TB genotyping, whole-genome sequencing, and molecular surveillance for recent transmission

Benjamin Silk, PhD, MPH
CDR, US Public Health Service
Lead, Molecular Epidemiology Activity, DTBE
National Tuberculosis Genotyping Surveillance Coverage* by Year: United States†, 2004–2018

* The proportion of positive cultures with at least one genotyped isolate.
† Includes 50 states and the District of Columbia.
§ For the year 2020, the national goal for TB genotyping surveillance coverage will change to 100%.

National Goal, 94%§

Proportion of culture confirmed TB cases genotyped (%)
Learning objectives

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

- Describe current uses of TB genotyping data for cluster alerting and detection
- Explain how to request and interpret WGS analyses for cluster investigation
- Describe the national transition toward use of WGS for TB molecular surveillance
TB Transmission and Course of Infection

~5% within 2 yrs. (recent transmission)

Active Disease (infectious)

~5% lifetime risk (reactivation)

Latent Infection (not infectious)

~95%
TB genotyping for cluster detection and alerting
TB Molecular Epidemiology: Targeting Recent Transmission

- **Goal**
  - Reduce the burden of TB by identifying where transmission is currently occurring and interrupting it

- **Challenge**
  - Distinguish recent transmission from cases infected long ago

- **Approach**
  - Combine molecular, clinical, and epidemiologic data to detect, investigate, and monitor recent TB transmission
Genotyping examines the DNA of \textit{M. tuberculosis} isolates from TB patients

- The \textit{M. tuberculosis} bacteria from a TB patient is called the patient’s isolate.
- Bacteria, including \textit{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 \textit{M. tuberculosis} is over 4.4 million nucleotides long.
Definitions for TB Genotyping in the United States

Spoligotype: 000000000003771

Initial 12-locus MIRU-VNTR\(^1\): 223325173533

PCRTypе: PCR00002

Additional 12-locus MIRU-VNTR (MIRU2): 444534423428\(^2\)

GENTypе: G00010

Sequentially assigned for each unique spoligotype and initial 12-locus MIRU-VNTR combination

Sequentially assigned for each unique spoligotype and 24-locus MIRU-VNTR combination

---

\(^1\) Mycobacterial interspersed repetitive unit–variable number tandem repeat.

\(^2\) The complete set of 24 loci is referred to as 24-locus MIRU-VNTR and is used for GENTypе designation for genotype in the United States.
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
Detecting Clusters of Recent Transmission using Genotyping

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

CDC cluster detection methods
- LLR cluster alerts: Unexpected increase in concentration of a genotype in a jurisdiction during a 3-year time period
- Large outbreak surveillance (LOTUS): 10 or more cases in a 3-year period related by recent transmission
County-based log-likelihood ratio (LLR)

- Compares the concentration of a genotype in a county compared with the rest of the country during a 3-year period

\[
LLR = a \times \log \left( \frac{\text{Obs}_\text{inside}}{\text{Exp}_\text{nation}} \right) + c \times \log \left( \frac{\text{Obs}_\text{outside}}{\text{Exp}_\text{nation}} \right)
\]

<table>
<thead>
<tr>
<th>County</th>
<th>Cluster genotype</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Inside</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Outside</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

- Observed prevalence of genotype in county (\(\text{Obs}_\text{inside}\)) divided by the expected prevalence in nation (\(\text{Exp}_\text{nation}\))
  
  \[
  \frac{a}{(a+b)} \\
  \frac{(a+c)\times N}{(a+b)\times N}
  \]

- Observed prevalence of genotype outside county (\(\text{Obs}_\text{outside}\)) divided by the expected prevalence in nation (\(\text{Exp}_\text{nation}\))
  
  \[
  \frac{c}{(c+d)} \\
  \frac{(a+c)\times N}{(c+d)\times N}
  \]
Alert levels based on LLR

- Higher LLR, greater likelihood of geographic clustering, suggestive of recent transmission

- TB GIMS generates alert levels based on LLR
  - No alert: LLR < 5
  - Medium alert: LLR 5 –< 10
  - High alert: LLR ≥ 10
Number of County-based TB Genotype Clusters* by Cluster Size, United States, 2016–2018

*Genotype cluster is defined as two or more cases with matching spoligotype and 24-locus MIRU-VNTR (GENType) within a county during the specified 3-year time period.
TB Genotype Clusters by TB GIMS* Alert Levels†, United States, 2016–2018

- No Alert, 957 (71%)
- Medium Alert, 306 (23%)
- High Alert, 86 (6%)

*Tuberculosis Genotyping Information Management System
†Alert level is determined by the log likelihood ratio statistic (LLR) for a given cluster, identifying higher than expected geospatial concentrations for a TB genotype cluster in a specific county, compared to the national distribution of that genotype; TB GIMS generates alert level notifications based on this statistic: “No alert” is indicated if LLR is between 0 – <5, “medium” is for LLR of 5 – <10 and “high” alert is for clusters with LLR ≥ 10.
Prioritizing TB Genotype Clusters

Outbreak Detection Work Group
Martin Clint, M.P.H., M.S., California Department of Public Health
Jason Cummins, M.P.H., Tennessee Department of Health
Wendy Cronin, Ph.D., M.S., Maryland Department of Health and Mental Hygiene
Lauren D'Immediato, Dr.P.H., M.S.P.H., Georgia Department of Public Health
Richard Duggan, Florida Department of Health
Cheryl Kearns, M.P.H., New York State Department of Health
Laura Lane, M.P.H., Texas Department of State Health Services
Nancy Robinson, M.P.H., Alabama Department of Public Health
Lynn Sosa, M.D., Connecticut Department of Public Health
Trudy Stein-Herr, M.S., Tennessee Department of Health
Tambol Shaw, M.P.H., California Department of Public Health
Jeanne Sullivan-Molisse, M.P.H., New York City Department of Health and Mental Hygiene
Shamee Poonsiri, M.P.H., Los Angeles County Department of Public Health

Division of Tuberculosis Elimination
Benjamin Silky, Ph.D., M.P.H.
Anne Marie France, Ph.D., M.P.H.
Tracie Gardner, Ph.D., M.S.
Smita Ghosh, M.S.
J. Stave Kammerer, M.B.A.
Lauren A. Lambert, M.P.H.
Clinton McDaniel, M.P.H., CPH.
Krista Powell, M.D., M.P.H.
Sandy Alphonson, M.P.H.
Maryam B. Haddad, M.S.N., M.P.H., F.N.P.
Jonathan Worthing, M.D., F.A.A.P.
Mark Miner, B.S.Ed.

https://www.cdc.gov/tb/programs/genotyping
Purpose

- Describe how to set up a routine systematic cluster assessment and prioritization system to review TB genotype clusters that would help:
  - Determine clusters that may indicate recent transmission
  - Identify, treat and prevent missed contacts and prevent infection
  - Identify opportunities to prevent bad outcomes (e.g., death, diagnosis delays, MDR-TB)
  - Identify locations where transmission may be occurring
  - Save resources by focusing on higher priority clusters
Considerations to Set Up Cluster Prioritization

- Identify key staff and establish roles
- Determine which clusters likely represent recent transmission or concerning characteristics
- Establish key criteria for cluster review, review frequency, and a process to prioritize for public health action
Steps and Outcomes of the Cluster Prioritization Process

- Step 1: Identify readily available data sources for genotype cluster review
- Step 2: Establish the current priority level of the cluster
- Step 3: Determine action items and next steps
- Step 4: Obtain additional information that is not readily available
- Step 5: Identify resource needs and key partners
- Step 6: Document review and decisions
- Step 7: Follow up and reconsider cluster prioritization as applicable
Considerations for whole-genome sequencing (WGS) to help focus public health action

- WGS may provide additional information to inform public health action:
  - Providing increased molecular resolution for a cluster of cases with a genotype that is common in the population or area;
  - Identifying a subset of cases where recent transmission is more likely to be occurring during an outbreak investigation;
  - Providing additional information that can distinguish cases attributable to recent transmission from cases that are due to reactivation of latent TB infection; and
  - Identifying or refuting possible epidemiological links.
Current *M. tuberculosis* genotyping is based on only ~1% of the genome

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

Genotyping provides low resolution for examining genetic relatedness of isolates

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

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

Whole-genome single nucleotide polymorphism (wgSNP) analysis

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

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

SNPs relative to H37Rv are identified

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

High-quality SNPs are mapped on to a phylogenetic tree

- SNPs in all isolates
- SNPs due to assembly errors
- Low confidence SNPs
Guide for interpreting the phylogenetic tree

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

MRCA = Most Recent Common Ancestor

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

Hypothetical Node
- Branching point with no circle
- Represents a hypothetical genome type
- No actual isolate with this genome type in the analysis
Guide for interpreting the phylogenetic tree

\[ \times \] = genetically distant, and unlikely to be involved in recent transmission

Closely related isolates, which may be involved in recent transmission
Guide for interpreting the phylogenetic tree

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

Consideration #1: Directionality cannot be inferred because cases involved in transmission may not be included on tree
Phylogenetic tree is not the same as a transmission diagram

Consideration #2: Directionality cannot be inferred because genetic changes could occur between the time of transmission and collection of the patient’s sample.
Phylogenetic tree is not the same as a transmission diagram

Consideration #2: Directionality cannot be inferred because genetic changes could occur between the time of transmission and collection of the patient’s sample.
Recent transmission is easier to rule out than to confirm with WGS

- Even isolates that are closely related or identical by WGS can be due to reactivation
  - This is because mutations may not occur as frequently during latent infection and therefore SNPs may not accumulate

- The phylogenetic tree should be used in conjunction with clinical and epidemiologic information to assess recent transmission
Case study on use of WGS for a confirmed outbreak in a high TB incidence jurisdiction
Background

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

= County A

MRCA

6 outbreak TB cases with epidemiologic links

Isolate from TB case with possible epi link to County A
Interpretation

- = County A

= Isolates from TB cases excluded from outbreak investigation because they are genetically distant

6 outbreak TB cases with epidemiologic links

Isolate from TB case with possible epi link to County A
New Clustered TB Case

= County A

Isolate from a recently diagnosed patient in County A
Public Health Outcomes

- Avoided unnecessary investigation of 7 cases, including 5 residing in different counties outside of County A
- WGS results enabled continued focus on 6 cases linked by recent transmission
- County A intensified work to identify, evaluate, and treat contacts to outbreak cases
- County A also investigating the new patient whose TB is genetically closely related to the outbreak to determine if/how linked to outbreak
National transition to WGS
Universal prospective WGS began in 2018

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

TB Genotyping Methods and Data Flow (2018 – 2020)

DNA

WGS data ➔ Spoligo

MIRU-24

- GENType ➔ TB GIMS
- Cluster Alert

- Yes
- No

wgMLST ➔ wgMLSType (TB GIMS – 2021)

wgSNP ➔ Phylogenetic Tree
## Analysis of clustering using WGS data: wgMLST vs. wgSNP

<table>
<thead>
<tr>
<th></th>
<th>wgMLST (whole-genome multilocus sequence typing)</th>
<th>wgSNP (whole-genome single nucleotide polymorphism)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of analysis</td>
<td>all isolates</td>
<td>isolates in a cluster</td>
</tr>
<tr>
<td>Use</td>
<td>assigning isolates to a wgMLSType that can be used for cluster alerting</td>
<td>examining genetic relationships among isolates</td>
</tr>
<tr>
<td>Output</td>
<td>wgMLSType (short string of numbers similar to a GENType)</td>
<td>phylogenetic tree</td>
</tr>
</tbody>
</table>
wgMLSType will replace GENType for cluster alerting in 2021

TB Genotyping Methods and Data Flow (2021)
Application of molecular surveillance
Genotyped Tuberculosis Cases Estimated to be Attributed to Recent Transmission, United States, 2017–2018

Recent Transmission*
12.6% (n=1,712)

Limited Recent Transmission

Extensive Recent Transmission†

8.3% (n=1,123)

4.3% (n=589)

Not Recent Transmission§
(n=11,889)

* A TB case is designated as attributed to recent transmission if a plausible source case can be identified in a person who i) has the same *M. tuberculosis* genotype, ii) has an infectious form of TB disease, iii) resides within 10 miles of the TB case, iv) is 10 years of age or older, and v) was diagnosed within 2 years before the TB case.

† A TB case is designated as attributed to extensive recent transmission when the criteria above for recent transmission are met, and furthermore the case belongs to a plausible transmission chain of six or more cases. Otherwise, the case is designated as attributed to limited recent transmission.

§ Cases not attributed to recent transmission may be misclassified in children <5 years old or indeterminate in persons with a recent U.S. arrival due to limitations of the plausible-source case method.
Genotyped Cases Estimated to be Attributed to Limited and Extensive Recent Transmission, United States, 2015–2018

*A TB case is designated as attributed to recent transmission if a plausible source case can be identified in a person who i) has the same *M. tuberculosis* genotype, ii) has an infectious form of TB disease, iii) resides within 10 miles of the TB case, iv) is 10 years of age or older, and v) was diagnosed within 2 years before the TB case.

† A TB case is designated as attributed to extensive recent transmission when the criteria above for recent transmission are met, and furthermore the case belongs to a plausible transmission chain of six or more cases. Otherwise, the case is designated as attributed to limited recent transmission.
Percentages of Tuberculosis Cases Estimated to be Attributed and Not Attributed to Recent Transmission, by Origin of Birth*, 2017–2018

* Cases with unknown origin of birth not shown (n=21).
† A TB case is designated as attributed to recent transmission if a plausible source case can be identified in a person who i) has the same *M. tuberculosis* genotype, ii) has an infectious form of TB disease, iii) resides within 10 miles of the TB case, iv) is 10 years of age or older, and v) was diagnosed within 2 years before the TB case.
§ Cases not attributed to recent transmission may be misclassified in children <5 years old or indeterminate in persons with a recent U.S. arrival due to limitations of the plausible-source case method.
This warning banner provides privacy and security notices consistent with applicable federal laws, directives, and other federal guidance for accessing this Government system, which includes (1) this computer network, (2) all computers connected to this network, and (3) all devices and storage media attached to this network or to a computer on this network. This system is provided for Government-authorized use only. Unauthorized or improper use of this system is prohibited and may result in disciplinary action and/or civil and criminal penalties. Personal use of social media and networking sites on this system is limited to not interfere with official work duties and is subject to monitoring. By using this system, you understand and consent to the following:

- The Government may monitor, record, and audit your system usage, including usage of personal devices and email systems for official duties or to conduct HHS business. Therefore, you have no reasonable expectation of privacy regarding any communication or data transiting or stored on this system. At any time, and for any lawful Government purpose, the government may monitor, intercept, and search and seize any communication or data transiting or stored on this system.
- Any communication or data transiting or stored on this system may be disclosed or used for any lawful Government purpose.

TBGIMS Rules of Behavior

Use of TB GIMS data from TB program jurisdictions outside your own jurisdiction for the purpose of research or public presentations or for release to outside parties (including the news media), must be approved in advance by the states that provided the data as well as CDC.

Disclaimer - CDC does not warrant or assume any legal liability for the accuracy or completeness of the data.

CLA Disclaimer - Genotyping data are obtained for surveillance purposes only. Such testing may not be performed in compliance with CLIA (compliance with CLIA is not a requirement because the testing is for surveillance purposes rather than diagnostic) and therefore these data cannot be used for patient diagnosis or treatment.
TB GIMS Watch List

- Available to all TB GIMS users for their jurisdiction

- Saved search on a specific genotype and jurisdiction
  - Notifies users when isolate or a linked patient record is added to TB GIMS
    - In their own jurisdiction (state) or outside their jurisdiction (national)
    - At individual level or group level for institutional memory and continuity of operations
Thank you!

For more information, contact CDC
1-800-CDC-INFO (232-4636)

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.