the gaps filter
- This removes positions where at least one isolate in the analysis has a gap, indicated here with a bracket
- And remember it could be a gap because of an actual insertion or deletion event or because there is little or no coverage at that position
- These positions are removed because whole-genome SNP comparison only considers SNPs, not insertions or deletions
- Also, since there is no base call for that position in the isolate with the gap, we would not be able to assign it as either matching to the reference or having the SNP

Uninformative and unreliable SNPs are filtered out to produce a list of "high-quality" SNPs

SNPs relative to H37Rv → Filter out
- SNPs in all isolates
- SNPs due to assembly errors
- Low confidence SNPs

Unreliable Bases Filter



Unreliable bases filter: Removes positions where at least one isolate in the analysis has an unreliable base

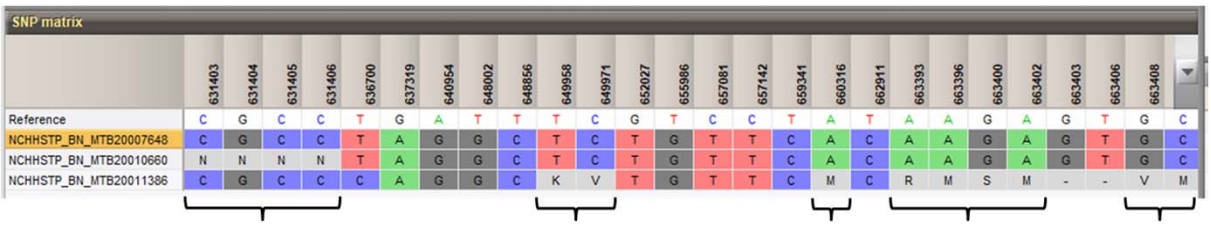
Rationale: We don't have enough confidence to call a base for that position

- There's also a filter for unreliable bases
- This removes positions where at least one isolate in the analysis has an unreliable base
- The base gets called as unreliable because the quality scores are low for that position
- So we basically just don't have enough confidence to call a base for that position

Uninformative and unreliable SNPs are filtered out to produce a list of "high-quality" SNPs

SNPs relative to H37Rv → Filter out
- SNPs in all isolates
- SNPs due to assembly errors
- Low confidence SNPs

Ambiguous Bases Filter



Ambiguous bases filter: Removes positions with at least one ambiguous base (a mixture of A, C, T, or G)

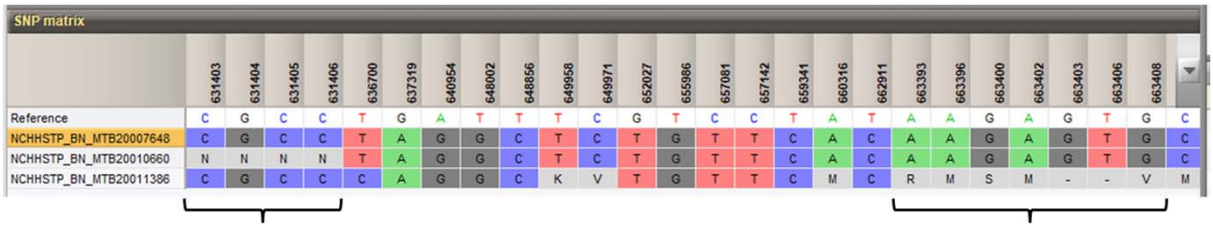
Rationale: Mixed base call could be due to sequencing errors or assembly errors. Mixed base call could also be due to a mixed infection and isolate cannot be designated as either matching the reference or having the SNP

- There's also a filter for ambiguous bases
- These are the positions that get called as a mixture of A, C, T, or G
- For a position to be called unambiguously as an A, C, T, or G, it needs to be in over 75% of the sequence reads
- A mixed base call could be due to sequencing errors or assembly errors
- A mixed base call could also be due to a mixed infection and the isolate cannot be designated as either matching the reference or having the SNP which has implications for where the isolate is mapped on the phylogenetic tree

Uninformative and unreliable SNPs are filtered out to produce a list of "high-quality" SNPs

SNPs relative to H37Rv → Filter out
- SNPs in all isolates
- SNPs due to assembly errors
- Low confidence SNPs

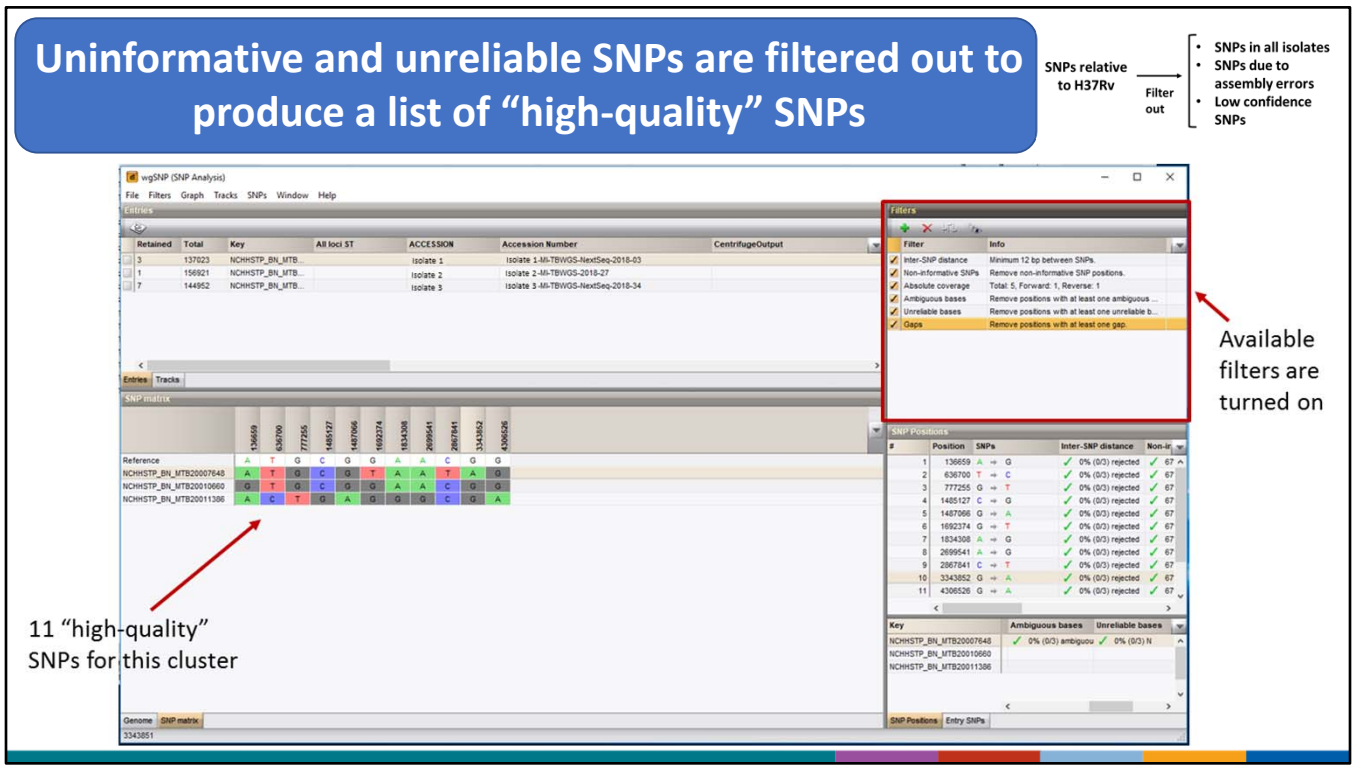
Inter-SNP Distance Filter



Inter-SNP distance filter: Removes positions where another SNP is present within a distance of 12 base pairs

Rationale: SNPs that are close together could be due to assembly errors (sequence reads were aligned to the wrong part of the genome)

- And finally the last filter is for inter-SNP distance
- This removes positions where another SNP is present within a distance of 12 base pairs
- These get removed because SNPs that are close together could be an indication of assembly errors (meaning the sequence reads were aligned to the wrong part of the genome)

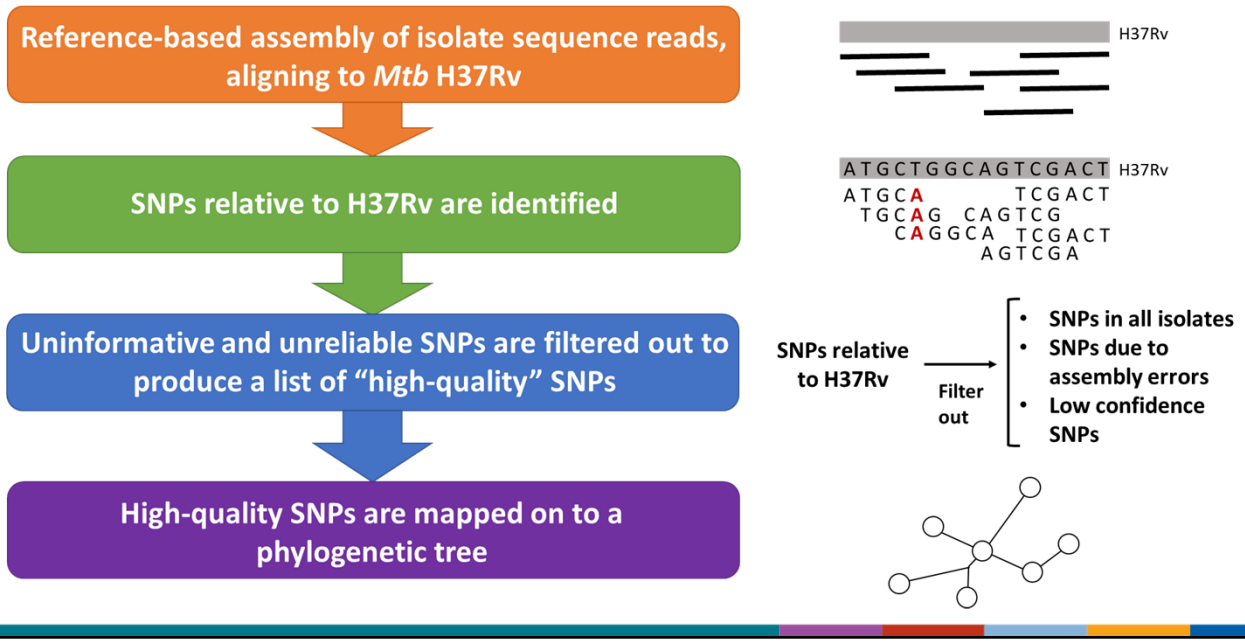


- So then if we put on all the filters (you can see in the red box on the right that they are all checked now), we narrow down to a short list of informative and reliable high-quality SNPs
- In this case we are left with 11 high-quality SNPs for this cluster

Building a phylogenetic tree and determining the placement of the most recent common ancestor (MRCA)

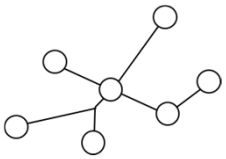
- In the next section I will describe how we build a phylogenetic tree and determine the placement of the most recent common ancestor (or MRCA)

Whole-genome SNP comparison



- So far, we aligned the sequence reads of the isolates to the reference H37Rv, identified SNPs, and then applied the filters to produce a list of high-quality SNPs
- For this cluster of three isolates, we were left with 11 high-quality SNPs and now the last step is to map the high-quality SNPs on to a phylogenetic tree

High-quality SNPs are mapped on to a phylogenetic tree



Tree-building methods vs. **Tree display options**

Maximum parsimony

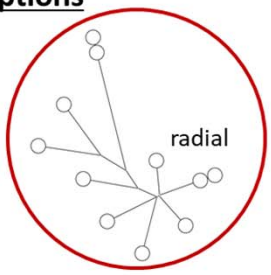
UPGMA

Neighbor-joining

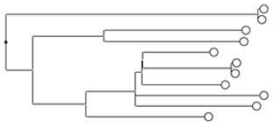
Minimum-spanning

Maximum Likelihood

circular

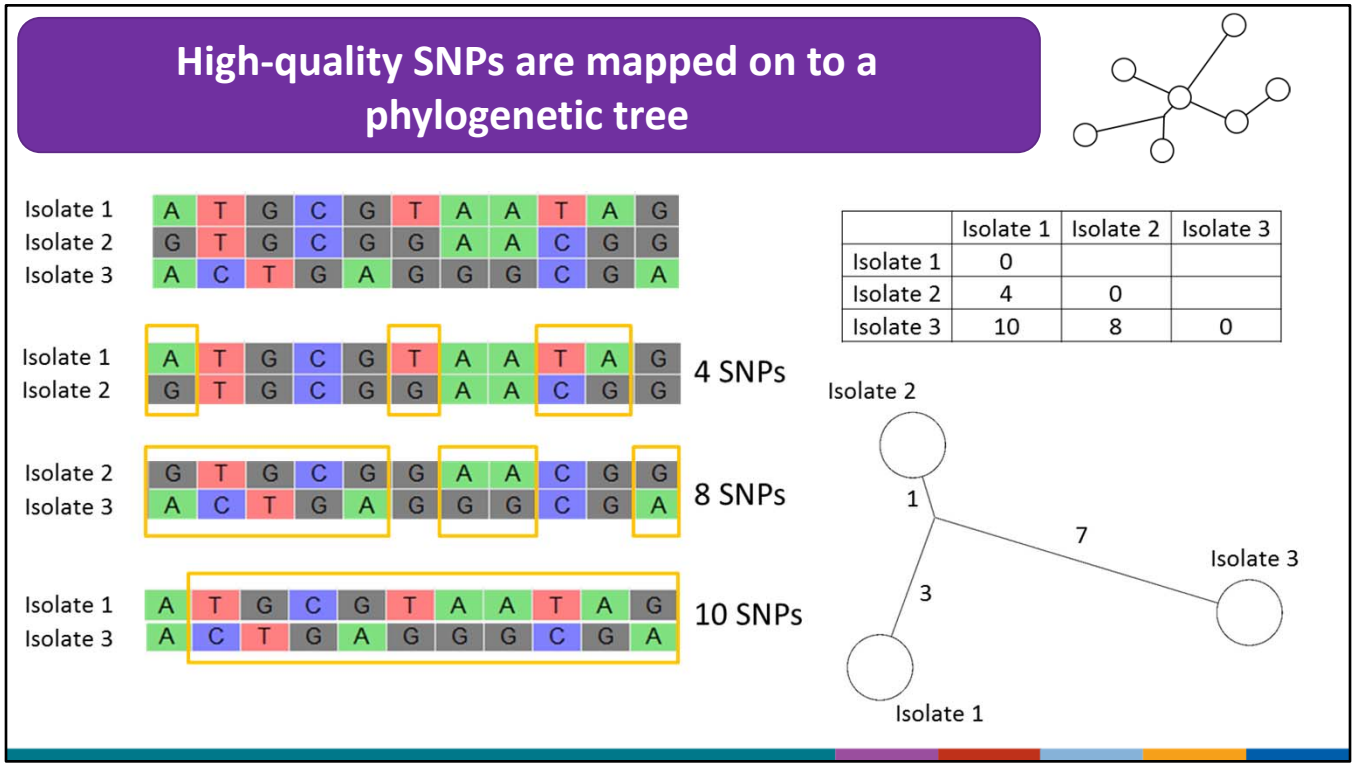


radial



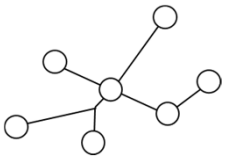
dendrogram

- We map them onto a phylogenetic tree so we can visualize the evolutionary relationships among the isolates
- There are a lot of different tree-building methods, some of which are listed here. We use the neighbor-joining method
- There are also a lot of different ways to display the trees. I show three here: radial, circular, and dendrogram
- We use a radial display but we could also display the neighbor-joining tree as a dendrogram or a circular layout
- So it's not a neighbor-joining tree because of how it looks
- It's neighbor-joining because of the underlying method that's used to figure out where the branch points are

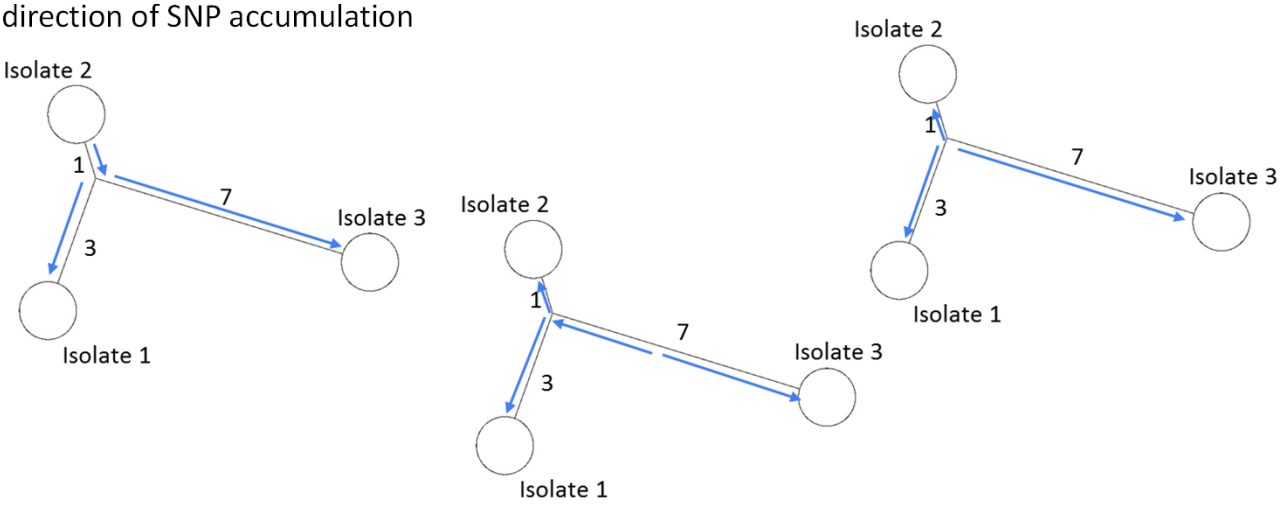


- A pairwise comparison of the isolates is done looking at the sequence at those 11 positions for each isolate
- This tells us the SNP distance between each pair
- If we look at Isolate 1 compared to Isolate 2, we see that these two isolates differ at 4 of the 11 positions
- Looking at Isolate 2 and Isolate 3, these two isolates differ at 8 of the 11 positions
- And Isolate 1 and Isolate 3 differ at 10 of the 11 positions
- This pairwise comparison between the isolates gives us a SNP distance matrix for the three isolates, shown here on the upper right
- To construct the phylogenetic tree, first one pair is joined together
- For example, Isolate 1 and Isolate 2. They are 4 SNPs apart
- Then you add in another neighbor, in this case Isolate 3, which is 10 SNPs from Isolate 1 and 8 SNPs from Isolate 2
- And you get a diagram that shows the SNP distance between each isolate but notice that there's no MRCA

High-quality SNPs are mapped on to a phylogenetic tree

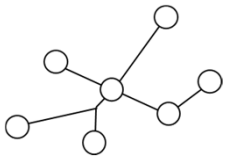


Important point: without the MRCA, we can't tell the direction of SNP accumulation



- Without the MRCA, we can't tell what the direction of SNP accumulation is
- This is showing just three of the many possible scenarios
- The blue arrows show the different possibilities of the direction of SNP accumulation
- Since SNPs generally do not revert, the direction of SNP accumulation can provide clues about the underlying chains of transmission when interpreting the phylogenetic tree

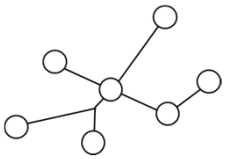
High-quality SNPs are mapped on to a phylogenetic tree



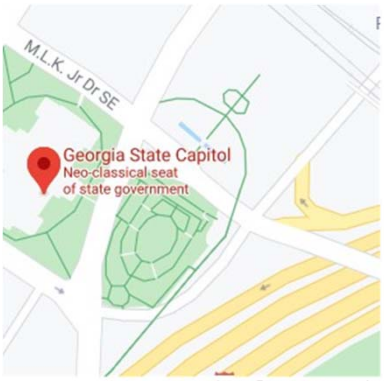
- **Rooting a tree and adding an MRCA**
  - Need to put your sequences of interest into a larger context to orient yourself as to where you are in evolutionary events

- To add the MRCA, you need to do what's called rooting the tree
- The point is that you need to put your sequences of interest into a larger context to orient yourself as to where you are in evolutionary events
- You can think of it like zooming out on the big picture of evolution

High-quality SNPs are mapped on to a phylogenetic tree

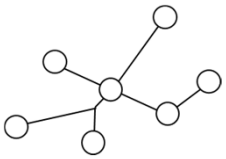


- **Rooting a tree and adding an MRCA**
  - Need to put your sequences of interest into a larger context to orient yourself as to where you are in evolutionary events



- An analogy would be if you were visiting Atlanta and wanted to go see the State Capitol
- This super zoomed in map wouldn't be very helpful in trying to figure out how to get there because you can't see where it is in relation to where you're coming from

High-quality SNPs are mapped on to a phylogenetic tree

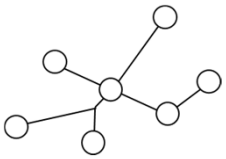


- **Rooting a tree and adding an MRCA**
  - Need to put your sequences of interest into a larger context to orient yourself as to where you are in evolutionary events

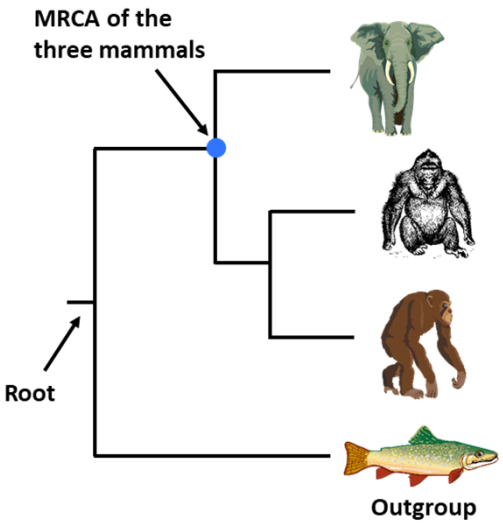
Zoom out

- But if you zoom out, you're better able to see which direction you'll be coming from so you'll know which way to turn as you get closer

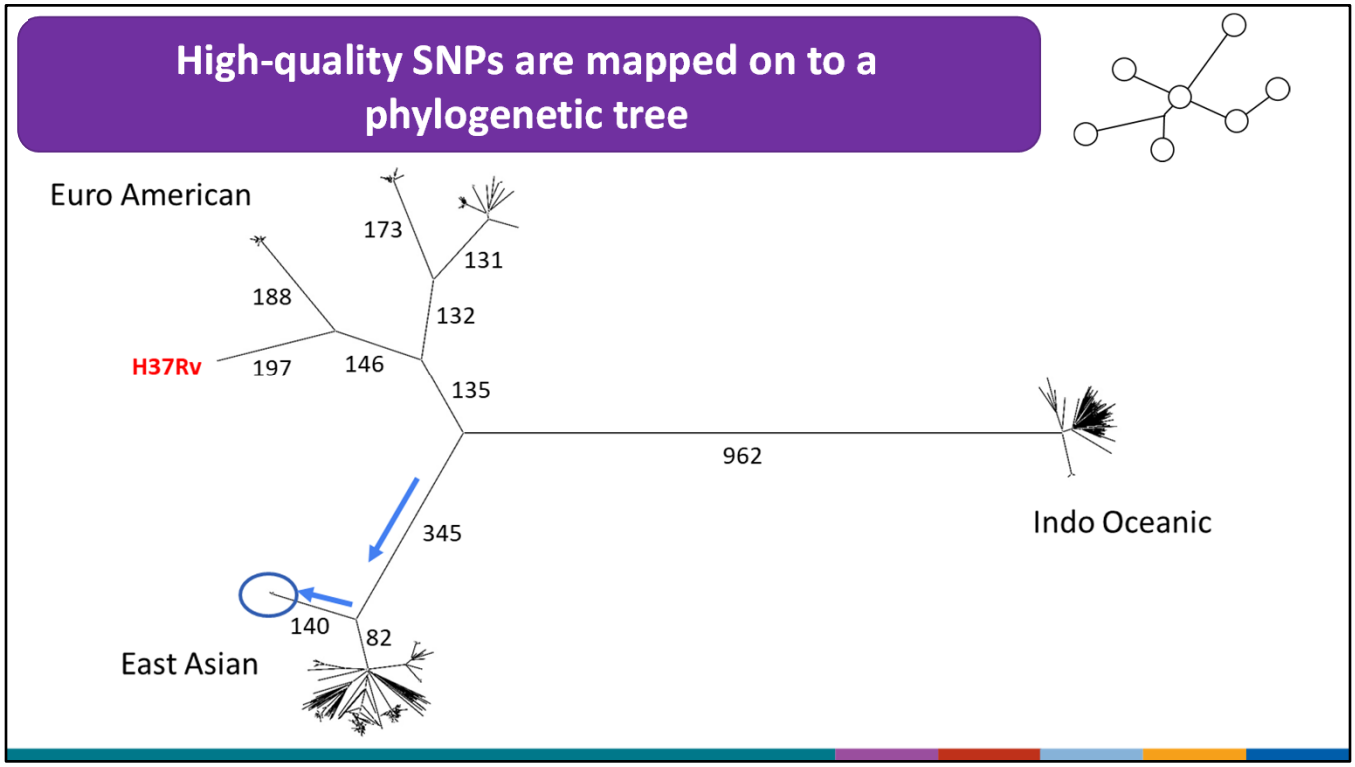
High-quality SNPs are mapped on to a phylogenetic tree



- **Rooting a tree and adding an MRCA**
  - This is usually done using an "outgroup" (something that is relatively distantly related to what you are analyzing)
  - Since we are analyzing very closely related *Mtb* isolates (matching genotypes) we can use a distantly related *Mtb* isolate
    - We use H37Rv

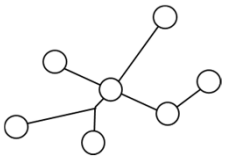


- This zooming out is usually done using an outgroup
- An outgroup is just something that is relatively distantly related from what you are analyzing
- Here in this example, we have a few mammals (an elephant, a gorilla, and a chimp) and the outgroup is a fish
- By rooting the tree using the fish as an outgroup, we can determine that the MRCA for our mammals of interest would be the point on the tree where this blue dot is
- Since we are analyzing very closely related Mtb isolates, we can use a distantly related Mtb isolate as an outgroup and we use the reference strain H37Rv

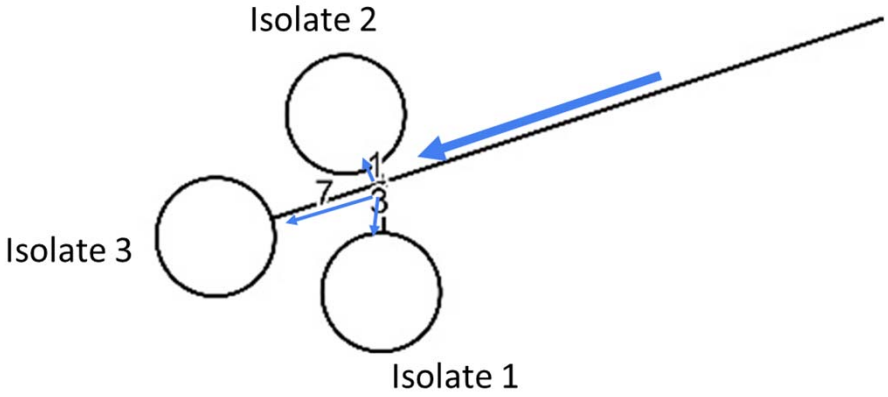


- This is the big picture, zoomed out view of where our three isolates of interest (shown circled in blue) fit in the context of other Mtb strains of various lineages including the reference strain H37Rv
- When we look over a much larger evolutionary time scale, the direction of genetic change (shown with the blue arrows) is more obvious
- So we can use something that's relatively distant (like H37Rv) to infer the most recent common ancestor of our three isolates of interest and the direction of genetic change

High-quality SNPs are mapped on to a phylogenetic tree

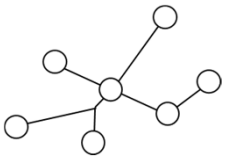


- **Using tree structure to determine placement of MRCA**

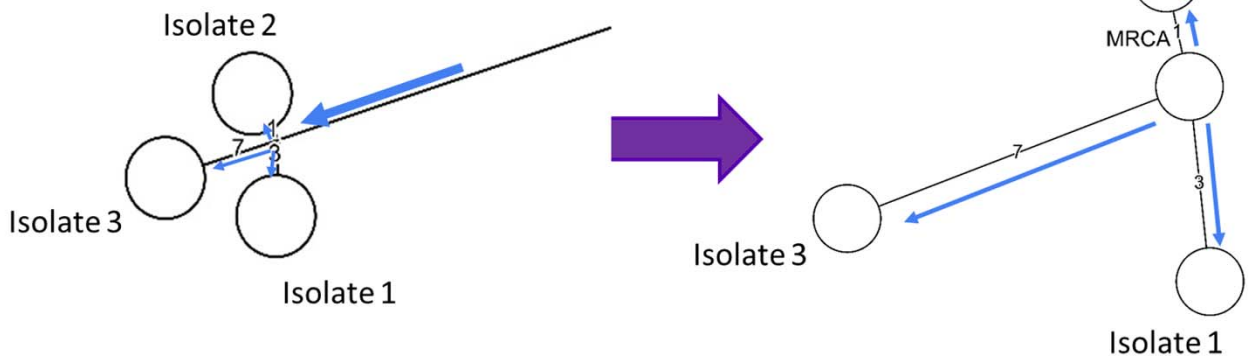


- Remember before we couldn't tell what the direction of genetic change between our three isolates was
- But if we zoom in on the isolates, it's now clear that SNPs are accumulating in the direction indicated by the blue arrows

High-quality SNPs are mapped on to a phylogenetic tree

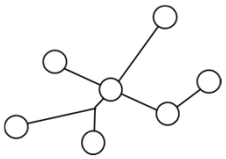


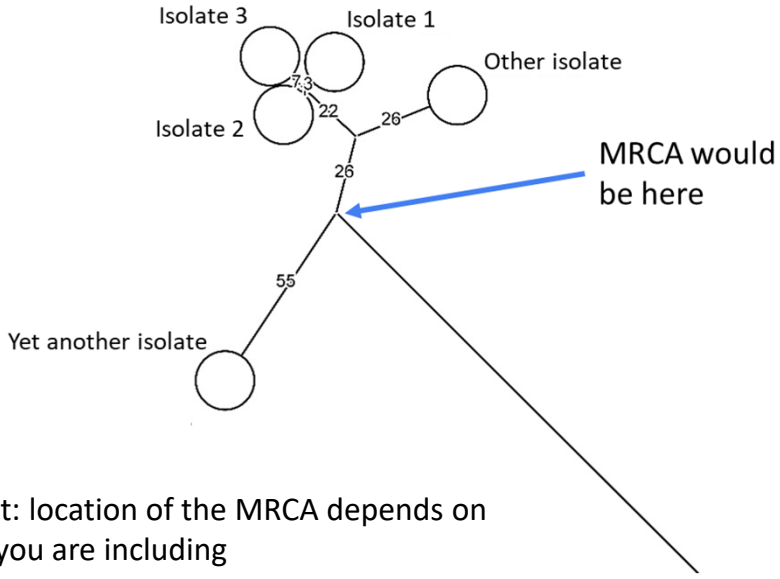
- **Using tree structure to determine placement of MRCA**



- And then we remove the branch with H37Rv and add the MRCA

High-quality SNPs are mapped on to a phylogenetic tree

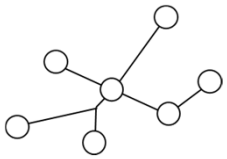




Important point: location of the MRCA depends on which isolates you are including

- The location of the MRCA depends on which isolates are included in the comparison
- If we add two more isolates to this comparison, the MRCA for these five isolates would now move to the point indicated on the tree
- You can think of it like you would a family: for example, if we were considering just me and my brothers, our most recent common ancestors would be our parents. But if we were considering me and my brothers and my cousins, then the most recent common ancestors for that group would be our grandparents

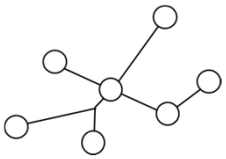
High-quality SNPs are mapped on to a phylogenetic tree



- **Using SNPs to determine placement of MRCA**
  - For each position with a "high-quality" SNP, use the base in H37Rv to represent the ancestral state
  - Allows us to infer the direction of genetic change
  - Rationale
    - 4.4 million bases in the *Mtb* genome
    - Isolates of interest and H37Rv have only diverged by ~1000 SNPs
    - H37Rv probably has not had changes at the same 11 positions as the isolates of interest
    - Expect base in H37Rv to represent ancestral state

- So that was one way of thinking about it where we made a tree with an outgroup, H37Rv, and then looked at the structure of the tree to see where the MRCA for the isolates of interest would be
- Now I'm going to show another way to think about it, looking at the actual SNPs themselves
- We're doing the same thing, I'm just approaching the explanation from a different perspective
- Here we're going to go through each position with a high-quality SNP that differentiates the isolates in the comparison and use the base in H37Rv to represent the ancestral state
- This allows us to infer the direction of genetic change at each of these positions
- The rationale for this is: if we consider that there's 4.4 million bases in the Mtb genome and our isolates of interest and H37Rv have only diverged by about 1000 SNPs, then we can reason that H37Rv probably has not had basepair changes at any of these same 11 positions that we identified in these three isolates of interest
- Therefore, we can expect that the base in H37Rv at these 11 positions represents the ancestral state at that position

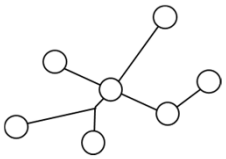
High-quality SNPs are mapped on to a phylogenetic tree



H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

- I will walk through this process, using the analysis of the three isolates as an example
- For each of the 11 SNPs that we identified in this analysis, we can look at what nucleotide H37Rv has at that position and, from that, determine the direction of genetic change

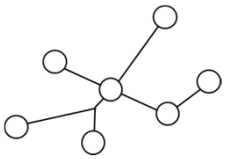
High-quality SNPs are mapped on to a phylogenetic tree



H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

- If we look at this first position boxed in orange for example, two of the isolates had an A and one has a G

High-quality SNPs are mapped on to a phylogenetic tree

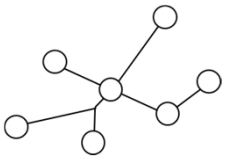


H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

A
↓
G

- H37Rv has an A so we are going to conclude that the direction of change here was A to G, and Isolate 2 had an A to G mutation at that position

High-quality SNPs are mapped on to a phylogenetic tree

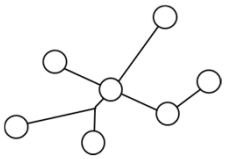


H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

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

- And then we would go through that same thought process for each of the 11 SNPs

High-quality SNPs are mapped on to a phylogenetic tree



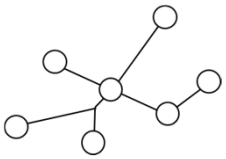
H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

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

H37Rv	0	0	0	0	0	0	0	0	0	0	0
Isolate 1	0	0	0	0	0	1	0	0	1	1	0
Isolate 2	1	0	0	0	0	0	0	0	0	0	0
Isolate 3	0	1	1	1	1	0	1	1	0	0	1

- Next, I'm just going to convert this to 1's and 0's to make it easier to look at
- If it has the ancestral base (same as H37Rv), then I'm going to give it a 0 and if it has the mutation then I will mark it as 1
- Then, we can start to draw the tree

High-quality SNPs are mapped on to a phylogenetic tree



H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

A T G C G G A A C G G

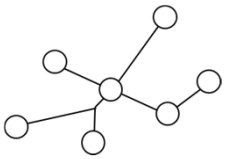
↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

G C T G A T G G T A A

H37Rv	0	0	0	0	0	0	0	0	0	0	0
Isolate 1	0	0	0	0	0	1	0	0	1	1	0
Isolate 2	1	0	0	0	0	0	0	0	0	0	0
Isolate 3	0	1	1	1	1	0	1	1	0	0	1

- First, we'll start with the MRCA
- This is just designating the genomic starting point before any of these 11 SNPs happened

High-quality SNPs are mapped on to a phylogenetic tree



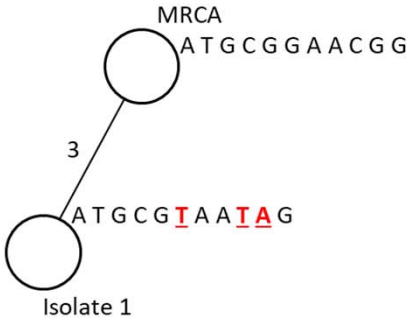
H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

A T G C G G A A C G G

↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

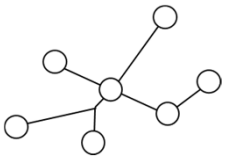
G C T G A T G G T A A

H37Rv	0	0	0	0	0	0	0	0	0	0	0
Isolate 1	0	0	0	0	0	1	0	0	1	1	0
Isolate 2	1	0	0	0	0	0	0	0	0	0	0
Isolate 3	0	1	1	1	1	0	1	1	0	0	1



- Then, let's add the first isolate, Isolate 1
- It has three SNPs, so we'll draw a branch, put a 3 next to it and draw the node

High-quality SNPs are mapped on to a phylogenetic tree



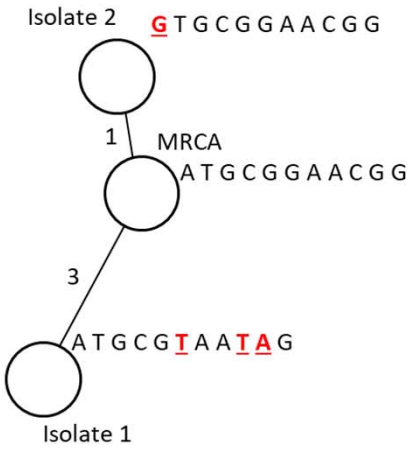
H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

A T G C G G A A C G G

↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

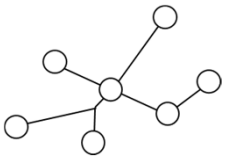
G C T G A T G G T A A

H37Rv	0	0	0	0	0	0	0	0	0	0	0
Isolate 1	0	0	0	0	0	1	0	0	1	1	0
Isolate 2	1	0	0	0	0	0	0	0	0	0	0
Isolate 3	0	1	1	1	1	0	1	1	0	0	1



- Next we can add Isolate 2
- This one has one SNP and it's not the same SNP as any of the three that we've already mapped, so we will draw it coming off the MRCA on a different branch

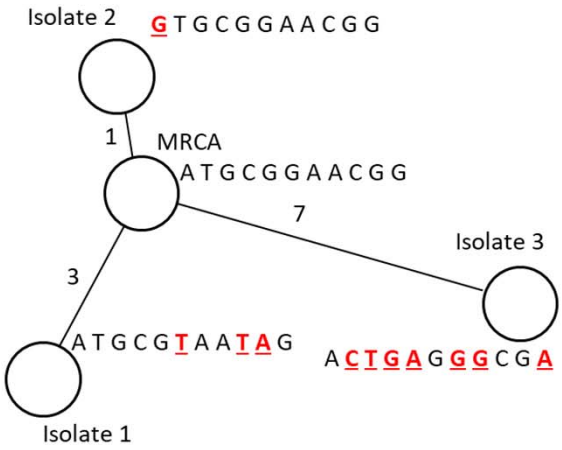
High-quality SNPs are mapped on to a phylogenetic tree



H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

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

H37Rv	0	0	0	0	0	0	0	0	0	0	0
Isolate 1	0	0	0	0	0	1	0	0	1	1	0
Isolate 2	1	0	0	0	0	0	0	0	0	0	0
Isolate 3	0	1	1	1	1	0	1	1	0	0	1



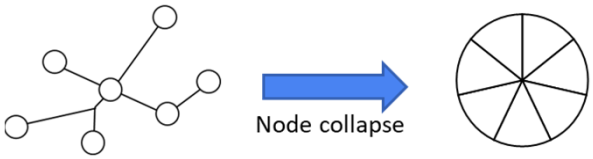
- And lastly, Isolate 3 has 7 SNPs and they are not the same as any of the others that we have already mapped so we can draw its branch coming off the MRCA in another direction

Considerations for how adding or removing isolates from the comparison affects results

- With that understanding of the details of the methods behind whole-genome SNP comparison, let’s look at some considerations for how adding or removing isolates from the comparison can affect the results

Excluding isolates from the comparison

- **Low coverage**
  - The average depth of coverage across the whole genome is low
  - If the average is low, there will be a lot of positions that don't pass the low coverage filter
  - Many informative, high-quality SNPs could be lost if an isolate with low coverage is added to the comparison
  - Nodes collapse



- Sometimes isolates need to be excluded from the comparison because of quality issues
- One issue is low coverage
- When we exclude an isolate because it has low coverage, that means that the average depth of coverage across the whole genome is low
- If the average depth of coverage across the genome is low, there will be a lot of positions that don't pass the low coverage filter
- That means many informative, high-quality SNPs could be lost if an isolate with low coverage is added to the comparison
- And when you lose SNPs, you lose branches and the nodes collapse, making the isolates appear to be more closely related than they are

Excluding isolates from the comparison

- **Contaminated isolates**
  - Isolate is contaminated with something non-*Mtb*
  - Fastq file will contain non-*Mtb* sequence reads
  - These reads could possibly align to the H37Rv reference sequence
    - SNPs that are in the contaminant → isolate looks more distant than it really is
    - Mixed SNPs leading to loss of SNPs → node collapse

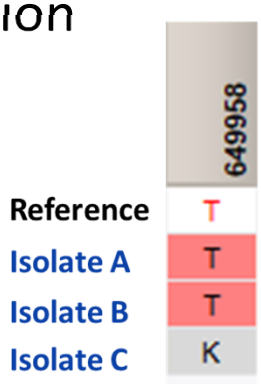
- We also exclude isolates if they are contaminated with something other than Mtb
- If they are contaminated, then the fastq file will contain non-Mtb sequence reads and these could possibly align to the H37Rv reference sequence
- This could have two different problematic outcomes
- One is if there's SNPs that are in the contaminant that are able to pass all the filters then these SNPs will get mapped on to the tree and the isolate will look more distant than it really is
- The other is that there could be a lot of ambiguous bases (or mixed SNPs) because of the contaminant and that would lead to loss of SNPs and node collapse, causing isolates to look more closely related than they really are

Sometimes SNP distances will decrease when an additional isolate is added to the comparison

- **Reasons for loss of SNPs**
  - Low coverage at that position
  - Unreliable base call at that position
  - Gap at that position
  - Another SNP within 12 base pairs of that position
  - Ambiguous (mixed) base call at that position

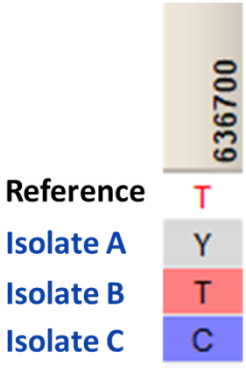
- Even though isolates with WGS quality issues like low coverage or contamination are excluded from the comparison, there can still be changes in the SNP distances when a new isolate is added to the comparison because of the filtering criteria
- If a SNP that passed all the filters in a previous comparison does not pass all the filters in the newly added isolate, it will be filtered out
- There could be low coverage, an unreliable base call, a gap, another SNP within 12 base pairs, or an ambiguous or mixed base at that position in the isolate that is being newly added to the analysis

Ambiguous bases can sometimes be due to mixed infection



Ambiguous base is the only difference from reference at this position

Likely due to sequencing or assembly errors

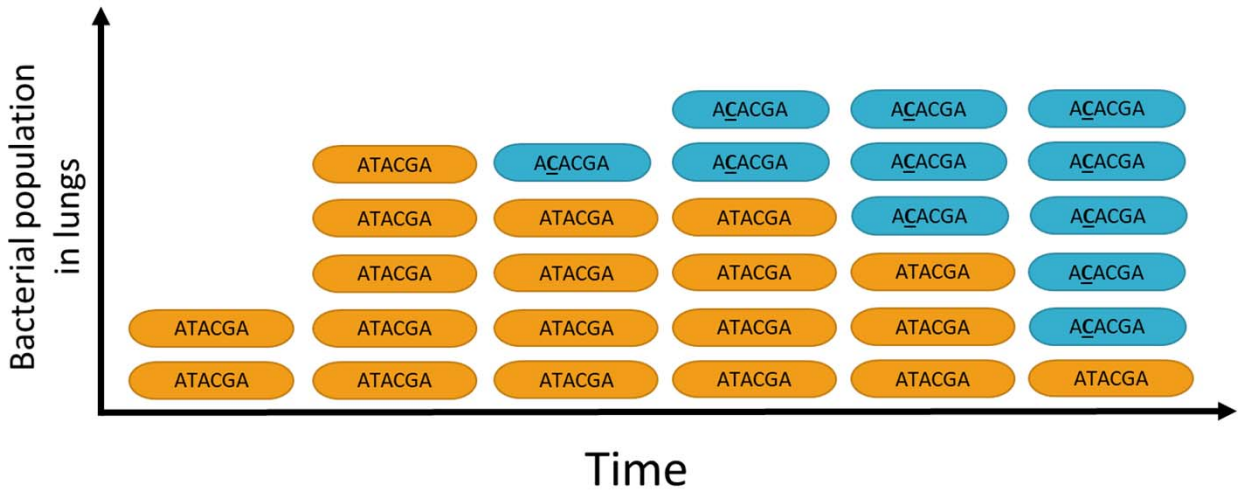


Ambiguous base is mixture of reference and SNP that is variably present in other isolates

Ambiguous base could indicate mixed infection

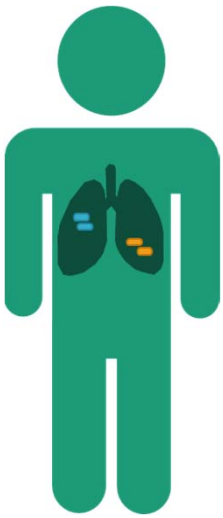
- A mixed base call could be because of sequencing or assembly error or because the patient has a mixed infection of Mtb
- In the case on the left, the ambiguous base is the only difference from the reference at this position
- This situation is common and often just due to sequencing errors or assembly errors
- But sometimes we see the case on the right where the ambiguous base is a mixture of reference and a SNP that is variably present in other isolates
- In this example, Isolate A has a Y at this position which stands for a mix of T and C, and these base calls are seen in other isolates in the analysis
- In this case there is a higher likelihood that the ambiguous base call indicates an actual mixed population of bacteria in the patient's sample because the patient has a mixed infection

Mixed infections: subpopulations can emerge over time



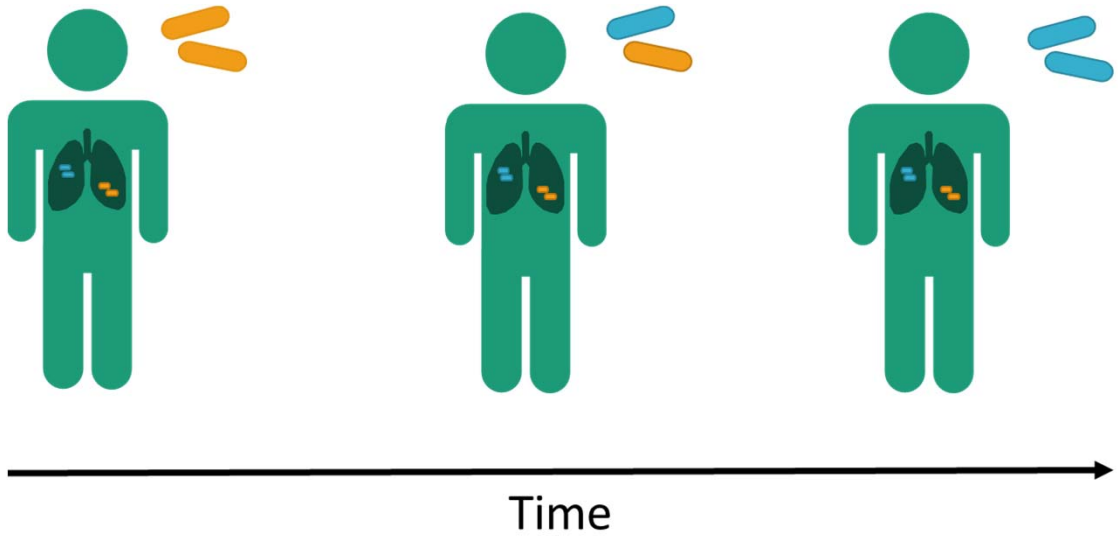
- So let's think some more about mixed infections
- A mixed infection in a patient can be because the person was infected with a mixed population or infected multiple times with different strains
- A mixed infection in a patient can also result from SNPs occurring and subpopulations with those SNPs emerging over time

Mixed infections: subpopulations are not uniformly distributed throughout the lungs



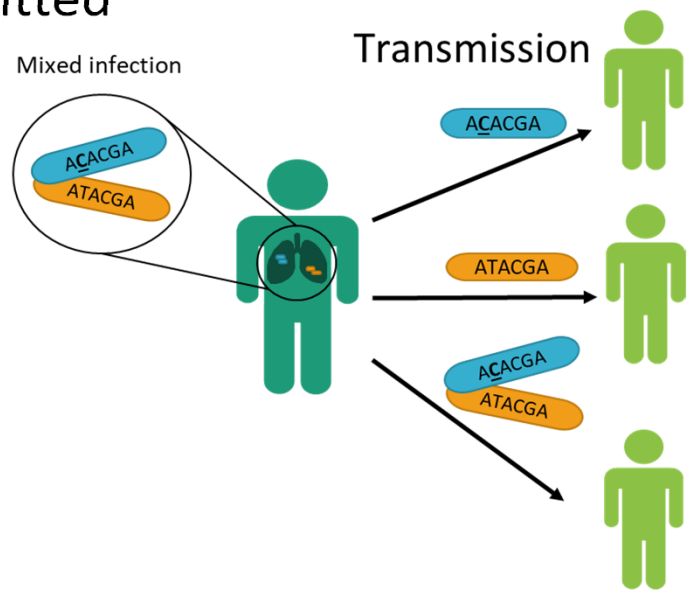
- And these subpopulations are not uniformly distributed throughout the lungs, so you could have different subpopulations in different lesions in the lungs

Mixed infections: lesions may be active at different times



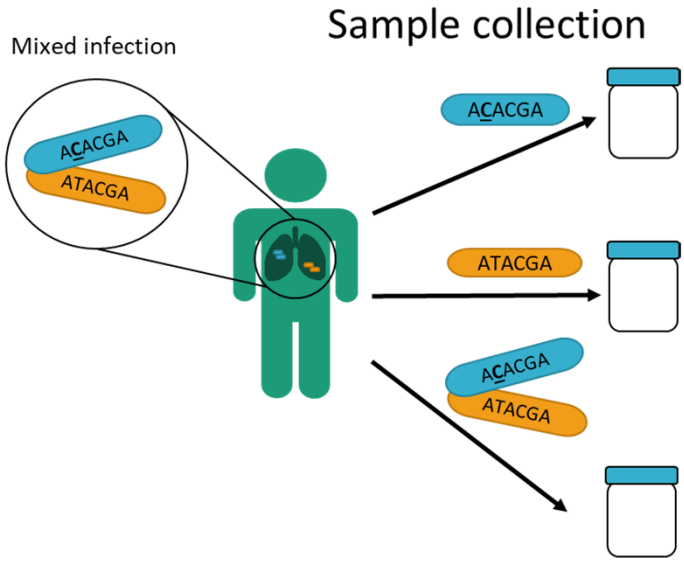
- Furthermore, these lesions might be active at different times, so different subpopulations might be coughed up at different times during a patient's infectious period

Mixed infections: one or more subpopulations can be transmitted



- Since different subpopulations may be coughed up at different times, a person with mixed infection can transmit one of the subpopulations or a mixture of the subpopulations

Mixed infections: sample can contain one or more subpopulations

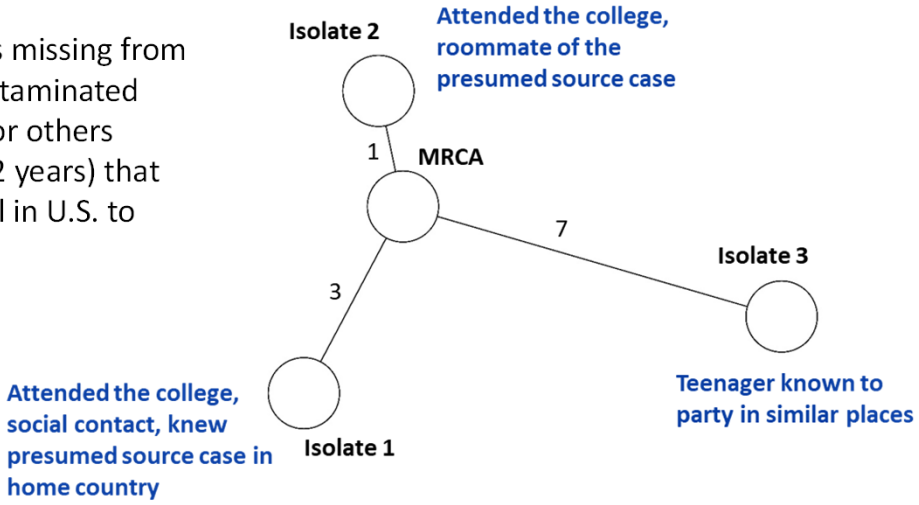


- Similarly, the sample that is collected from the person could have one or both subpopulations

Case study: Initial whole-genome SNP comparison

Isolate from a fourth case is missing from the tree because it was contaminated

- Presumed source case for others
- Long infectious period (2 years) that started soon after arrival in U.S. to attend college

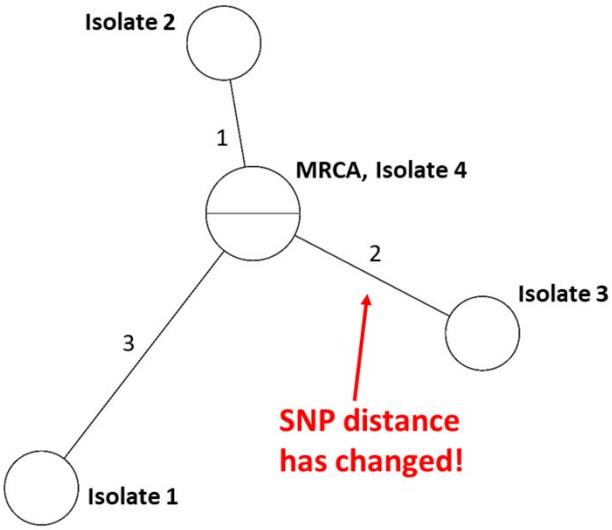


Important points:

- Tree should be interpreted in the context of known epidemiologic information
- Consider if isolates are missing from the tree

- As an example, let's consider the phylogenetic tree that we generated earlier in this training
- This example is particularly interesting because these results were unexpected so it's a good example of taking into account the known epi information and also considering if isolates are missing from the tree when interpreting the results
- This also serves as an example of how SNP distances can change depending on what isolates are in the comparison because of how the SNP filtering is done
- This cluster alerted with four cases, but the isolate from one of the cases was contaminated so it was left off the tree
- Unfortunately the contaminated isolate was from the presumed source case who had a two year infectious period that started soon after arrival in the U.S. to attend college
- The reason the result was unexpected was because one of the isolates (Isolate 3) was somewhat distant from the other isolates and this isolate was from a teenager who was known to party in the same social network
- Also, although this cluster shared a fairly common genotype, the cluster happened in a small town and several cases occurred all at the same time, so it would be surprising if this case in the teenager wasn't actually related to the other cases
- So CDC and the state program agreed not to rule out the case's involvement in this cluster
- The state was able to send an additional isolate for the presumed source case to replace the contaminated one, and this was included in an updated whole-genome SNP comparison

Case study: Updated whole-genome SNP comparison



Important point:

- SNP distances can change depending on which isolates are in the comparison

- When we added the new isolate from the presumed source case (Isolate 4), it ended up being in the same node as the MRCA
- And now the branch that had 7 SNPs in the previous phylogenetic tree has 2 SNPs

Case study: Why did the SNP distance change?

Some of the high-quality SNPs in the initial comparison were mixed in Isolate 4

H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 4	A	Y	G	S	R	G	R	R	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

Y = T or C
S = C or G
R = A or G

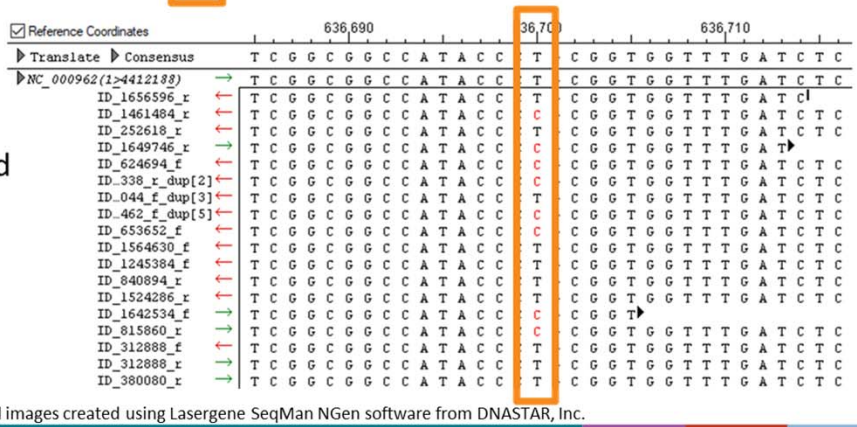
- So why did the SNP distance change?
- It happened because some of the 11 high-quality SNPs that we identified in the initial whole-genome SNP comparison were mixed in the newly added isolate 4, so they ended up getting filtered out in the updated comparison

Case study: Why did the SNP distance change?

H37Rv	A	T	G	C	G	G	A	A	C	G	G
Isolate 4	A	Y	G	S	R	G	R	R	C	G	G
Isolate 1	A	T	G	C	G	T	A	A	T	A	G
Isolate 2	G	T	G	C	G	G	A	A	C	G	G
Isolate 3	A	C	T	G	A	G	G	G	C	G	A

Y = T or C
S = C or G
R = A or G

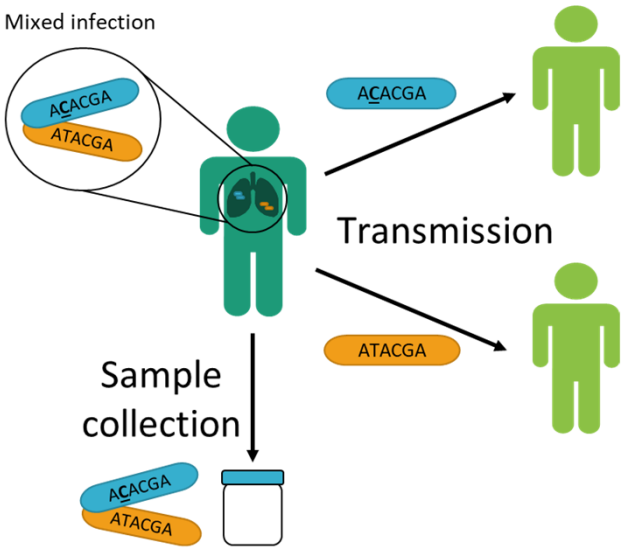
Sequence read alignment for Isolate 4



Alignment results obtained and images created using Lasergene SeqMan NGen software from DNASTAR, Inc.

- For example, if we go back to that SNP at position 636700 (boxed here in orange) that was a C in Isolate 3 from the teenager, it's getting called as a Y in Isolate 4, which is a mix of C and T
- And if we look at the alignment for Isolate 4 (the isolate from the presumed source case that we just added), we can see at that position (also boxed in orange) that some of the reads have a C and some have a T
- There were 5 SNPs like this that were present in isolate 3 from the teenager and mixed in Isolate 4 from the presumed source case
- So the SNP distance from Isolate 3 to the MRCA changed from 7 to 2 because 5 SNPs got filtered by the ambiguous bases filter

Case study: Presumed source case appeared to have a mixed infection and different subpopulations were transmitted



- In this example, the presumed source patient with the long infectious period appeared to have a mixed infection
- While a mixture of the subpopulations was present in the sample that was collected as indicated by the mixed SNPs, this source patient could have transmitted different subpopulations to others in the cluster

Summary

- **WGS of an isolate results in short sequence reads that get mapped to the H37Rv reference genome**
- **Positions that differ from the reference in any isolate in the comparison are identified**
- **Six filters are applied to remove positions with uninformative or unreliable SNPs**
  - If any isolate in the comparison does not meet all the filtering criteria, then the SNP is filtered out
  - SNP distances can change when new isolates are added to the comparison because of the filtering criteria

- In summary, WGS of an isolate results in short sequence reads that get mapped to the H37Rv reference genome
- Positions that differ from the reference in any isolate in the comparison are identified
- Six filters are applied to remove positions with uninformative or unreliable SNPs
- If any isolate in the comparison does not meet all the filtering criteria, then the SNP is filtered out
- SNP distance can change when new isolates are added to the comparison because of the filtering criteria

Summary (cont.)

- **Phylogenetic trees are constructed using the neighbor-joining method**
- **The reference strain H37Rv is used to infer the placement of the MRCA on the tree and the direction of genetic change**
  - The MRCA represents a genomic starting point before the SNPs that differentiate the isolates in the comparison occurred

- Phylogenetic trees are constructed using the neighbor-joining method
- The reference strain H37Rv is used to infer the placement of the MRCA on the tree and the direction of genetic change
- The MRCA represents a genomic starting point before the SNPs that differentiate the isolates in the comparison occurred

Acknowledgements

- **DTBE Applied Research Team**
  - Jamie Posey
  - Lauren Cowan
- **BioNumerics**
  - Hannes Pouseele
- **DTBE Molecular Epi Activity**
  - Ben Silk
  - Kala Raz
  - Clint McDaniel
  - Lydia Rautman
- **Michigan State Public Health Laboratory**
- **Wadsworth Center**
- **Association of Public Health Laboratories**

- We would like to thank members of the Applied Research Team and Molecular Epidemiology Activity at DTBE, BioNumerics, the Michigan State Public Health Laboratory, Wadsworth Center, 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.